

# ბავშვთა და მოზარდთა ფსიქოლოგიური შეფასება და დიაგნოსტიკა 2

რიდერი

შემდგენელი ლუიზა არუთინოვა

## სარჩევო

1. Cecil R. Reynolds, Randy W. Kamphaus - Handbook of psychological and educational assessment of children :
2. personality, behavior, and context
3. DAVID W. BARNETT, GREGG M. MACMANN, FRANCIS E. LENTZ, JR. - Personality Assessment Research:Applying Criteria of Confidence and Helpfulness
4. ***PSYCHOMETRIC FOUNDATIONS OF ASSESSMENT***
5. KELLY ROBINSON - Assessment of Childhood Anxiety
6. THE BECK SCALES IN CLINICAL PRACTICE
7. CYNTHIA A. RICCIO CECIL R. REYNOLDS- The Assessment of Attention via Continuous Performance Tests
8. **Global or Big-Five Structure**
9. **The Eysenck Personality Scales: The Eysenck Personality Questionnaire-Revised (EPQ-R) and the Eysenck**
10. **Personality Profiler (EPP)**
11. The Sixteen Personality Factor
12. ***MINNESOTA MULTIPHASIC PERSONALITY INVENTORY—ADOLESCENT***
13. **The Beck Depression Inventory-II (BDI-II),Beck Hopelessness Scale (BHS), and Beck Scale for**
14. **Suicide Ideation (BSS)**
15. JENNIFER THORPE RANDY W. KAMPHAUS CECIL R. REYNOLDS - The Behavior Assessment System for Children
16. TIMOTHY A. CAVELL BARBARA T. MEEHAN SAMUEL E. FIALA - Assessing Social Competence in Children and Adolescents

17. TIMOTHY A. CAVELL, BARBARA T. MEEHAN, SAMUEL E. FIALA - Children and Adolescents

***18. SENTENCE COMPLETION METHODS***

Handbook of psychological and educational assessment of children :  
personality, behavior, and context / edited by Cecil R. Reynolds and  
Randy W. Kamphaus.— 2nd ed.p. cm.  
DAVID W. BARNETT GREGG M. MACMANN FRANCIS E. LENTZ, JR.

Personality assessment has had a tentative status within professional psychology for the past several decades and, as we indicated in the last edition, an important question for many professionals in light of the difficulties is “Why hold on?” Our answer was not optimistic—not with regard to the construct of *personality* per se but to the professional *uses* of constructs typically associated with personality measurement. The proliferation of seemingly shallow approaches to assessment and intervention is cause for concern, and perhaps all would agree that the ultimate criteria for the evaluation of clinical and educational decisions should rest on enhanced personal and social development (i.e., Seligman & Csikszentmihalyi, 2000) rather than the more easily measurable objectives ordinarily associated with test development.

An underlying theme of the earlier chapter was the need for new research consistent with the goal of defining “an epistemology of practice” (Schön, 1983)—research that involves the analysis of successful practice in real-life situations. Maintaining that emphasis in this chapter, we offer an opinion about the kinds of research information that would be helpful in developing professional practices for the assessment of child and adolescent personality. However, in a natural evolution of our views in regard to the most appropriate criteria for personality measurement, our current discussion of the issues is framed in terms of desirable measurement qualities for professional practice, namely, decision confidence and helpfulness in natural setting contexts.

#### AN OVERVIEW OF THE CHAPTER

Despite the enormous controversies, personal and social development remains the broadest context for assessment and intervention design that serves to organize an array of professional practices involving clinical, educational, and vocational decisions about problems of children and adolescents. Guiding principles are clearly needed to enable effective outcomes in the difficult and ambiguous situations in which professionals need to apply the tools of their trade with confidence. However, almost since their inception, personality measures have been

subject to extensive criticism as summarized in the following paragraphs. First, the difficulties in “knowing” the personality of an individual are fundamental

3

# 1

## Personality Assessment Research: Applying Criteria of Confidence and Helpfulness

DAVID W. BARNETT GREGG M. MACMANN FRANCIS E. LENTZ, JR.

and raise serious questions related to the degree of *confidence* that can be justifiably entertained in making decisions during problem solving—a process limited in part by error rates as well as by meaningful predictions. Competing theories and measurement differences make any consensus about appropriate assessment practices among professionals unlikely. The crucial procedure for practice becomes one of empirically demonstrating useful information through probabilistic statements—a question of *helpfulness* in decisions. However, the need to account for “person” variables is unassailable. This is an important distinction we return to later. Second, personality assessment for children has special difficulties (again, both theoretical and practical). The issues are made evident by research that deals with the continuity of personal and social development and the processes of psychosocial change. Evidence exists for both the coherence and continuity of development; yet the consistencies described confidently for groups are not easily defined or measured for individuals and individuals may change significantly (Block & Block, 1980; Lewis, 1997; Moss & Sussman, 1980). When viewed from the perspective of individual development, Emmerich’s (1966) review has stood the test of time: “Impressive as the evidence for the early determination of personality may seem, there are also grounds for believing that personality ordinarily remains open to change over extended periods” (p. 237). The task for professional practice is one of enhancing the developmental trajectory of high-risk or challenging children, and surprisingly little is known about the process. A major goal of personality research has been the identification of traits or person variables, or pathological markers, that

would improve the selection of specific and effective treatments. Such attributes remain elusive thus far, and individual applications are tenuous, based on understanding problem situations. Intervention design is in its infancy, and the course of personal and social development is complexly determined by psychological, biological, and accidental events (Bandura, 1986; Plomin, DeFries, McClearn, & Rutter, 1997).

A third issue has been the debate concerning the relative importance of situational versus personal determinants of behavior. The focal points have been the obvious control that situations exert on behavior, and the relatively modest predictive power of many personal variables. War was waged over this issue in the 1960s (Mischel, 1968; Peterson, 1968), but most have assumed an interactionist position (Magnusson & Endler, 1977; Plomin et al., 1997). Perhaps the debate has been put to rest: "Behavior, cognitive and other personal factors, and environmental influences all operate interactively as determinants of each other" (Bandura, 1986, p. 23). The relative influence of specific factors in any unique situation will certainly vary across persons, situations, and behaviors. It would be easy simply to state that more research is needed in personality assessment techniques. In the past 10 years, that is where the field has gone, but without much apparent gain. Research is still needed to verify useful constructs and to clarify the potential applications of personality measurement to individual children and intervention design. In other words, decisionmaking research guided by the natural questions of psychological practice, such as how to help in natural problem contexts, is still the most critical gap at present.

In this chapter, we first analyze personality measurement and related topics of psychopathology within a unified validity framework that encompasses the traditional categories of reliability and validity. Our focus is on central theoretical and practical issues defined in terms of confidence and helpfulness of decision outcomes related to test use. Second, we present fundamental dilemmas for research and professional practices associated with personality measurement. Third, we make suggestions in the form of component strategies for personal and social assessment. The challenges of behavioral assessment also are integrated

into the discussion because they are inherently related to child intervention outcomes.

## CRITICAL CONCEPTS OF PSYCHOMETRIC QUALITY

The psychometric qualities of measurement traditionally have been analyzed via two separate but intertwined constructs—reliability and validity (duly compartmentalized in terms of the “three c’s” of construct, con-

### 4 I. GENERAL ISSUES

tent, and criterion-related validity). These measurement qualities are likewise treated separately in the technical manuals of most tests. Following Messick (1989, 1995), we frame our discussion of psychometrics within a “unified” theory of validity. Consistent with that framework, the adequacy and appropriateness of measures can be evaluated

through a series of pointed questions, as follows:

1. *What is the meaning of an inference made from a test score?* This is easily recognized as an aspect of validity evidence

(American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) and is critical within the unified view of validity. Perhaps especially within the assessment of personality variables it is imperative that the meaning of any score derived from a test be clearly established to determine the appropriateness of its use to make assessment decisions.

Score meanings can be supported through a number of related considerations such as the relevance and representativeness of test content, nomological relationships with other construct measures, and generalizability of scores within the content domain (as shown through reliability analyses).

However, beyond their theoretical or substantive import, scores have social meaning as well (derived from the consequences of score use). We believe that the latter idea has not received sufficient emphasis in either test development or use.

2. *Given an understanding of the meaning of a score, can you use it to meet some purpose (answer some assessment question)?*

All assessment should be driven by questions and purposes. This is the characteristic toward which traditional validity and reliability estimates have been somewhat loosely aimed. For example, the reliability of some test inference is important in evaluating whether you can consistently make a particular decision. Likewise, various means of assessing

validity have to do with how well an assessment decision meets a defined purpose.

3. *Should you use a metric to meet some defined purpose?* Many methods for judging validity and reliability have been developed in regard to the first two sets of emphases (shown previously). This question more clearly places emphasis on the *consequences* of psychological decision making in regard to desirable outcomes for clients and includes analyzing the values that accompany measurement (e.g., labels applied to a client because of test results).

A set of critical assumptions underlie this perspective, as follows:

1. Perhaps the most important assumption is that validity can only be assessed within the boundaries of some clearly understood assessment purpose. This is (and has been) the position of major professional organizations in regard to appropriate test interpretation (AERA et al., 1999). That means that the concept of *test* validity per se is inappropriate, and that inferences from a test score(s) may be made validly or not *given some particular purpose*. If the intent is to use a test score for multiple purposes, these purposes must be articulated and examined separately. One of the biggest shortcomings of test manuals is that purposes intended for derived scores are only vaguely outlined, and evidence concerning validity is not typically tied to any specific purpose. This is a dangerous practice in that it may lull practitioners into ill-supported decisions about their use of test scores.

2. Treatment (or other) decisions are nearly always the result of *multiple* assessment activities. This is not the way validity has been analyzed within the psychometric literature, nor within test manuals where single instruments are analyzed for general psychometric qualities as opposed to qualities tied to a specific test purpose.

3. Even given multiple measures, decisions are based on the inferences from test scores. It is these *inferences*, not particular test scores, that ultimately must be validated.

4. Psychometric qualities such as reliability (a critical facet of unified validity) must be judged within coherent units of assessment decisions (e.g., diagnosis, screening, developing problem hypotheses, and selecting appropriate treatments), and based on what aspect(s) of a test will guide a particular decision. For example, if a profile of subtest

scores is used to make a diagnostic inference, it is the reliability of that *profile* that is critical, not the set of subtest reliability estimates or the overall measure of test

1. Personality Assessment Research 5 reliability. Using diagnosis as a continuing example, the meaningfulness of a diagnostic category, accuracy of classification decision, and the consequences of making the diagnosis must be analyzed to judge if one can and should use test data to make the diagnostic inference. If a single test item were used to make an inference, it is the reliability of that inference (and the single item) that is important. Decision-based reliability indices may bear little resemblance to traditional reliability estimates.

5. Judgments about client outcomes as the critical arbiters of validity are complex and must ultimately involve subjective analysis based on clearly described values.

6. Analysis of the contribution of any single measure within an array of measures used in making a decision is most closely related to what has been called incremental validity (Sechrest, 1963), and can be profitably analyzed. For example, if a single measure adds some “amount” to improved outcomes for children, it adds some increment of validity. Validity remains a unified overall construct.

It is in this spirit of incremental validity that we turn to the traditional constructs of reliability and validity. Inferences derived from a personality instrument will still need to be consistent in order to decide that a score can be used to help make some decision. Thus, reliability as consistency can still be, and should be, analyzed as an important facet of validity. As we will see, traditional views of reliability leave something to be desired, even in consideration of incremental improvement to overall validity.

## ESTIMATING THE RELIABILITY OF SINGLE SCORES

“Reliability” has generally pertained to the consistency of measurement and enables an estimation of test error. As stated earlier, the real issue for practice is the estimation of the stability of an inference or a decision based on an inference. The reliability of a test or score is not the critical level of psychometric analysis, even though this is the way such data are nearly always presented.

We see the construct of reliability as an important facet of validity, and most directly

tied to the second evaluation question defined previously (“Can you use a score?”). Unfortunately, classic numerical estimates of reliability can only *indirectly* inform a practitioner about the actual consistency of an inference or decision. The only direct information concerning the confidence one might place in a diagnostic inference concerns the probability of making the same inference or decision after a second measurement occasion with the same or different examiners, across practitioners given the same information, or other combinations of possibilities. These sorts of probability analyses are seldom provided to assessors by test developers. Further, if decision consistency became the basis of reliability, it would obviate many of the arguments about which reliability coefficient is appropriate. Given these caveats, it is still important to review the typical ways in which reliability is examined, and the issues which need to be understood by practitioners as they make decisions about instrument use.

#### **Classic Construct of Test Reliability**

The classical reliability model developed by Spearman at the turn of the 20th century assumed that each score has two components: a “true score,” which is estimated from a person’s obtained or “observed score” on a measure, and “test error.” Traits are assumed to be stable, and error is assumed to be random or uncorrelated. The reliability of a test may be defined by the variation (or variance) in true scores divided by the variance in observed scores, or the ratio of true score to observed-score variance (e.g., Thorndike, 1982). The definition of reliability from classical test theory typically makes the assumption of hypothetical repeated measurements of an individual’s performance on parallel tests, so that fluctuations in scores enable the estimation of error.

Although intuitively attractive and amenable to statistical manipulations, the classical model has significant complexities that have been debated for nearly as long as the model has dominated measurement theory (Lord & Novick, 1968). The greatest concern is that there is no direct link between the actual observed score and the hypothetical true score; in psychology, the true score is always inferred, and for many traits critics have argued that the true score may

#### **6 I. GENERAL ISSUES**

exist more vividly in the minds of test developers

than in behavioral evidence. The definition of reliability based on classical test theory cannot be directly tested, and there are significant theoretical divisions in reliability theory. Further, there is no single method of estimating the reliability of a test score. A variety of methods to estimate reliability in a classical framework exists based on both logic and specific statistical and experimental procedures.

Some think that the concept of true score should be abandoned (e.g., Lumsden, 1976). However, the classical model and its derivatives have dominated test theory, and therefore test development (Ghiselli, Campbell, & Zedeck, 1981). Thus, many contemporary practices associated with test use are based on the assumptions that stem from this model.

For any test, reliability can be estimated in numerous ways; however, the methods account for different sources of error. Traditionally, experimental and statistical procedures are used to present reliability data in four ways: internal consistency, test–retest reliability, alternate forms, and analysis-of-variance procedures. Many discussions stop at this point. However, practical problems are associated with the fact that different estimates of reliability address different aspects of the consistency of measurement.

Perhaps the most important criticism related to any view of reliability as a quality of a score or as a quality of a decision (these are related) is that estimates of reliability based on a single procedure are nearly always underestimates of actual instability of both scores and inferences. As we discuss, realistic numerical estimates of score reliability must take into account different sources of error available only through simultaneous consideration of reasonably composed estimates of different reliability facets. In addition, estimates of the reliability of a *decision* based on a score can only be estimated by examining studies of various aspects of actual decision probabilities, not just errors associated with a score.

### **Estimation of Reliability as Internal Consistency**

The reliability coefficient of a scale can be estimated from a single administration. Statistical analyses are used to determine the degree of “internal consistency” of responses to test items. An early procedure was to (1) “split” the test into equivalent parts, (2)

correlate the halves, and (3) “correct” the obtained correlation that resulted from the briefer scale to estimate the reliability coefficient for the full-length test. The Spearman–Brown prophecy formula, the basic method to estimate the resulting reliability from the test halves, has other uses as well and is discussed in a subsequent section concerned with aggregation (see Equation 3). A basic problem is that the resulting reliability coefficients derived from splitting the test in various ways (e.g., odd–even and first half–second half) will vary.

Coefficient alpha (Cronbach, 1951) provides a more general solution for internal consistency reliability, in effect representing the average of all possible item splits. It is derived from the variance of the items relative to the variance of the total test, or from the average interitem correlations, and establishes the “upper limit” for the test’s reliability based on content sampling (Nunnally, 1978, p. 230). If coefficient alpha for a measure is low, the items have little in common and/or are too few in number.

Internal-consistency estimates are widely used in practice and are especially important for constructs hypothesized to change or fluctuate over time. Although it appears straightforward, coefficient alpha does present interpretative difficulties. Because coefficient alpha increases as a function of the magnitude of interitem correlations and number of correlated items, scales with similar reliabilities may have different properties. High reliabilities may be achieved through a small number of highly correlated items or through large numbers of items with modest or even low interitem correlations.

Tests that have high estimates of internal consistency may be measuring a homogeneous trait, asking redundant questions, or asking highly related questions repeatedly.

### **Estimation of Reliability through Test–Retest and Alternative Forms**

Equally straightforward in description, the test–retest method involves a second administration of a scale within a specified inter-

1. Personality Assessment Research 7  
val (usually about 2 weeks). The results of the two administrations are correlated; thus, an estimate of the stability of the measurement is provided. It has been argued that test–retest reliability is less important to personality measurement because of a number of confounding variables: the unknown

effects of two administrations of the questions or test stimuli, unknown changes in life circumstances during the retest interval, changes that occur in the testing situations, and the unknown influence of these factors on estimates of the stability of the trait.

Reliability estimates also can be derived from alternative or equivalent forms of a scale, although this method is less frequently used in personality measurement. Simply, two (or more) measures of a trait developed from an item pool are administered and the results correlated to produce reliability estimates. When this method is used to examine test-retest reliability, the effects of memory in responding to specific items are eliminated.

From the viewpoint of professional practices, the stability estimates based on test-retest procedures are often important.

Practitioners are quite concerned about potential fluctuations of trait estimates over brief periods, and the corresponding degree to which intervening variables may influence interpretations of various constructs and thus decisions that stem from test usage.

Time-limited decisions are troublesome (even embarrassing) for many professional practices. A low test-retest reliability coefficient is only problematic if the behavior measured is presumed to be relatively stable.

In this case, questions can be raised about the measurement procedure, the trait, or the power of intervening variables.

#### **Estimation of Reliability through Analysis-of-Variance Procedures**

Related to the classical model, analysis-of-variance procedures (Thorndike, 1967, 1982) and generalizability theory, advanced by Cronbach, Gleser, Nanda, and Rajaratnam (1972) have been proposed. Logically, the procedures enable questions to be raised about both the true-score and error components. Thorndike (1982) writes from a psychometric perspective: "What is included under the heading 'error variance' depends on how the universe that the test score is presumed to represent is defined, with certain sources of variance being treated as error under one definition . . . and as true score under another definition" (p. 157).

The *realities* are that when tests are administered on different occasions, under different conditions, or through different procedures, variation in true-score estimates is

a rule rather than an exception. Generalizability theory recasts the reliability problem “into a question of accuracy of generalization” (Cronbach et al., 1972, p. 15). An analysis is made of how test scores that result from a particular measure *generalize* to other subjects and to the universe of possible conditions. The conditions that are potentially examined include tasks, time of observations, situations or settings, and the performance of different raters. Classical notions of test error have been challenged in a similar way by behaviorists. The questions raised are important because the criteria typically used are tied directly to issues of usefulness by way of intervention design. Concerns related to classical measurement assumptions are described by Cone (1981; see also Johnston & Pennypacker, 1993): “Error is just a blanket way of referring to a host of ‘don’t knows,’ none of which are random. . . . [T]he term ‘random’ is really a pseudonym for ‘haven’t found out yet’ . . . since all behavior is lawful whether we know the controlling variables or not” (p. 61). Analysis-of-variance procedures were for a time deemed of interest to behavioral assessors. However, their overall contribution is likely to be limited, as the central tenets are different and behaviors of concern may be idiosyncratic, may have a low rate of occurrence, and may be episodic (Hayes, Nelson, & Jarrett, 1986). The hallmark of behavioral assessment, functional analysis, requires that a wide range of antecedent and consequent events be potentially amenable to appraisal as well. Behavioral assessment approaches have not resolved the basic measurement dilemmas, but the topics are dissimilar. In general, the most critical reliability and validity questions in behavioral assessment apply to (1) problem identification, (2) treatment selection, (3) treatment adherence, and (4) outcome measurement

### 8 I. GENERAL ISSUES

of the therapeutic process (Barrios & Hartmann, 1986). These are not well treated through analysis-of-variance procedures. However, some studies are quite useful. For example, Hansen, Tisdelle, and O’Dell (1985) used generalizability procedures to establish the overall equivalence of audio versus direct observation of parent–child interactions, although low-frequency behaviors likely to have clinical significance had

to be dropped from the analysis (e.g., crying and destructiveness).

Generalizability studies pertinent to professional practices are few (cf. Jones, Reid, & Patterson, 1975), and the methods based on analysis-of-variance procedures typically have rigorous design requirements that hamper many applications. Multifaceted studies require large samples, and clinical realities introduce the potential for confounding design factors that are difficult to control, especially with more than two facets (Thorndike, 1982). The conceptual issues, however, are clearly critical to individual assessment within the model of unified validity. Consistent with Messick's (1989, 1995) perspective, with generalizability theory, the distinction between reliability and validity is deliberately blurred.

The model enables questions to be raised about "error." Sources of error become information from an experimental view. The study by Hansen and colleagues (1985) illustrates both the power of the techniques associated with analysis-of-variance procedures and their difficulties: General relationships may be studied, but specific behaviors require idiographic measurements.

#### **Applying Estimates of Reliability Standards for Reliability**

Well-established conventions for reliability coefficients have been described as follows (Nunnally, 1978; Stagner, 1948). Reliabilities for research purposes should range from .70 to .80 to be deemed "acceptable." For individual decisions, it is suggested that reliabilities should be .90 at a minimum, with coefficients of .95 the desirable standard.

These standards for individual decisions are commonly ignored by scale developers and consumers. However, as it turns out, even scales that meet these conventional standards may not work as expected in applied decision contexts (see "Reliability of the Inference").

Reliabilities can be directly interpreted as explained (or true) variance. However, for individual decisions, the "standard error of measurement" generally provides a more meaningful way of representing measurement precision in practice (though often biased by incomplete representation of the sources of variability in scores). Conceptually, the standard error of measurement is the standard deviation or variability of observed scores, given a true score, based on the hypothetical repeated measurement of

an individual with parallel tests. In practice, an individual's true score is not known but is estimated from a fallible observed score. The appropriate statistic for this situation is the "standard error of estimation," or standard deviation of true scores given an observed score (Dudek, 1979). The standard error of estimation is used to establish a "confidence interval," which indicates the probability that an individual's true score lies within a specified range over repeated "theoretical" test occasions; estimates of true score (based on observed scores) vary around the actual true score because of error, and thus presumably because of chance factors.

The conceptualization of standard error is of more practical significance than a coefficient of reliability because it shows a practitioner something about score stability and, indirectly, the stability of a decision. One of many significant problems occurs when some sort of cut or criteria score is used to make a decision (school-based screening, for example). Such decisions are basically dichotomous and any sort of standard error is insufficient to describe the likelihood of making the same decision across practical facets (other examiners, forms, occasions, etc.). Although there is a well-described literature concerning this type of issue (and the related issue surrounding ideas of "false positives" and "false negatives"), information does not seem available for most common practices. As we discuss later, when there is information about decision analysis of cut score use, it often has been discouraging with respect to the adequacy of decisions.

Three issues concerning standards for reliability are of primary importance for research on professional practices. The first two are considered here; the third bridges the gap between traditional views of reliability and validity.

First, confidence intervals that apply to the prediction of true scores should be established around the "estimated true score."

Nunnally (1978) points out the "obtained scores are biased estimates of true scores. Scores above the mean are biased upward, and scores below the mean are biased downward" (p. 217). More extreme scores contain more "error," and thus more bias (i.e., the error component for a score that is 2 standard deviations from the mean is twice

as large as the error component for a score that is 1 standard deviation from the mean). Obtained scores easily can be adjusted to enable the use of an estimated true score ( $x_{\text{t}}$ ) for creating confidence intervals, as follows:

$$x_{\text{t}} = X_{\text{t}} + (r_{xx}) (x - X_{\text{t}}) \quad (1)$$

where  $X_{\text{t}}$  is the mean of  $X$ ,  $r_{xx}$  is the reliability of  $X$ , and  $x$  is the observed score. Although this issue is often trivial in a numerical and statistical sense (scores may be “adjusted” by small amounts), an important exception to the triviality occurs when specific cutting scores are used for professional decisions—implicitly or explicitly, a common practice.

Second, the calculation of standard error of measurement should be placed within the context of a specific reliability estimate, intended for a specific function (Sechrest, 1984). Many test authors use internal estimates of reliability (e.g., coefficient alpha) to establish confidence intervals for test users. This type of reliability only takes into account the consistency and number of items within the domain *as represented by the particular test developer*. Confidence intervals to describe the stability of the behavior over time, and the generalizability of estimates across potential raters and alternative measures within the same domain (as defined by other test developers), are often unexamined. For these situations, the “standard error of prediction,” or standard deviation of expected scores across raters or occasions, is the appropriate statistic (Dudek, 1979; Lord & Novick, 1968; Schulte & Borich, 1985). Third, and most important, reliability must be viewed within a context of validity. For example, the standard errors of prediction associated with independent estimates of the construct or behavior of interest need to be considered—a discussion resumed below in regard to multitrait–multimethod analyses.

#### Reliability of Inferences from Test Scores

The errors of measurement reported in manuals have a combined effect with regard to a score inference. Anastasi has pointed this out in many editions (e.g., Anastasi & Urbina, 1997). The focus on specific estimates of reliability, which may be high, tends to underestimate the problems of individual interpretation of scores.

To estimate the extent of the problem, a theoretical approximation of the combined effect of measurement errors can be obtained

through variance components analysis (L. Feldt, personal communication, September 26, 1991), in which the reliability coefficients for independent sources of variability in scores (e.g., content, time, and examiners) are represented by  $r_{yy}$  (subscript 1 through  $n$ ):

1

$$1 + \frac{r_{yy1}}{n}$$

$r$

—

$yy$

$r$

1

$$r_{yy1} + \frac{r_{yy2}}{n}$$

$r$

—

$yy$

$r$

2

$$r_{yy2} + \dots + \frac{r_{yyn}}{n}$$

$r$

—

$yy$

$r$

$n$

$r_{yyn}$

(2)

As an example, a popular rating scale that measures internalizing problems has a coefficient alpha reliability of .78 and a test-retest reliability of .76 for teacher ratings at the elementary school level. From the aforementioned formula, the combined effects reliability is .62, which would be the reliability most appropriately associated with an inference about internalizing behaviors for an individual child based on the score. Depending on the decision purpose, the scale user would then build a “confidence” interval that would show how much confidence there may be in an inference related to “internalizing.” However, this reliability estimate is still optimistic because other reliability facets could also be addressed (i.e., other potential raters), as well as comparisons with other instruments purporting to measure the same construct, a matter we will turn to again in the section on validity.

#### 10 I. GENERAL ISSUES

One would not have much confidence in the inference of “internalizing.”

Aggregation to Improve Reliability

## of Inferences

The fundamental method to increase the reliability of a measure is simply to increase the number of items. The relationship between test length and reliability is defined by the Spearman–Brown formula:

$$r_{xx} = \frac{k r_{xx}}{1 + (k - 1) r_{xx}} \quad (3)$$

where  $r_{xx}$  is the reliability of expanded test form,  $r_{xx}$  is the original reliability of measure  $X$ , and  $k$  is the proportionate increase in test length. For example, in determining split-half reliability,  $k = 2$ .

Through a series of widely cited studies, Epstein (1984) has applied this basic relationship to issues concerning personality measurement. Rushton, Brainerd, and Pressley (1983) also have reviewed the principle of aggregation and reached a similar conclusion: “Many weak or unconvincing correlational relationships found in the personality, social, and cognitive development literature are consequences of failure to aggregate” (p. 18).

Epstein (1984) theorized that “personality coefficients,” a disparaging term applied to low but significant correlations of .20–.30 often found in personality research, might be the results of a failure to aggregate based on the analysis of units of behavior that were simply too small; these results might be similar to the results achieved when looking at item correlations. He summarized the effects of different aggregation strategies in the following manner:

1. Aggregation over stimuli and situations reduces error variance associated with the uniqueness of particular stimuli and situations.
2. Aggregation over occasions and over trials within occasions reduces error variance associated with changes over time.
3. Aggregation over judges or raters reduces error variance associated with individual differences among judges.
4. Aggregation over modes of responses reduces error variance associated with different response modes. . . . (pp. 260–261)

$$r_{xx} = \frac{k r_{xx}}{1 + (k - 1) r_{xx}}$$

Through aggregation, Epstein was able to achieve respectable correlations (as high as .80 and .90) for traditional constructs associated with personality measurement. Epstein (1984) summarized the results in the following way: “Stable response dispositions can be demonstrated when responses are averaged over adequate samples of behavior but not necessarily when single instances

of behavior are observed” (p. 214).

The respectable stability coefficients obtained through aggregation of responses, he argued, lead to improved evidence for validity across a number of constructs and measures.

Through aggregation, Epstein found support for the following statements: “(a) behavior is situationally specific, (b) behavior is general across situations, and (c) people have broad cross-situational response dispositions” (p. 263).

Of course, nothing is this simple in personality measurement, and aggregation is not a panacea, as pointed out by Epstein (1984) and others (Bandura, 1986; McFall & McDonel, 1986; Mischel, 1984b; Mischel & Peake, 1982). First, aggregation leads to a corresponding loss of information.

If observations are aggregated over any of the four dimensions listed previously, variance attributed to the dimensions is termed “error.” For example, if observations are aggregated over situations, then information related to the potential variance of situations may be less readily apparent.

Second, Mischel (1984b) argues that Epstein’s research does not deal with the central issues of “the classic personality debate”:

the manner in which individuals *discriminate* situations in relationship to their social behavior and the usefulness of inferring traits in predicting actions in a given context. In Bandura’s (1986) words,

“Aggregation inflates correlations but yields indeterminate or empty predictions” (p.

10). Furthermore, a high degree of behavioral consistency logically cannot be used to argue for the existence of traits without other considerations pertaining to validity (see McFall & McDonel, 1986, for a discussion of these concerns in addition to other related problems).

Finally, aggregation as discussed by Epstein (1984) relates to aggregation of observations of behavior, whereas personality has been much more typically measured in prac-

1. Personality Assessment Research 11  
tice by administration of a test of one sort or another. Thus, the idea of aggregation in practice appears most useful in basic research regarding propensities to respond or possibly an analysis of construct facets. Although aggregation is perhaps useful in some conceptual or theoretical analyses, it would seem of little use in examining the reliabilities of decisions based on typical personality

measures.

Aggregation may elucidate one area of difficulty but, in doing so, obscure others. Although aggregation may better reveal the coherence in some behaviors related to personality constructs, social and cognitive behaviorists argue that personality processes should be linked to psychological processes that result in a better understanding of individual patterns of behavior within socially important contexts.

### ESTIMATION OF THE RELIABILITY OF INFERENCES FOR DIFFERENCE SCORES AND PROFILES

The error involved in interpreting a difference between two measures is larger than the error in each measure: More error is involved and the difference score itself has a separate meaning. The “difference-score” problem is widely known in the field of psychological and educational measurement

(AERA et al., 1999, Standard 2.3). The basic question is this: Given two measures for an individual, how reliable is the difference between the two scores? The reliability of a difference is related to (1) the reliability (or error) of each measure, (2) the intercorrelations between measures, and (3) possible normative difference between the measures.

As the average reliability of the two measures approaches the intercorrelation between the two measures, the reliability of the difference score ( $r_{DIFF}$ ) approaches zero:

$$r_{DIFF} = (4)$$

where  $r_{xx}$  and  $r_{yy}$  are the reliabilities of measures  $X$  and  $Y$ , respectively, and  $r_{xy}$  is the correlation between  $X$  and  $Y$ .

The problem of the reliability of differences has been discussed primarily in terms of learning disabilities (Salvia & Ysseldyke, [1/2( $r_{xx} + r_{yy}$ )] -  $r_{xy}$  \_\_\_

$$1 - r_{xy}$$

1998; Thorndike, 1963; Thorndike & Hagen, 1961), but it is also a serious problem in any profile interpretation (Barnett & Macmann, 1992b; Macmann & Barnett, 1985) in which the objective is to identify “severe discrepancies” among correlated variables, as discussed in the next section

### The Reliability of Profiles

Profile interpretation is the hallmark of many techniques, yet it raises a host of thorny issues because multiple difference scores are used to build personality profiles.

Profile interpretation depends on scale construction

employing appropriate methods and on validity studies, although neither may be sufficient to enable one to interpret individual profiles with great confidence. We define “profile interpretation” in two related ways: the meaning that can be attributed to the pattern of scores, and whether or not the pattern is related to a defined taxonomic group. A “taxonomic group” refers to “psychologically” similar individuals who occur with sufficient frequency in a population to permit ostensibly reliable predictions and thus meaningful comparisons (Wiggins, 1973). Three dimensions of profiles are important (Nunnally, 1978): (1) the level or elevation of scores, (2) the dispersion or scatter of scores, and (3) the shape or features of the profile.

The following difficulties are associated with profile interpretation in professional practice (Barnett & Zucker, 1990):

1. The reliability of subscales, the validity of subscales, and the reliability and validity of patterns are likely to vary within a scale. Some profiles or patterns may ultimately be useful; others may require further research or may lack validity support.
2. The reliability of the difference between two correlated scores is lower than the reliability of either score alone (see the preceding section on the reliability of differences). This problem is severely compounded when one considers the relationship among many correlated scores.
3. The stability of the profile may be unknown. Because of all of the factors described above, the overall pattern may be quite unstable. To the extent that there is

#### 12 I. GENERAL ISSUES

an important general factor in a scale, a common finding (Macmann & Barnett, 1999), peaks and valleys in the profile may just connote trivial differences.

4. As the number of comparisons increases, differences due to chance are more likely to occur. Furthermore, most comparisons are unplanned. Profile interpretations for individuals capitalize on chance occurrences of behavior.
5. Extreme scores are most often interpreted; these have the most error.
6. Subscales may have very different meanings (a) at different elevations, and (b) within unique or overall profile patterns.
7. Unusual or idiographic patterns will not be represented in taxonomies.

8. Support for the utility of profiles in intervention decisions badly needs critical empirical evaluation.

9. Profile “classification” must also be understood in terms of error rates.

One of the best known examples of profile development is that of Achenbach and Edelbrock (1983). They write what amounts to a prototypical description of professional practices of this genre: “Practitioners should adapt our materials to their own situations and integrate them with other types of data. . . . The essence of clinical creativity is to synthesize diverse and imperfect tools and data into practical solutions suited for each individual case” (p. 113).

However, the potential cumulative error rates for assessment inferences in this type of recommendation are not even estimable. Suffice it to say that the difficulties in personality measurement have not been redressed by profile development. Based on measurement principles, namely, the theoretical behavior of correlated variables (Macmann & Barnett, 1999), profiles basically will have unknown reliability when applied to individual children, except for the proviso that, to the extent that inferences derive from multiple difference score analyses, they will be much less than the reliabilities of the individual scales on which they are built.

More important than the determination of statistical significance of profile types, and their reliability, the psychological meaning of profiles must be determined. Profile interpretation has a long history within personality assessment; think, for example, of the extremely large number of interpretive profiles for the Minnesota Multiphasic Personality Inventory (MMPI). However, we argue here that the meaning of profiles or their contribution to assessment decisions is too often a moot issue because of the severe problems surrounding the *stability* of profile interpretations for individuals. If an individual’s profile is seen as “psychologically significant,” yet the profile is based on chance relationships among subtests for a single occasion, decisions cannot be made with confidence.

### **The Reliability of Projectives**

The term “projective” carries negative connotations for many, based on years of high hopes followed by disappointing research and subsequent reviews. Although the development of some projective techniques

has yielded reliability estimates that approximate those of so-called objective measures, and the interest in covert processes remains lively, many challenges to their use in professional practice still exist.

A review by Karon (1968) challenged the classical assumptions associated with the reliability of projective instruments. Karon argued that traditional notions of reliability have resulted in a "paradox": Reliability estimates associated with temporal consistency are not applicable because the experiences are expected to fluctuate, and similar arguments may be applied to internal-consistency estimates. Thus, Karon argued that the validity of projective techniques is not necessarily bound by reliability. Despite the fact that some reliability estimates (e.g., interrater reliabilities) reach satisfactory levels, the aforementioned problems associated with the reliability of the inference have not been addressed by Karon or others. Moreover, the use of the information concerning reliability for intervention design is a necessary and independent criterion yet it remains unavailable.

Although projectives have been widely criticized for their limited success in reliably assessing personality traits or structures, a gap still exists in the study of "private" thought processes and automatic (or in a different paradigm, unconscious) influences. Many constructs have been applied, such as depression or anxiety, as have terms such as "fantasies," "affect," and so on. Furthermore, there continues to be interest in subjective or private inner experiences and their function as potential causal factors, including motivational and emotional processes, from diverse sources (Bandura, 1986; Friman, Hayes, & Wilson, 1998; Mahoney, 1980; Meichenbaum & Gilmore, 1984).

The primary technical problems are verifications of inner experiences (Johnston & Pennypacker, 1993) and, ultimately, utility.

The potential contributions of private processes, including inner states not amenable to direct report, cannot be readily dismissed; to do so would be to deny the significance, as well as creative applications, of cognitive and affective theory (Bandura, 1986; Friman, et al., 1998).

Major advances are less likely to be based on new instruments that follow tradition than on procedural safeguards and research-

based methods for incorporating the information about private thoughts and those out of awareness into testable plans for professional practices. Mischel (1973) has described personality appraisal in a manner consistent with this direction and we review his recommendations in a later section.

#### **CONFIDENCE IN PROFESSIONAL PRACTICE: DECISION RELIABILITY**

Although addressed by the *Standards for Educational and Psychological Testing* (AERA et al., 1999), more attention should be given to the reliability of decisions. "Decision reliability" (e.g., the reliability of a psychological classification or diagnosis) is amenable to study and merits formal practical consideration (e.g., Cronbach & Gleser, 1965; Livingston, 1977). To examine the reliability of professional decisions, one can approximate the professional or natural decision process as closely as possible (e.g., the child fits a classification or does not; a targeted variable is deemed important or not, and the child would benefit from intervention or would not).

Studies of decision reliability for high incidence educational classifications and clinical profiles have produced alarming results (Barnett & Macmann, 1992a; Macmann & Barnett, 1999). As a principal finding, we have found three to five cases of disagreement for every case of agreement across many types of inferences (i.e., learning disabled or not learning disabled), using different samples, computer simulations, constructs, and methods. Although some of these studies have focused on scales not traditionally thought of as "personality measures" (i.e., achievement) or are marginally associated with personality (i.e., intelligence as a "trait" or personal competence), the patterns of findings apply because they are based on measurement theory and principles, not on specific constructs. Also, many researchers and professionals anchor personal competence to these constructs as some traditional sources have created momentum to infer personality-related constructs from intelligence profiles (e.g., Kaufman, 1979).

There are many reasons for these pessimistic decision reliability findings, and some are well traveled (i.e., approaches to norm development) or are discussed in other sections (i.e., base rate and selection rate,

extreme scores, and difference score reliability, depending on the decision being analyzed). The procedures for analyzing the reliability of decisions, however, are basic and fit practical choices. Alternate forms, raters, observational systems, and decisions implied by test outcomes can be compared (e.g., parent vs. teacher ratings and parent vs. clinician ratings) with the outcomes dichotomized (e.g., eligible vs. not eligible for services, internalizing disorder or not, and attention-deficit/hyperactivity disorder or not). For example, Ronka and Barnett (1986) investigated the reliability of decisions based on different adaptive behavior instruments and raters and found kappa coefficients ranging from .00 to .51. In another example, the kappa coefficients revealed in comparisons of parents' versus clinicians' ratings of profiles and of internalizing and externalizing syndromes ranged from .30 to .90 (Achenbach & Edelbrock, 1983). Decisions based on multifaceted evaluations can also be subjected to analysis. It is reasonable to require criterion levels for decision reliability that are equal to those for other areas of reliability (e.g., .90). We close our reliability section by simply pointing out that *reliability is a far more vexing problem for professional practice*

#### 14 I. GENERAL ISSUES

*than typically implied in test manuals and many texts.* The interpretation of reliability estimates reported in manuals generally is too optimistic: The error rates associated with inferences derived from scores are much higher than commonly discussed, even by many critics. Of course, reliability estimates limit validity inferences.

#### EVALUATING ASPECTS OF VALIDITY

As stated earlier, validity is a unified concept tied to judgments about the quality of decisions and the inferences made from test scores. Following Messick (1989, 1995), we outlined validity as related to a sequence of decisions (what is being measured; can you use the measure to meet some purpose; should you use it). Research results can only validate a specific use for a test, not the test itself (Nunnally, 1978). In this conception, there are many facets of validity that can and should be explored separately and simultaneously. In a unified view of validity, many of the methods that have traditionally been seen as different types of validity (construct, content, criterion, etc.) are best conceptualized

as related facets or considerations that define the meaning of a construct in both theory and practice. In Messick's model, this broader notion of multiple facets of construct evidence is the first of several necessary decisions about overall validity. Any facet of validity can be examined conceptually, logically, and through experimental procedures. There is no consensus on the methods that are used to validate inferences derived from tests. Tests and test scores cannot have "high" or "low" validity in this general sense because it is the decisions and inferences based on scores that must be evaluated. General guidelines and principles for assessing various validity facets have, however, been developed and are easily incorporated into a unified validity concept. To be meaningful, validity information must be qualified and scrutinized through consideration of some specific purpose, test standardization procedures and samples, and the actual experimental validation procedures used by the test developers.

Personality scales were intended as a short-cut method, to save time and money in contrast to other intensive methods (Peterson, 1968). However, in considering major scales, one often finds that test development has occurred over decades through the efforts of many—a process clearly revealed through the history of any major personality measure (e.g., Schinka & Greene, 1997). It is an expensive endeavor (Burisch, 1984). Validation efforts are conducted not only by test developers but subsequently, even more important, by independent researchers. Therefore, validity evidence is subsumed by the body of research that follows the publication of a scale. However, Cronbach and Meehl (1955) pointed out that there is no one "scientific approach" that can completely legitimize a construct in a manner that would rule out scientific disputes and varying viewpoints.

We discuss here the traditional validities called construct, criterion-related validity, and content as methods used to validate constructs within the unified validity model. Validating the meaning of a constructed measure that is used to make an inference relates to the first question we used to define unified validity—"What is the meaning of the construct being measured?" In terms of the second question ("Can you use a test score to make an inference to guide some

decision?”), we discuss several facets that have typically been called incremental validity and criterion-related validity. In addressing the issue further, we extend our discussion of the relationship of reliability to unified validity. Given the criteria of confidence and helpfulness, we conclude with the ultimate defining quality of validity—the quality of client outcomes resulting from assessment (the facet of consequential validity) and the relationship to the practice of personality measurement. Unfortunately, the incorporation of the criteria of meaningful outcomes resulting from assessment decisions has been well developed and examined only within behavioral models (i.e., Baer, Wolf, & Risley, 1968; Wolf, 1978), not within traditional uses of personality assessment.

### The Facets of Constructs

#### Nomological Network as the Foundation

Constructs are central to much of scale development and use. Test developers should

1. Personality Assessment Research 15  
place the construct measured by a scale within a conceptual framework that defines the meaning of the construct and its relationships to other constructs and observable behaviors or variables—a process referred to as the construction of a “nomological network” (Cronbach & Meehl, 1955; see also Messick, 1989). This network should explicitly translate theoretical constructs into verifiable relationships among operations. It is certainly dynamic during instrument development but should be coherent and well developed at the point of making an instrument available for clinical use. We believe that the nomological network is a critical construction if practitioners are to adequately use personality measures. A clearly described nomological network would allow developers to organize and present their data (to potential users) related to verifying the operational meaning of the constructs measured by an instrument and supporting a set of uses. Unfortunately, test manuals seldom, if ever, explicitly offer a proposed network. Furthermore, data collected during scale development are often presented in a fractured manner, related to separate validity concepts, and not clearly connected to meaningful inferences or to specific uses. Although there are many statistical approaches to scale construction that address aspects of validity (Schinka & Greene, 1997), we present two well-articulated

methods typically used in construct validation studies that could easily fit within an explicit nomological network: the multitrait–multimethod matrix and factor analysis.

**The Multitrait–Multimethod Correlation Matrix**

Kenny (1995) wrote: “The MTMM matrix represents one of the most important discoveries in the social and behavioral sciences” (p. 123). A logical approach that stems from a classic paper by Campbell and Fiske (1959), the “multitrait-multimethod correlation matrix” (MTMM) can be employed whenever there are at least two constructs measured by two methods. An MTMM analysis includes two features of validity. First, it examines “convergent validity.” Logically, the correlations between the methods intended to measure the same trait should “converge” or be higher than those developed to measure different traits. “Discriminant validity” provides an indication of the predicted divergence of the traits. Dissimilar traits should have lower intercorrelations than those found for similar traits. Each measurement procedure yielding a score is considered to be a “trait-method” unit in order to evaluate the possible contributions of various methods used to estimate the trait. “Method bias” is indicated when correlations between different traits (measured by the same method) are higher or similar to the correlations between different traits (measured by different methods). Correlations between supposedly related traits may be substantial because of shared method variance (e.g., self-report methods), although ideally the correlation between two traits should not be a function of the methods. In contrast, the use of distinctly different methods (e.g., observations and behavior ratings) may attenuate (or minimize) estimates of the same trait (Campbell & O’Connell, 1982).

Validity coefficients are represented by correlations across methods for the same trait. To demonstrate evidence for constructs, same-trait or validity coefficients must be higher than correlations between different traits. A number of alternative statistical approaches to MTMM analysis have been proposed (Kenny, 1995; Millsap, 1995; Wothke, 1995), but none exactly match the rationale and logic of MTMM analyses. Separating trait from method remains challenging (Millsap, 1995).

The MTMM analysis is quite appropriate

to professional practice questions associated with the interpretation of psychological constructs. Given two (or more) methods of measuring various personality constructs (i.e., temperament, children's self-concepts, or anxiety), a basic question pertains to the decisions that would be made depending on the various alternative assessment methods. Typically, the MTMM analysis has revealed a number of problems when employed (e.g., Macmann & Barnett, 1984). The validity coefficients, even when significant, may be low or moderate, and logical or expected relationships may not be upheld. The effects of undesirable or unpredicted method similarities and differences may be pronounced.

Too frequently, scale developers have tried to measure constructs that do not differ

#### 16 I. GENERAL ISSUES

much from other constructs (Kenny, 1995).

Practically speaking, the ambiguity of interpretation that is revealed through MTMM analyses is considerable (Poth & Barnett, 1988).

In Fiske's (1982) review of the frequent disappointments revealed by the MTMM procedure, he pointed out that convergentdiscriminant validation represented "*modest* criteria" (p. 90; emphasis added). The ultimate goal is convergence not only across different methods but also across different research strategies. Campbell and O'Connell (1982) argued that the MTMM matrix may overemphasize problems with traits and that measurement methods require the same scrutiny as traits when one is analyzing the disappointing results of construct validation.

#### Factor Analysis

"Factor analysis" represents a family of techniques that have as the common goal the simplification of complex correlation matrices, or interrelationships between items (or behaviors), in order to reveal the major dimensions that underlie a set of items. Factor analysis can also be used to test theory and clarify the meaning of personality constructs. The logic is easy to comprehend, but the use of the techniques requires a great deal of sophistication. Furthermore, the statistical approaches that are used rely on judgments and assumptions made in carrying out the necessary steps, and ultimately in interpreting and giving meaning to the results.

A "factor" is a construct that seems to

best represent the structure or relationship between variables or clusters of variables. Factors are named by the researcher or scale developer through a subjective labeling or judgment process. "Factor analysis is only a prelude to more systematic investigations of the constructs" (Nunnally, 1978, p. 330). Despite the vast array of techniques, there is increasing agreement on the basic procedures. Stevens (1996) recommends the use of principal-components analysis to first describe and "enumerate" the underlying constructs for a set of variables—a procedure that requires fewer assumptions but often yields results similar to other statistical procedures. Confirmatory methods such as structural equation modeling (e.g., Crowley & Fan, 1997; Judd, Jessor, & Donovan, 1986), may be used to evaluate the degree of correspondence between the obtained factor structure and theory as can other factor-analytic methods.

There are two major areas in which the use of factor analysis is relevant to the topics subsumed in the present chapter: research in personality and research in psychopathology. An overarching difficulty is the variation in terms (defined by the results of factor analysis with different samples, items, methods, and researchers) that have been used to describe similar behaviors (Sells & Murphy, 1984).

One of the best examples of factor analysis for the purpose of personality description stems from decades of research by Cattell (e.g., Cattell, 1982; Dielman & Barton, 1983). In numerous studies, Cattell has factor-analyzed self-report inventories, biographical inventories, and observations. The work is based on Allport and Odbert's (1936) research on basic descriptive adjectives that define human characteristics. Cattell has typically discussed 16 source traits of normal personality for adults and 14 for children. However, other researchers stress far fewer factors (e.g., Costa & McCrea, 1986; Goldberg, 1995), which they think describe fundamental features of personality in a reasonable manner. Although there are important sources of agreement, the descriptive research has been tangential to professional practices associated with the diagnosis or classification of psychopathology, and especially to research concerning intervention design.

The study of child psychopathology

through the use of factor-analytic procedures is exemplified by the work of Achenbach and Edelbrock (e.g., Achenbach, 1985; Achenbach & Edelbrock, 1978, 1983; Achenbach & McConaughy, 1997). (For other similar lines of research, see Reynolds & Kamphaus, 1992; Wirt, Lachar, Klinedinst, & Seat, 1984.) In a comprehensive review, Achenbach and Edelbrock (1978) examined the literature for empirical syndromes used in classifying childhood psychopathology. The studies included ratings by mental health workers, teachers, and parents. Syndromes with at least five items and with “factor loadings or average inter-  
1. Personality Assessment Research 17  
correlations of .30 or higher” were included in the analysis (p. 1284). They found four “broad-band” syndromes that were replicated across studies, or that had “counterparts” in at least two studies. Sex differences were found to be important, with the findings most often based on boys. Empirically based syndromes found in two or more studies included the following: Overcontrolled, Undercontrolled, Pathological Detachment, and Learning Problems. In addition, 14 “narrow-band” syndromes were identified (e.g., Aggressive, Depressed, Hyperactive, and Anxious).

Achenbach and Edelbrock (1983) developed the Child Behavior Checklist (CBCL) and the associated Child Behavior Profile to provide “standardized descriptions” of problems and social competencies of children. A principal-components analysis with a sample of clinically referred children ( $n = 2,300$ ) yielded Internalizing and Externalizing syndromes, and narrow-band scales varying by age and sex (Depressed, Uncommunicative, Social Withdrawal, etc.). This methodology has also been extended to the analysis of teachers’ ratings and adolescent self-reports, as well as observations, into a complex assessment system (Achenbach & McConaughy, 1997).

From this brief review, a number of professional practice questions concerning factor analysis become evident. The first involves determining the primary purpose of the factor analysis. Some purposes involve basic research and are tangential to professional practice issues. Second, the technical adequacy of the procedures should be evaluated. Whether the factor structure is stable and replicable is partly determined by the

adequacy of the measure, but also by sample size and other sample characteristics. Evidence for cross-validation should be presented (replications on an independent sample). Third, the importance of each factor should be considered, estimated by the percentage of variance explained by the factor solution. Fourth, evidence to support various interpretations of the factors should be presented. This would include MTMM logic to consider the adequacy of the nomological network and relationship between similar and dissimilar constructs and measures. MTMM analyses are also important because most factor-analytic solutions are based on rating scale techniques, but intervention design and evaluation are founded on direct observation. Finally, because scales derived from factor analysis often are used to derive profiles during assessment, the serious problems of profile reliability (see earlier) may severely limit their contributions to reliable decisions about children. What may appear to describe meaningful groupings of childhood problems based on research with groups may not yield information that provides reliable information in making decisions about individuals. Basic research in childhood psychopathology is clearly still needed, and empirical research will permit a better study of etiology and of the relationship among differentiable traits, empirical syndromes, or constellations of behaviors and well-developed interventions (Hibbs & Jensen, 1996; Kazdin, 1985a, 1985b). However, important factors have been traditionally omitted from factor-analytic solutions: "the stability or predictability of dysfunction" (Kazdin, 1985a, p. 38) and contextual appraisals of dysfunction. These are the factors that lead to confidence and helpfulness in decisions. In summary, factor solutions that result in glee for researchers may imply headaches for practitioners when the magnitude of support for various interpretations is translated into actual professional practice decisions and corresponding error rates. Although factor-analytic procedures represent an important contribution to childhood personality and psychopathology, applications to intervention design have not been directly addressed (see Haynes, 1986).

**Criterion-Related Validity Methods within a Unified Validity Model**

Traditionally, “criterion-related validity” has focused attention on the degree to which a test predicts behavior (or classification status) on an independent criterion.

The term “concurrent validity” is used to describe studies involving the child’s status at the time of testing (e.g., a diagnosis), whereas the term “predictive validity” is used to examine the relationship between a score and future status (e.g., temperament and later adjustment). “Postdictive studies” examine the relationship between a score

#### 18 I. GENERAL ISSUES

and past status on a criterion. Interventionists apply other criteria to outcome measurement (Komaki, 1998; Wolf, 1978).

Criterion-related validity thus involves predictions between one set of variables (the predictor) and a second set (the criterion).

We believe that in many cases, these relationships are most appropriately considered within the nomological network that gives operational meaning to a construct. It may even be difficult to distinguish methods subsumed under multimethod–multitrait procedures from criterion-related methods. However, there are important situations when methods associated with criterion-related validity facets have been used to evaluate some actual use of a test score.

Correlations typically used to estimate the association between variables have been termed “validity coefficients,” although within our analyses this may be viewed as an overstatement. In addition to estimation of the correspondence between variables, the correlation coefficient permits an estimation of the standard error of prediction.

The Pearson correlation, perhaps most often used, indicates the strength of a linear relationship. When squared, it reveals the proportion of accountable variation between two variables, or the variance that may be reliably predicted (Wiggins, 1973).

Factors that limit validity coefficients are the same as those that influence correlations: whether the measures have a linear relationship, are normally distributed, and have equal variances throughout the range of scores (homoscedasticity). Furthermore, sample characteristics can limit (or exaggerate) the size of the correlations (Nunnally, 1978).

The professional practice issues in studying criterion-related validity facets are formidable, but criterion-related validity lies at

the heart of many professional practice questions. A fundamental issue may be referred to as the "criterion problem." Wiggins's (1973) warning still applies: "Criterion analysis has proved to be the most recondite and vexing issue confronting personality assessment" (p. 39). The reasons are obvious. Many of the inherent problems in scale construction have counterparts in criterion construction. The criterion itself must be evaluated in terms of its reliability and validity. Frequently, personality scale developers have used similar rating scales for concurrent validity studies, although these fall into our analysis as being related to construct evidence. Even though this practice confounds shared method with validity and promotes circularity, results are often modest. For example, correlations of internalizing scores for two different rating scales on internalizing or externalizing disorders are likely to range from .3 to .7. Criteria often include such ambiguous and ill-defined behaviors as "adjustment," "emotional disturbance," or "risk" measured at one point in time. Some relationships have proven to be limited because of moderating variables apart from the predictor or criterion measure. For example, fantasy or other indirect indications of "aggressiveness" may not predict overt aggressive behaviors because of social sanctions or inner controls, not because they do not exist or because they are trivial. Moreover, forms of aggression may be varied and subtle in their expression. Because of these and other theoretical factors, modest correlations are often expected.

#### Decision versus Criterion-Related Validity

The criterion-related validity "problem" can positively be refocused to that of "decision accuracy and validity," or the analysis of correct and incorrect decisions. This refocusing clearly is a better fit for a unified model of validity that ultimately hinges on the outcomes of assessment for clients. Decision theory is concerned not only with accuracy but also with the utility of decisions, and with values that can be tied to different outcomes. The decisions with respect to personal and social assessment are typically in the form of diagnostic systems: whether groups or individuals differ in measurable ways, and whether the groups or individuals should receive special "treatments." The prototypical study of decision accuracy is

depicted in a  $2 \times 2$  matrix whereby valid positives are successfully screened, with high scores on both the screening measure and criterion; valid negatives are also successfully screened, with low scores on both measures; false positives have high scores on the screening measure but do not “succeed” according to the criterion measure; and false negatives have low scores on the screening

1. Personality Assessment Research 19  
measure but do “succeed” according to the criterion measure. “Cutting scores” on the test or predictor are used to guide decisions. The empirical basis for cutting scores should be clearly specified (AERA et al., 1999). As the cutting score is “moved” either higher or lower by the investigator, the type of errors change. Classic sources for decision theory include Cronbach and Gleser (1965), Cronbach and colleagues (1972), Meehl and Rosen (1955), Mischel (1968), and Wiggins (1973). Contemporary approaches to decision theory emphasize attempts to come closer to the natural process (i.e., context specific setting and organizational factors such as teaming and multiple goals, feedback loops, and realistic or high stakes) (Zsombok & Klein, 1997). Professionals’ situational awareness based on problem structuring is a key construct.

#### **Additional Facets of Unified Validity Incremental Validity**

The correlations between a test and a criterion measure applied in the analysis of criterion-related validity can be misleading to a large degree. Other factors have to be taken into account. There should be clear benefits for using the targeted measure over alternative procedures in terms of cost, intrusiveness, acceptability, concordance with intervention design, and other factors. The outcomes of actual decisions based on a comparison of procedures should be assessed. Introduced earlier, “incremental validity” (Sechrest, 1963) suggests the need to evaluate the contributions that a test-based inference makes to a decision, over and above information that is already known or readily available. Incremental validity is concerned with the improvements in assessment decisions that occur by the use of a procedure, compared to those that can be made without it. The question is critical to personal and social assessment. For example, after interviewing parents, teachers, and child, and conducting observations across settings,

the professional psychologist would ask: “What nonredundant information will be gained by adding other procedures?” and “Will adding these procedures help my client?”

Among a number of factors, the incremental validity of an assessment procedure depends on base rates for the problem to be identified, evaluated within an appropriate population. Mischel (1968) offered the following example: “If 95 percent of the inpatients of a particular hospital are diagnosed ‘schizophrenic,’ a test-derived statement that predicts this label correctly 80 percent of the time is less useful than systematically calling every patient in the hospital schizophrenic” (p. 106). Thus, base rates limit the association that can be demonstrated between a predictor and criterion. Alternatively, modest validity coefficients may be useful in certain situations, because the “usefulness” depends on information that is actually gained through a measurement procedure.

#### **Content Validity**

Content validity is concerned with the adequacy of the items and is based on the systematic examination of content, selected from an identifiable “universe” of content. Typically, content validity has been applied to academic tests, or tests with similar properties, because of the difficulties of sampling items from a universe or domain related to personality or social constructs. When the point of emphasis is shifted to the functional analysis of behavior, however, principles associated with content validity can have considerable relevance for personal and social assessment (see Linehan, 1980).

#### **Subjective Validity**

Although “subjective validity” is not a traditional topic in measurement theory, the personal or subjective elements of test usage are considerable. There are many lines of reasoning necessary for the consideration of subjective validities, but most point toward an increasing ecological perspective (Cronbach, 1984, p. 571). A judgment that assessment leads to better outcomes for clients innately involves values and subjective viewpoints. Within a unified model of validity, these subjective decisions must be made and clearly stated.

Messick (1980, 1989) pointed out that tests are evaluated on the basis of measurement characteristics, whereas test applica-

## 20 I. GENERAL ISSUES

tions require evaluation through potential social outcomes that include both normative and subjective comparisons. Social outcomes cannot be studied without reference to actual behaviors in real-life settings, followed over long periods. In assessment related to intervention design, social validity (Kazdin, 1977; Wolf, 1978) has been described through the identification of (1) socially significant problems for behavior change, and (2) socially acceptable methods with respect to the immediate social community. The results of the intervention should have practical significance (Baer et al., 1968) and should be evaluated over time (Schwartz & Baer, 1991) and in multiple ways (Willems, 1977). As broad and illdefined as they appear in contrast to other measurement topics, subjective validities are necessary and basic considerations in the assessment-intervention process.

### Modern Test Theory

“Modern test theory” involves the latent-trait approach to test development (e.g., Lord, 1980). The procedure involves the study of the relationship between individual items and an underlying (thus latent) trait through mathematical models. Two major aspects of latent-trait models include (1) the assumption of a unidimensional trait with homogeneous items, and (2) mathematical functions that depict the relationship between item choices and the trait or ability measured (Ghiselli et al., 1981). A “latent attribute” refers to “the hypothesized but unobserved characteristics that account for a particular set of consistencies within and differences among persons” (Thorndike, 1982, p. 5). The mathematical functions are expressed as item-characteristic curves. In test design, items are selected that satisfactorily are related to the latent attribute and represent different levels of the trait. Thorndike (1982; see also Wainer & Braun, 1988) presents an introductory but comprehensive description of the latent-trait approach to test construction.

Although modern test theory has been used increasingly for test construction, the potential issues on a practical level remain unknown. Psychometrically, the theory has attractive features. Items developed through latent-trait procedures can be thought of as “calibrated,” presumed to be a major advantage. Nunnally (1978) points out that although

the theory is termed “modern,” its basic principles were developed in the 1950s, with the ideas present long before that period. We believe that if there are implications for personality research, they likely lie within the facet of validating personality constructs, but the potential use of the constructs for individual decisions must be addressed independently.

## CONTINUING THEMES IN PERSONALITY MEASUREMENT

The purpose of this section is to review major themes in personality measurement that have created dilemmas for both researchers and professional psychologists. In the final section, we turn to component strategies for sound professional practices associated with personal and social assessment.

### Idiographic or Nomothetic?

Unique life experiences combined with inherited characteristics lead to “idiographic” patterns of behavior. “Nomothetic” pertains to general lawfulness of behavior. Although some view the relationship between the two as conflicted (Nunnally, 1978), the differences (as in many “dilemmas” in psychology) often define various goals in research and professional practices, and not fundamental conflicts. Allport (1937) argued for the importance of both approaches. Although “nomothetic” defines basic goals of behavioral research, there has been a continued interest in idiographic approaches and their relationships to general laws. Furthermore, it is also viewed as a “false dichotomy” (McFall & McDonel, 1986). All assessments involve idiographic and nomothetic objectives and procedures.

### The Idiographic Approach

An interest in the in-depth understanding of individuals is shared by psychodynamic, phenomenological, behavioral, and cognitive-social learning theorists. Mischel (1968, 1973, 1984a, 1984b) has developed the most coherent description of the idiographic approach to the study of personality from a social and cognitive viewpoint, as well as of its relationship to nomothetic principles.

Mischel (1984b) wrote:

Deeply impressed by George Kelly’s (1955) thinking, I was sensitive to the fact that clients-like other people don’t describe themselves with operational definitions. They invoke motives, traits, and other dispositions as ways of describing and explaining their experiences

and themselves. Much of the assessor's task . . . should be to help people in the search for such referents for their own personal constructs, instead of forcing the assessor's favorite dispositional labels on them. (p. 280)

Mischel (1973, p. 265) suggested five "person" variables that serve as guidelines for assessment-intervention design: (1) the individual's competencies to construct (generate) diverse behaviors under appropriate conditions, (2) the encoding or categorization of events, (3) expectancies about outcomes, (4) the subjective values of different outcomes, and (5) self-regulatory systems and plans.

### The Nomothetic Approach

Two distinct lines of research can be identified within the nomothetic approach: those that have been concerned with the identification of traits (e.g., Costa & McCrea, 1986), and those that have been concerned with lawful behavioral principles (e.g., Cone, 1986; Mischel, 1968). We briefly describe the "trait" approach here. The applications from behavioral and from social and cognitive-behavioral approaches are integrated into the suggested components discussed in the final section of this chapter.

From Allport's early work, Stagner (1984) defines a trait in the following manner:

"A trait is a consistent and persistent pattern of behavior and experience (cognitive and affective) characteristic of a particular individual" (p. 7). Theoretically, traits may be unique to the individual, may be shared with others, or may be common. They are hierarchically structured, and function to control lower-level processes in behavior. Thus, a trait is associated with various probabilities of occurrences of specific behaviors or classes of behaviors in situations. "Trait-situation interactions are considered to be the rule" (p. 34).

Although there are far too many traits (or factors) to review, convergence has been a recent goal of researchers (Costa & McCrea, 1986; Sells & Murphy, 1984). Reviewed briefly under the topic of factor analysis, a well-established Big Five taxonomy similar to one proposed a number of years ago by several researchers (e.g., Fiske, 1949; Norman, 1963; Tupes & Christal, 1961) is presented by Costa and McCrea (1986, p. 410; see also Costa & Widiger, 1994): Neuroticism (e.g., Calm-Worrying), Extraversion (e.g., Reserved-Affectionate),

Openness (e.g., Down-to-Earth–Imaginative), Agreeableness (e.g., Ruthless–Softhearted), and Conscientiousness (e.g., Negligent–Conscientious). It is ironic that two very different Big Fives in personality measurement have stood the test of time (i.e., Mischel’s).

Potential effects of a better understanding of traits include (1) more adequate research in personal and social development, (2) a better understanding of expressions and meanings of various symptoms, and (3) further insights into mechanisms of stability and change. Costa and McCrea (1986) argue for stability in personality development through adulthood, and for the proposition that “human nature is by no means easily changed” (p. 420). This “proposition” has been well analyzed (e.g., Lewis, 1997; Moss & Sussman, 1980), and although no one would disagree with the difficulties of change, others would argue that the complexities of intervention design and execution have hampered unequivocal statements regarding possibilities for change.

#### Interviews, Observations, or Ratings?

##### Interviews?

Many regard the interview as the most important “instrument” in assessment. An extensive and diverse literature on interviews stresses anything from communication skills, unstructured or structured clinical diagnosis, private experiences, and intervention roles, to therapeutic approaches. The topic has also been addressed with respect to interviews directed to children and to all others associated with the assessment-intervention process; these vary widely by function (Barnett & Zucker, 1990). Many ques-

##### 22 I. GENERAL ISSUES

tions persist concerning the technical adequacy and use of interview procedures (Gresham, 1984). Despite problems in describing the interview in a way that would satisfy all uses, it may be best to state that it is integral to assessment and intervention and all stages therein. All sources of assessment error are potentially relevant as well.

Peterson (1968) described a prototypical behavioral interview. He pointed out that the distinguishing feature of the interview is the relationship between interviewer and interviewee.

Without the relationship, the assessment–intervention process is impaired and ultimately terminated. The future of the interview with respect to studies of

technical adequacy rests with the study of the interview's interrelated but specialized functions: (1) as a technique of behavior analysis (e.g., Kanfer & Grimm, 1977; Peterson, 1968), (2) as a formal diagnostic tool, (3) as a method to study private experiences, and (4) as a data-gathering technique necessary to the assessment-intervention process.

### **Ratings or Observations?**

As a reaction against the use of both projective and objective techniques for the study of personality structures and processes, there has been renewed interest in behavior rating scales and direct observations. They are not alternative ways of assessing the same information, nor are they complements of each other; they have different objectives (Cairns & Green, 1979). Rating scales have been developed to assess a broad range of factors: personality, psychopathology, problem behavior, and social emotional functioning (Edelbrock, 1983).

"Direct observations . . . can be the key for identifying how actual behaviors are elicited, maintained, and organized" (Cairns & Green, 1979, p. 222). Observations are essential for the analysis of interactions (Bakeman & Gottman, 1986). Cairns and Green (1979) have examined the issue in depth. In their words,

The distinguishing characteristic of rating scales is that they involve a social judgment on the part of the observer, "or rater," with regard to the placement of an individual on some psychological dimension. (p. 210)

The distinguishing property of behavior observations is that they involve an attempt to record the actual activities of children as opposed to offering a judgment about children's personal dispositions or the quality of their relationships. (p. 213)

### **Professional Practice Issues**

In summarizing the topics presented in this section, one problem stands out: The use of all the techniques described earlier must be guided by a template that defines decisions related to use, including interpretations and outcomes. "Professional judgment" is the term used to bridge the gap between the knowledge base and actual decisions concerning the use of various techniques (Barnett, 1988). Professional judgment itself involves a host of unknowns and depends on a personal model of professional practice. Unfortunately, professional judgment has not served well as a template, because the

decision process has not been successfully explicated and has not been amenable to study. Although concerns have been voiced most often about projectives, the same arguments apply in varying degrees to all techniques, including interviews, observations, “objectives,” and behavior rating scales.

### **CONFIDENCE AND HELPFULNESS: STRATEGIES FOR PERSONAL AND SOCIAL ASSESSMENTS**

Throughout the previous sections, many questions have been raised concerning personality measurement and professional practice issues. The final section addresses two underlying themes: that test or scale interpretation would be best guided by the criteria of confidence and helpfulness. We think that these are addressed by applying appropriate theory related to behavioral change, and by analyzing problem-solving procedures that serve as the basis for professional practices.

#### **What Can Be Said with Confidence?**

##### **Applying a Coherent Theoretical Model**

Which factors guide decisions in ambiguous and complex real-life assessment situations?

Decisions are guided by theory that may be

1. Personality Assessment Research 23  
deliberate and formal, or may be internalized and personalized adaptations of a venerable theory (e.g., Freudian theory and social-cognitive behaviorism), or may be eclectic or ad hoc. To the degree that factors that guide the assessment of personality remain unspecified, the process cannot be studied. Within any endeavor related to the assessment of personal and social functioning from the interpretations of projectives, objectives, rating scales, interviews, and observations, the potential exists for significant judgmental differences that can lead to idiosyncratic outcomes for clients.

Many personality “theories” remain viable in the literature or in the minds of professional psychologists (see Loevinger, 1987). For the purposes of this chapter, they warrant appraisal to the extent that they guide actual professional practice decisions.

Bandura’s (1986) criterion is pertinent:

“The value of a theory is ultimately judged by its usefulness as evidenced by the power of the methods it yields to effect psychological changes” (p. 4). Theories of behavioral change are different from trait theories in that different variables are considered.

Rather than resolving theoretical debate, a

more promising tactic is clarifying technical adequacy characteristics for the key variables used to guide intervention plans.

### Applying Psychometric Concepts to Professional Practices

As discussed in prior sections and other work (i.e., Barnett, Lentz, & Macmann, 2000; Macmann & Barnett, 1999), professionals relying on test or scale use to improve professional judgments, defended by traditional information about traditional measurement quality but without regard to decision research, problem context, or consequence, may have unwarranted confidence in decisions. The data on psychometric quality presented in test manuals does not sufficiently fit the real-world decision contexts faced by professionals, parents, and children, nor do reliability and validity data collected during instrument development adequately generalize to questions of problem identification or classification, or problem solving.

The *intent* of psychometric methods to improve decision making still must be met, but in ways that lead to confidence in understanding problem context and decision consequence. Practitioners must consider the psychometric qualities of *decisions* across the sequence of problem solving. Technical adequacy questions may be addressed by directly collecting ongoing data on *both* problem variables and the quality of those measures in the natural problem settings that constitute a referral. In other words, behaviors, other key variables (environmental events and interventions), and their technical adequacy can be all measured or sampled in phases of problem solving.

The fundamentals of psychometric quality should have to do with measuring (1) meaningful child-related variables, (2) within specific educational or other natural contexts, and (3) for the purpose of guiding important professional decisions. The many complexities, challenges, and criticism are reviewed in Nelson and Hayes (1986) as well as other sources.

Model mixes, such as adding traditional personality or behavior rating techniques to direct assessments of problem situations, which also may be described as a professional eclecticism, are not panaceas; they may lead to heightened sense of professional confidence at the same time that error rates may be increased, decision processes may be

obfuscated, and time and resources may be depleted. The dilemmas of model mixes or nonsystematic eclecticism have been examined often (Haynes & Uchigakiuchi, 1993; Williams & Thompson, 1993).

In summary, to achieve confidence, the study of technical adequacy of assessment for specific professional practices is needed; this would be the technical adequacy of problem solving in natural contexts (Barnett et al., 2000; Macmann et al., 1996).

#### What Can Be Said That Is Helpful?

An aspect of professional judgment is to determine the correspondence between personality constructs (or syndromes, traits) that appear to apply to the client and those that have an empirical foundation. Based on the characteristics of one's client, logical generalizations can be made from researched interventions with children having similar characteristics. Many personality variables suggested by idiographic tech-

#### 24 I. GENERAL ISSUES

niques are widely researched (e.g., anxiety, depression, and aggression), and logical comparisons can be made between the idiographic client-based understanding of the construct, or how the client sees him- or herself, and its researched dimensions, the most likely nomothetic link for professional practice but one that does not necessarily meet the criterion of helpfulness. In other words, the client leads the way with regard to personality description (Mischel, 1973), but helpfulness is judged by outcomes. The "criterion" with respect to the adequacy of plans is that of resulting changes in behaviors. Intervention design typically involves a process of *sequential decisions* (Bandura, 1969) in contrast to diagnostic or classification decisions: "By retaining flexibility in the selection, sequencing, and timing of objectives, the treatment program remains highly sensitive to feedback from resultant changes. . . . Successful treatment . . . requires the selection and attainment of a variety of specific objectives rather than single omnibus outcome" (pp. 103–104).

Time-series methods associated with single-case experimental designs can be used flexibly to analyze the effects of plans with a wide range of problem behaviors, interventions, and even "philosophies" (Barlow, Hayes, & Nelson, 1984). Replications are critical to the process (see Kazdin, Kratochwill, & VandenBos, 1986). Hayes and

colleagues (1986) discuss specific research strategies that enable the analysis of the effects of assessment information on intervention decisions.

### Consequential Validity

The major theme from our analysis is that psychometric issues are best examined by focusing primarily on the effectiveness of context-specific decisions that are made during problem-solving processes, and with an ultimate regard to the *consequences* of these decisions (Messick, 1995). Validity is a unitary concept that is studied within a specific decision context. The validity of an assessment procedure is ultimately judged by whether it accomplishes some clear purpose, *and* whether *consequences* for those assessed are positive. He argues that there is a sequence of decisions that needs be made about whether any metric should be used to help make a specified decision, a sequence that begins with demonstration that a construct to be assessed is meaningful. The characteristic of a meaningful, interpretable *construct* organizes and subsumes the various types of validities that have been discussed in the literature

Once construct meaning is demonstrated within Messick's model, evidence supporting specific use for measuring the construct is needed. This is the point at which traditional psychometrics have typically stopped. However, Messick strongly advocates that there are values and consequences associated with assessment that are the final arbiters of whether assessment data *should* be used to make some decision. Thus, a construct that may be measured with meaningfulness, and useful for some purpose, may not have ultimate validity if consequences of use are detrimental.

### CONCLUSION: NEEDED RESEARCH FOR PROFESSIONAL PRACTICES

Long sought after, a major goal of personality research has been to meet the criteria of helpfulness: The hallmark has been the identification of traits, with the hope of identifying strong trait  $\times$  treatment interactions. The difficulties are numerous: (1) limited consensus on the classification and measurement of traits; (2) limited consensus on intervention design; (3) treatment integrity issues; (4) experimental design issues, including the development of adequately sized groups for relatively rare syndromes, adequate control groups, and

random assignment to groups; and (5) unique moderating effects for individuals and subgroups even within defined taxonomic groups or classifications. This has been a worthy goal, but it has been constrained by another highly significant problem: the inability to use personality appraisal techniques at appropriate levels of professional confidence. Furthermore, assessment practices need to encompass methods of assessing the potential impact of life events that alter developmental trajectories, including those that are accidental. New research is needed to address the criteria of confidence and helpfulness when

1. Personality Assessment Research 25

methods are suggested for personality and social appraisals. This would be the psychometry of professional practices based on problem solving.

## REFERENCES

- Achenbach, T. M. (1985). *Assessment and taxonomy of child and adolescent psychopathology*. Beverly Hills, CA: Sage.
- Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical methods. *Psychological Bulletin*, 85, 1275–1301.
- Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., & McConaughy, S. H. (1997). *Empirically based assessment of child and adolescent psychopathology* (2nd ed.). Thousand Oaks, CA: Sage.
- Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Wiley.
- Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycholexical study. *Psychological Monographs*, 47(1, Whole No. 211).
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.
- Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Baer, D. M., Wolf M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91–97.
- Bakeman, R., & Gottman, J. M. (1986). *Observing interaction: An introduction to sequential analysis*. New York: Cambridge University Press.
- Bandura, A. (1969). *Principles of behavior modification*. New York: Holt, Rinehart & Winston.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The scientist-practitioner: Research and accountability in clinical and educational settings*. New York: Pergamon Press.

- Barnett, D. W. (1988). Professional judgment: A critical appraisal. *School Psychology Review, 17*, 656–670.
- Barnett, D. W., Lentz, F. E., Jr., & Macmann, G. M. (2000). Psychometric qualities of professional practice. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 355–386). New York: Guilford Press.
- Barnett, D. W., & Macmann, G. (1992a). Decision reliability and validity: Contributions and limitations of alternative assessment strategies. *Journal of Special Education, 25*, 431–452.
- Barnett, D. W., & Macmann, G. (1992b). Discrepancy score analysis: Accuracy in analysis misdirected. *School Psychology Review, 21*, 494–508.
- Barnett, D. W., & Zucker, K. B. (1990). *The personal and social assessment of children: Current status and professional practice issues*. Boston: Allyn & Bacon.
- Barrios, B., & Hartmann, D. P. (1986). The contributions of traditional assessment: Concepts, issues, and methodologies. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 81–110). New York: Guilford Press.
- Block, J. H., & Block, J. (1980). The role of ego-control and ego-resiliency in the organization of behavior. In W. A. Collins (Ed.), *Development of cognition, affect, and social relations* (pp. 39–101). Hillsdale, NJ: Erlbaum.
- Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39*, 214–227.
- Cairns, R. B., & Green, J. A. (1979). How to assess personality and social patterns: Observations or ratings? In R. B. Cairns (Ed.), *The analysis of social interactions: Methods, issues, and illustrations* (pp. 209–226). Hillsdale, NJ: Erlbaum.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.
- Campbell, D. T., & O'Connell, E. J. (1982). Methods as diluting trait relationships rather than adding irrelevant systematic variance. In D. Brinberg & L. Kidder (Eds.), *Forms of validity in research* (pp. 93–111). San Francisco: Jossey-Bass.
- Cattell, R. B. (1982). *The inheritance of personality and ability: Research methods and findings*. New York: Academic Press.
- Cone, J. D. (1981). Psychometric considerations. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (pp. 38–68). New York: Pergamon.
- Cone, J. D. (1986). Idiographic, nomothetic, and related perspectives in behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 111–128). New York: Guilford Press.
- Costa, P. T., & McCrea, R. R. (1986). Personality stability and its implications for clinical psychology. *Clinical Psychology Review, 6*, 407–423.
- Costa, P. T. & Widiger, T. A. (1994). *Personality disorders and the five-factor model of personality*. Washington, DC: American Psychological Association.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–434.
- Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University

of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

## 26 I. GENERAL ISSUES

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crowley, S. L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. In J. A. Schinka & R. L.

Greene (Eds.), *Emerging issues and methods in personality assessment* (pp. 285–308). Mahwah, NJ: Erlbaum.

Dielman, T., & Barton, K. (1983). *Child personality structure and development: Multivariate theory and research*. New York: Praeger.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335–337.

Edelbrock, C. (1983). Problems and issues in using rating scales to assess child personality and psychopathology. *School Psychology Review*, *12*, 293–299.

Emmerich, W. (1966, March). Stability and change in early personality development. *Young Children*, 233–243.

Epstein, S. (1984). The stability of behavior across time and situations. In R. A. Zucker, J., Aronoff, & A. I. Rabin (Eds.), *Personality and the prediction of behavior* (pp. 209–268). New York: Wiley.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*, 329–344.

Fiske, D. W. (1982). Convergent-discriminant validation in measurements and research strategies. In D. Brinberg & L. Kidder (Eds.), *Forms of validity in research* (pp. 77–92). San Francisco: Jossey-Bass.

Friman, P. C., Hayes, S. C., & Wilson, K. G. (1998). Why behavior analysts should study emotion: The example of anxiety. *Journal of Applied Behavior Analysis*, *31*, 137–156.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.

Goldberg, L. R. (1995). What the hell took so long? Donald Fiske and the big-five factor structure. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske* (pp. 29–43). Hillsdale, NJ: Erlbaum.

Gresham, F. M. (1984). Behavioral interviews in school psychology: Issues in psychometric adequacy and research. *School Psychology Review*, *13*, 17–25.

Hansen, D. J., Tisdelle, D. A., & O'Dell, S. L. (1985). Audio recorded and directly observed parent-child interactions: A comparison of observational methods. *Behavioral Assessment*, *7*, 389–399.

Hayes, S. C., Nelson, R. O., & Jarrett, H. B. (1986). Evaluating the quality of behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 461–503). New York: Guilford Press.

Haynes, S. N. (1986). The design of intervention programs. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 385–429). New York: Guilford Press.

Haynes, S. N., & Uchigakiuchi, P. (1993). Incorporating personality trait measures in behavioral assessment:

Nuts in a fruitcake or raisins in a mai tai? *Behavior Modification*, 17, 72–92.

Hibbs, E. D., & Jensen, P. S. (Eds.). (1996). *Psychosocial treatments for child and adolescent disorders: Empirically based strategies for clinical practice*. Washington, DC: American Psychological Association.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Jones, R. R., Reid, J. B., & Patterson, G. R. (1975). Naturalistic observations in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 3, pp. 42–95). San Francisco: Jossey-Bass.

Judd, C. M., Jessor, R., & Donovan, J. E. (1986). Structural equation models and personality research. *Journal of Personality*, 54, 149–198.

Kanfer, F. H., & Grimm, L. G. (1977). Behavioral analysis: Selecting target behaviors in the interview. *Behavior Modification*, 1, 7–28.

Karon, B. P. (1968). Problems of validities. In A. I. Rabin (Ed.), *Projective techniques in personality assessment* (pp. 85–111). New York: Springer.

Kaufman, A. (1979). *Intelligent testing with the WISC-R*. New York: Wiley.

Kazdin, A. E. (1977). Assessing the clinical or applied significance of behavior change through social validation. *Behavior Modification*, 1, 427–452.

Kazdin, A. E. (1985a). Alternative approaches to the diagnosis of childhood disorders. In P. H. Bornstein & A. E. Kazdin (Eds.), *Handbook of clinical behavior therapy with children* (pp. 3–43). Homewood, IL: Dorsey Press.

Kazdin, A. E. (1985b). Selection of target behaviors: The relationship of treatment focus to clinical dysfunction. *Behavioral Assessment*, 7, 33–47.

Kazdin, A. E., Kratochwill, T. R., & VandenBos, G. H. (1986). Beyond clinical trials: Generalizing from research to practice. *Professional Psychology: Research and Practice*, 17, 391–398.

Kelly, G. A. (1955). *The psychology of personal constructs* (Vols. 1 & 2). New York: Basic Books.

Kenny, D. A. (1995). The Multitrait–Multimethod Matrix: Design, analysis, and conceptual issues. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske* (pp. 111–124). Hillsdale, NJ: Erlbaum.

Komaki, J. L. (1998). When performance improvement is the goal: A new set of criteria for criteria. *Journal of Applied Behavior Analysis*, 31, 263–280.

Lewis, M. (1997). *Altering fate: Why the past does not predict the future*. New York: Guilford Press.

Linehan, N. M. (1980). Content validity: Its relevance to behavioral assessment. *Behavioral Assessment*, 2, 147–159.

Livingston, S. A. (1977). Psychometric techniques for criterion-referenced testing and behavioral assessment. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology* (pp. 308–329). New York: Brunner/Mazel.

Loevinger, J. (1987). *Paradigms of personality*. San Francisco: Freeman.

### 1. Personality Assessment Research 27

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. (1968). *Statistical theory of mental test scores*. Reading, MA: Addison-Wesley.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, 27, 251–280.

Macmann, G. M., & Barnett, D. W. (1984). An analysis of the construct validity of two measures of adaptive

behavior. *Journal of Psychoeducational Assessment*, 2, 239–247.

Macmann, G. M., & Barnett, D. W. (1985). Discrepancy score analysis: A computer simulation of classification stability. *Journal of Psychoeducational Assessment*, 4, 363–375.

Macmann, G. M., & Barnett, D. W. (1994). Structural analysis of correlated factors: Lessons from the Verbal–Performance dichotomy of the Wechsler Scales. *School Psychology Quarterly*, 9, 161–198.

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: reliability of interpretations for Kaufman’s “Intelligent Testing” approach to the WISC-III. *School Psychology Quarterly*, 12, 197–234.

Macmann, G. M., & Barnett, D. W. (1999). Diagnostic decision making in school psychology: Understanding and coping with uncertainty. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd ed., pp. 519–548). New York: Wiley.

Macmann, G. M., Barnett, D. W., Allen, S. J., Bramlett, R. K., Hall, J. D., & Ehrhardt, K. E. (1996). Problem solving and intervention design: Guidelines for technical adequacy. *School Psychology Quarterly*, 11, 137–148.

Macmann, G. M., Barnett, D. W., Sharpe, M., Lombard, T. J., & Belton-Kocher, E. (1989). On the actuarial classification of children: Fundamental studies of classification agreement. *Journal of Special Education*, 23, 127–149.

Magnusson, D., & Endler, N. S. (1977). *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, NJ: Erlbaum.

Mahoney, M. J. (1980). Psychotherapy and the structure of personal revolutions. In M. J. Mahoney (Ed.), *Psychotherapy process: Current issues and future directions* (pp. 157–180). New York: Plenum Press.

McFall, R. M., & McDonel, E. C. (1986). The continuing search for units of analysis: Beyond persons, situations, and their interactions. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 201–241). New York: Guilford Press.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.

Meichenbaum, D., & Gilmore, J. B. (1984). The nature of unconscious processes: A cognitive-behavioral perspective. In K. S. Bowers & D. Meichenbaum (Eds.), *The unconscious reconsidered* (pp. 272–298). New York: Wiley.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012–1027.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: MacMillan.

Messick, S. (1995). Validity of psychological assessment: Validity of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

Millsap, R. E. (1995). The statistical analysis of method effects in Multitrait-Multimethod data: A review. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske* (pp. 93–109). Hillsdale, NJ: Erlbaum.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

- Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 80, 252–283.
- Mischel, W. (1984a). Convergences and challenges in the search for consistency. *American Psychologist*, 39, 351–364.
- Mischel, W. (1984b). On the predictability of behavior and the structure of personality. In R. A. Zucker, J. Aronoff, & A. I. Rabin (Eds.), *Personality and the prediction of behavior* (pp. 269–305). New York: Academic Press.
- Mischel, W., & Peake, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. *Psychological Review*, 89, 730–755.
- Moss, H. A., & Sussman, E. J. (1980). Longitudinal study of personality development. In O. G. Brim, Jr. & J. Kagan (Eds.), *Constancy and change in human development* (pp. 530–595). Cambridge, MA: Harvard University Press.
- Nelson, R. O., & Hayes, S. C. (Eds.). (1986). *Conceptual foundations of behavioral assessment*. New York: Guilford Press.
- Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574–583.
- Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Peterson, D. R. (1968). *The clinical study of social behavior*. New York: Appleton-Century-Crofts.
- Plomin, R., DeFries, J. C., McClearn, G. E., & Rutter, M. (1997). *Behavioral genetics* (3rd ed.). New York: Freeman.
- Poth, R. L., & Barnett, D. W. (1988). Establishing the limits of interpretive confidence: A validity study of two preschool developmental scales. *School Psychology Review*, 17, 322–330.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Service.
- Ronka, C. S., & Barnett, D. W. (1986). A comparison of adaptive behavior ratings: Revised Vineland and AAMD-ABS-SE. *Special Services in the Schools*, 2, 87–96.
- Runyon, W. M. (1983). Idiographic goals and methods in the study of lives. *Journal of Personality*, 51, 413–437.
- Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The
- 28 I. GENERAL ISSUES**
- principle of aggregation. *Psychological Bulletin*, 94, 39–53.
- Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed). Boston: Houghton Mifflin.
- Schinka, J. A., & Greene, R. L. (1997). *Emerging issues and methods in personality assessment*. Mahwah, NJ: Erlbaum.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.
- Schulte, A. C., & Borich, G. D. (1985). *Using confidence intervals appropriately*. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.
- Schwartz, I. S., & Baer, D. M. (1991). Social validity assessment: Is current practice state of the art? *Journal of Applied Behavior Analysis*, 24, 189–204.
- Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, 23, 153–158.

- Sechrest, L. (1984). Reliability and validity. In A. S. Bellack & M. Hersen (Eds.), *Research methods in clinical psychology* (pp. 24–54). New York: Pergamon Press.
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology. *American Psychologist*, *55*, 1–14.
- Sells, S. B., & Murphy, D. (1984). Factor theories of personality. In N. S. Endler & J. M. Hunt (Eds.), *Personality and the behavior disorders* (2nd ed., Vol. 1, pp. 39–72). New York: Wiley.
- Stagner, R. (1948). *Psychology of personality* (2nd ed.). New York: McGraw-Hill.
- Stagner, R. (1984). Trait psychology. In N. S. Endler & J. M. Hunt (Eds.), *Personality and the behavior disorders* (2nd ed., Vol. 1, pp. 3–38). New York: Wiley.
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.
- Thorndike, R. L. (1963). *The concepts of over- and underachievement*. New York: Teachers College, Columbia University.
- Thorndike, R. L. (1967). Reliability. In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 217–240). New York: McGraw-Hill.
- Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.
- Thorndike, R. L., & Hagen, E. (1961). *Measurement and evaluation in psychology and education* (2nd ed.). New York: Wiley.
- Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings*. (USAF-ASD Technical Report No. 61–97). Lackland Air Force Base, TX: U. S. Air Force.
- Tversky, A., & Kahneman, D. (1984). The framing of decisions and the psychology of choice. In G. Wright (Ed.), *Behavioral decision making* (pp. 25–41). New York: Plenum Press.
- Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Erlbaum.
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.
- Willems, E. P. (1977). Steps toward an ecobehavioral technology. In A. Rogers-Warren & S. F. Warren (Eds.), *Ecological perspectives in behavioral analysis* (pp. 39–61). Baltimore: University Park Press.
- Williams, D. E., & Thompson, J. K. (1993). Issues confronting clinical behavioral therapy: Are we up to the challenge? *Behavior Modification*, *17*, 4–7.
- Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1984). *Multidimensional description of child personality: A manual for the Personality Inventory for Children*. Los Angeles, CA: Western Psychological Services.
- Wolf, M. M. (1978). Social validity: The case for subjective measurement, or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*, 203–214.
- Wothke, W. (1995). Covariance components analysis of the multitrait-multimethod matrix. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske* (pp. 125–144). Hillsdale, NJ: Erlbaum.
- Zsombok, C. E., & Klein, G. (Eds.). (1997). *Naturalistic decision making*. Mahwah, NJ: Erlbaum.

## A PSYCHOMETRIC AND CLINICAL EVALUATION OF SELECTED STUDIES

A psychometric and clinical analysis of the

hundreds of projective drawing studies available in the literature necessarily involves an evaluation of each study. This is primarily because of the vast differences in psychometric and experimental quality among these studies, and because these studies provide such widely differing methodological details that one cannot assume that results are clinically accurate and socially valid without additional validation and replication. Despite this need for individual evaluations, few critical reviews of the literature are available for projective drawings. More common are (1) collections of projective characteristics and interpretations based on the work of individual clinicians, who have investigated self-selected populations and truncated samples that often are nonrandomly distributed, and/or (2) reviews that fail to integrate projective drawings into a multitrait, multisetting, multimethod assessment approach (Campbell & Fiske, 1959), which would demonstrate their ability to generate integrated hypotheses rather than isolated, relative meaningless results. The lack of a comprehensive literature review leaves the clinician without an objective and useful measure of the current status of projective research, and the field without a clear direction as to what future directions are most necessary and most relevant.

Relative to the psychometric qualities of projective drawings, most of the recent research has investigated the ability of projective drawings to identify and/or discriminate various specific clinical groups of children and adolescents—sexually abused children (e.g., West, 1998), mentally ill chemical abusers (Taylor, Kymissis, & Pressman, 1998), and mood- and anxietydisordered children (Tharinger & Stark, 1990). In addition, whereas West (1998) conducted a meta-analysis of studies assessing the efficacy of projective techniques in discriminating child sexual abuse and Handler and Habenicht (1994) reviewed the psychometric literature on the Kinetic Family Drawing Technique, few, if any, recent studies have specifically evaluated the reliability of projective drawings. Indeed, most of these studies have simply reported reliability data within the context of their broader research agenda.

More historically then, relative to reliability, Cummings (1986) evaluated a dozen

studies between 1968 and 1981 with respect to interjudge reliability and reported reliabilities ranging from .75 to .97, with the critical determinants of good interjudge agreement being (1) the specificity of the items, characteristics, or drawing styles to be scored and (2) the training and supervision of the various judges to insure similar scoring criteria and conditions. With respect to test–retest stability, Cummings critically reviewed nine studies with retest intervals from 1 day to 3 months and stability coefficients ranging from .69 to .91. With stability generally decreasing as the test–retest interval increased, the importance of simultaneously evaluating psychometric quality and clinical utility becomes apparent. Simplistically, this decreased stability could suggest that projective drawing techniques are unreliable over time, or it may be accurately reflecting a child’s change in social-emotional status. Or, the decreased stability may be due to a scoring system that attends more to structural changes in a child’s initial versus later drawings (e.g., the presence of excessive shading in the first drawing and lack of shading in the second drawing) than to drawing characteristics that reflect changes in a child’s clinical status (e.g., the presence of anxiety).

Regardless, it appears that with the right scoring criteria and evaluative approaches, acceptable levels of interrater reliability are possible with projective drawings (see, e.g., Handler & Habenicht, 1994, for a review with the K-F-D technique). With respect to test–retest reliability, the dilemma of evaluating the psychometric nature of projective drawings with their *clinical* meaning and validity becomes apparent. And, in the absence of (many) composite scale approaches to scoring projective drawings, calculations of their internal reliability are virtually impossible. These limitations, relative to determining a projective drawing technique’s reliability, explain why these approaches should be used only for the generation of social, emotional, or behavioral hypotheses and not to make clinical diagnoses or conclusion. Indeed, consistent with the multitrait–multimethod approach, hypothesis generation may be the most defensible and helpful contribution of projective drawings. The validation of projective drawing techniques has been extremely difficult to

document, and an extremely variable research literature exists in this area. Three types of validity are commonly found in this literature: concurrent validity, construct validity, and incremental validity. There are few studies of concurrent validity in the projective drawing literature. These studies typically correlate specific diagnostic indicators in projective drawings with more objective scales or indices that already are validated with respect to these indicators. Conceptually, this is an acceptable method of validation. Pragmatically, however, its success is based on the integrity of the methodological procedures used and the ability of the researchers to choose samples that have some level of generalizability to specific and important clinical groups or to more universal populations. More concurrent validity research with projective drawings is needed before definitive conclusions can be made as to their utility. Although the issue of interpretation accuracy is important with projective drawings, this issue is equally critical in the interpretation of many so-called objective personality assessment tools.

Studies of construct validity with projective drawing techniques have often focused on the “known-groups” method, which uses a validated anchor scale to separate a large sample of children into clinically differentiated subgroups. These subgroups then complete a projective drawing battery, which is analyzed for its ability to approximate the same clinical separations. A variation of this approach investigates the discriminant validity of projective drawings—that is, their ability to discriminate between samples that either represent substantively different clinical groups or involve a clinical group that is compared with a control or nonclinical group. This type of construct validity has been used in numerous studies of projective drawing techniques with equivocal or nonsignificant results. For example, Hibbard and Hartman (1990) compared the presence of Koppitz’s Emotional Indicators and categories in HFDs completed by alleged sexual abuse victims between the ages of 5 and 8 years and comparison (presumably nonabused) children. The results indicated that although more anxiety indicators were present in the sexually abused children’s drawings, no significant differences in specific indicators or categories

between the two groups was evident. Extending this research to include studies in which effect sizes could be calculated, West (1998) conducted a meta-analysis of 12 studies that used projective techniques (among them the Draw-A-Man, H-T-P, HFD, and K-F-D) to discriminate sexually abused versus nonabused children. While the results indicated that projective tests, in fact, did discriminate abused versus nonabused students, the specific contribution of projective drawings in this result could not be determined.

The lack of discriminant validity for projective drawings again was found by Tharinger and Stark (1990), who administered the DAP and K-F-D to 52 fourth through seventh graders with mood, anxiety, or mood and anxiety disorders and 13 nonclinical control students. Using the Koppitz scoring system for the DAPs and an adapted Reynolds scoring system for the KF-Ds, the results indicated that none of the 30 DAP or 37 K-F-D emotional indicators differentiated any of the three clinical groups from the normal controls. Finally, these results were generally supported by Handler and Habenicht (1994), whose review of the K-F-D literature found numerous studies in which the K-F-D did not consistently discriminate between different groups of children, including those who were well-adjusted versus those with specific clinical difficulties.

Overall, construct or discriminant validity is dependent on the scales or procedures used to initially separate a sample into clinical subgroups; the methodological characteristics and reliability of the projective drawing administration and scoring, respectively; and the statistical analyses used to as-

102 II. PROJECTIVE METHODS

sess the diagnostic "hit rates" and false-negative and false-positive decisions. Despite numerous studies failing to demonstrate the discriminant validity of projective drawing tests, Handler and Habenicht's (1994) recommendation that future studies use a holistic, integrated scoring approach rather than a specific, indicator-by-indicator approach is notable. In the end, more validity research using this integrated approach is needed before conclusive statements can be made.

The third type of validity, incremental validity, has focused on projective drawings'

ability to facilitate more accurate diagnostic and clinical judgments when used in conjunction with other assessment tools and approaches in a stepwise decision-making process. To date, few studies have demonstrated this type of validity with projective drawings, and some studies have reported that various clinicians were unable to discriminate clinically unremarkable versus clinically identified populations beyond the chance level with projective drawings (Cummings, 1986). Despite the absence of more recent incremental validity research, Hammer (1997) still defends their use and minimizes the results of the previous research, noting (1) that projective drawings provide hypotheses that are more descriptive than diagnostic, especially when many clinical groups share similar psychological characteristics and behavioral patterns; and (2) that projective drawings should not be used independently to discriminate clinical groups but as part of an integrated assessment battery involving numerous objective and other diagnostic approaches. Thus, the psychometric debate with respect to projective drawings continues. Some researchers decry their lack of psychometric integrity, whereas others argue that projective drawings should not and cannot be fairly evaluated in a traditional statistical and psychometric manner. This argument aside, the hypothesis-generating use of projective drawings and their ability to be interpreted within various psychological orientations remains both viable and defensible. Moreover, as noted previously, the generation of hypotheses tied to specific drawing characteristics is discouraged in favor of a more integrated, composite approach to projective drawing analysis and interpretation. This latter approach, however, still must be validated in the future research. To assist in the overall analysis and evaluation of the individual studies that make up the research with projective drawings, a review of many of the studies most often cited in the projective drawing literature are shown in Appendices 5.1 and 5.2 in this chapter. These tables analyze the predominant interpretations of projective drawing techniques and the BGT according to their experimental designs, their statistical and methodological characteristics, their dependent variables and results, and their generalizability to other clinical situations in the

field. The first part of each of these appendices reviews the research prior to 1990, while the second part of each reviews the research after 1990.

Although exhaustive descriptions and evaluations of the studies in Appendices 5.1 and 5.2 are beyond the scope of this chapter, some descriptive analyses of these studies are possible as one way to analyze the state of projective research to date. Within the projective drawing studies described in the first part of Appendix 5.1 (i.e., prior to 1990), 50% used ex post facto designs, 37% used descriptive “designs,” 9% used experimental designs, and 4% used case-study designs. Sixty-eight percent of the studies did no systematic type of sampling, whereas 22% used some type of randomized sampling and 10% used some type of matched sampling. Seventy-two percent of the studies did not use control groups, whereas 28% did; 56% analyzed their data using nonparametric statistics, whereas the other 44% used some type of parametric analysis. Finally, in an external analysis, 20% of the studies were deemed to have good generalizability to other samples, 51% of the studies were considered to have fair generalizability, and 29% of the studies were felt to have limited or no generalizability. Of the 12 projective drawing studies reviewed since 1990, 75% used ex post facto designs, 33% used some type of control or comparison group with some type of systematic sampling (no study used random sampling), 75% calculated interrater reliability (one calculated internal reliability), 83% analyzed their data using parametric statistics, and only one study was considered to have good generalizability. Notably, of the newer studies, one study did a literature review of the K-F-D research (Handler & Habenicht, 1994) and one completed a meta-analysis of projective tests with sexually and nonsexually abused children (West, 1998).

Of the pre-1990 BGT studies described in Appendix 5.2, 53% used ex post facto designs, 44% used descriptive “designs,” and 3% used experimental designs, and no casestudy designs were reviewed. Sixty-three percent of the studies did no systematic type of sampling, whereas 22% used some type of matched sampling and 15% used some type of randomized sampling. Fifty-three percent of the studies used control groups,

whereas the other 47% did not; 62% analyzed their data using parametric statistics, whereas the other 38% used some type of nonparametric analysis. Finally, in the same type of external analysis, 22% of the studies were deemed to have good generalizability to other samples, 51% of the studies were considered to have fair generalizability, and 27% of the studies were felt to have limited or no generalizability. Significantly, only two BGT studies were found after 1990, one with subjects ranging from 15 to 65 years old (Maloney & Wagner, 1991) and one with male adolescents in a residential treatment center (Nyfield & Patalano, 1998). Neither of these studies had good generalizability to broader populations. It is important to note the significant number of descriptive “designs” used in both the projective drawing and BGT studies, even though they were outnumbered by the ex post facto designs. Even more notable is the dearth of experimental studies, a gap that points to the need for far more sophisticated research in the future—research that should especially target the validity and clinical utility of projective drawings. Furthermore, the studies’ general lack of appropriate sampling and control groups and their dependence on the less robust nonparametric statistical analyses suggest that research on projective drawings has yet to fully evaluate their real potential and their actual impact. Finally, the relatively small percentage of studies considered to have good generalizability indicates that much of the projective drawing research cannot be used for differential diagnosis, or even for validation of specific clinical characteristics within a referred individual or an identified group. This reinforces the use of projective drawings as hypothesis-generating tools rather than hypothesis-validating tools. This use and perspective are further explicated later. Two final points, however, remain. First, it is interesting to note that the BGT studies (Appendix 5.2) used significantly more control groups and parametric statistical analyses than did the projective drawing studies, although this may be changing given the post-1990 studies (Appendix 5.1). Although this may be due to the BGT’s more explicit and investigated scoring systems, which have been standardized and normed around the country, the real reasons are not otherwise apparent. Second, few if

any of the studies reviewed used multitrait-multimethod designs. This is a significant flaw that again, in summary, points out the relative weakness of current projective drawing research. For the future, the areas of potential and needed research are both apparent and unlimited. Projective drawings cannot be evaluated on the basis of the present research; only after a great number of experimentally sound studies have been completed can these assessment tools and techniques be fairly critiqued.

### INTEGRATING PROJECTIVE DRAWINGS INTO COMPREHENSIVE PERSONALITY ASSESSMENT

When projective drawings are being integrated into a comprehensive personality assessment, the primary goals of such an evaluation need to be considered. Elsewhere (Knoff, 2002), in discussing the comprehensive personality assessment process and its goals, I have identified two critical conceptualizations of that process: (1) personality assessment is a hypothesis-testing, problemsolving process that works within a probability model; and (2) personality assessment uses a multitrait, multisetting, multimethod assessment approach, which is sensitive to the ecological and reciprocally determined nature of personality and behavior. Personality assessment, in its most basic form, exists to functionally evaluate (e.g., O'Neill et al., 1997) children who have been referred for some behavioral, affective, intrapersonal, and/or interpersonal difficulty that threatens some critical domain or facet of

104 II. PROJECTIVE METHODS  
their development (e.g., social-emotional, academic, and cognitive). The process involves systematic problem solving as it attempts, first, to identify and functionally analyze the primary referred (and other relevant) problems from a multifaceted, ecological perspective, and second, to intervene in the problem with appropriate and effective interventions. Evaluation of these identification, analysis, and intervention components occurs throughout as a way to maintain both the integrity and the utility of the entire process.

During problem analysis, the psychologist evaluates the referred problem and situation, using multitrait, multisetting, multimethod assessment approaches to generate hypotheses that best explain and facilitate an understanding of the significant variables

that are causing, supporting, maintaining, and/or encouraging the problem. Once these hypotheses are identified and refined, the analysis process continues at a higher level as the hypotheses are tested, using the most objective, empirically proven methods, for their validity and reliability across people, settings, time, and ecological circumstances.

Confirmed hypotheses that explain why a referred problem occurs (e.g., the child exhibits anxiety because he was once left home alone during a power loss) then links directly to interventions (e.g., relaxation therapy and/or systematic desensitization), while rejected hypotheses necessitate continued problem analysis.

Given their more clinical nature (both in the research and in actual practice), and their focus on children's covert processes, projective drawings probably are best used to *generate hypotheses* about a referral situation rather than to *validate those hypotheses*.

Indeed, at times, projective drawings may be unnecessary in the assessment battery, because other tools or techniques can identify and test the relevant hypotheses more validly and effectively. Within the personality assessment process, problem analysis is deemed successful when a functional understanding of a child's referred problem is reached and interventions to address the precipitating problem conditions or events are identified. In most cases, and in the face of their limited clinical validity, this suggests a limited role for projective drawings in this process.

After the referred problem has been analyzed to the extent that its primary determinants or contingencies are clear and the necessary interventions are apparent, the process of psychological change can begin.

Although the link between assessment and intervention is logical and methodologically necessary, the presence of unassessed and unknown intervening variables can decrease the predictive validity of intervention success.

Thus, personality assessment in general and the development of effective interventions in particular work within the context of probability. That is, given thorough and well-conceived analyses, we expect that our explanations of referred problems have the *highest probability* of accuracy, and that our interventions based on these analyses have the *highest probability* of success. At times, however, intervening variables (e.g., a

teacher's covert lack of commitment to the intervention) are missed, and the efficiency and efficacy of the process are diminished. In most cases, the intervening and disruptive variables will be identified through formative evaluations of the ineffective intervention program. In other cases, the intervening variable may be the intervention program *itself*, and the decrease in program effectiveness may indicate that some facet of the original referred problem has been solved and that a different problem, unaffected by the intervention, has taken its place. In both cases, the intervention must be reconceptualized, and the assessment-to-intervention link must be forged anew. At this juncture, problem analysis and intervention may focus only on observable and measurable behavior. Once again, beyond suggesting additional explanative hypotheses, projective drawings may be of little assistance at this level.

To summarize, personality assessment is a process, not a product. The real product, the primary goal of the process, is the treatment or resolution of referred behavioral or social-emotional problems so that a child's normal development and positive mental health can continue. Thus, it is not enough to *describe* or *understand* a child's socialemotional problems; we must move from problem analysis to intervention by *using* this understanding. On a different level, given the aforementioned goals of personality assessment, it should be apparent that *not*

5. Evaluation of Projective Drawings 105

*all personality assessment evaluations with a referred child necessitate the use of projective tests, or, more specifically, projective drawings.* In fact, psychologists should use only those personality assessment tools or techniques that are needed (1) to fully and validly identify and analyze the significant variables that are causing, supporting, maintaining, and/or encouraging the problem; and (2) to fully and validly identify those interventions that will effectively and efficiently resolve these significant variables and the referred problem. In most cases, these necessary and sufficient personality assessment tools and techniques will involve those that are behaviorally based, objective, standardized, and/or well researched. Projective drawings, then, may be used when a referred problem situation is particularly complex, when an in-depth evaluation of a

child's intrapersonal and cognitive-behavioral status is needed, and/or when behavioral and objective assessments have not provided an understanding of the problem situation that makes intervention success highly probable. Projective drawings *can be* important to the personality assessment process. They should be used strategically, however, and not randomly; they should be used to maximize problem understanding and treatment effectiveness, not to cloud the issues with unnecessary redundancies or irrelevancies.

On a more pragmatic level, when used in the personality assessment battery, projective drawing techniques do provide potentially unique information and samples of behavior that are unavailable through other approaches, be they behavioral, objective, or anecdotal. From a cognitive-behavioral perspective, projective drawings use administration approaches and test-related stimuli that allow children to introduce their own significant, egocentric beliefs, attitudes, expectations, or attributions into the assessment session. These cognitions generally are not elicited by the more standardized techniques, which are often so structured and so specific in the personality domains they are intended to evaluate that they either suppress children's more personal self-reflections and intricacies or assess in ways that appear not to relate to a child's actual or perceived difficulties. Projective drawings are among a handful of techniques that do not fully structure the assessment interaction, providing instead an opportunity for children to communicate, in their own way, the specific issues that are troubling them. In a sense, then, projective drawing techniques may be most similar to unstructured or semistructured clinical interviews, except that the psychologist initially asks the child to draw (or, in a sense, to discuss visually) some event or facet of his or her life. Later, during the Inquiry process, the child can describe or discuss the drawing and, with effective clinician rapport and interview skills, may reveal important personality- or behaviorally related issues.

To expand on this idea briefly, the clinical interview can be just as "projective" as the projective drawings. On one hand, some interview questions are more direct, and the interpretation of a child's "direct" responses may appear to be self-evident and objective. However, the child still can decide whether

or not to answer these direct questions honestly or completely. Thus, like the projective drawings, interview responses may only generate hypotheses that need additional, external validation. On the other hand, many interview questions are purposely global and ambiguous, in the hope that they will trigger some important memory, emotional reaction, or relevant response by the child, and/or so that the child has the opportunity to respond with personal issues that are of real and current concern. Again, this is similar to the projective drawings, except that the drawings use a visual modality as compared to the interview's verbal modality. To continue, and as noted, the success of the clinical interview in eliciting information of importance often depends on the rapport and trust between the child and the interviewer, and the ability of the interviewer to ask the right questions at the right time. The success of projective drawings is dependent on the same factors. Finally, and somewhat parenthetically, it is interesting to note that clinical interviews suffer the same potential problems with respect to reliability and validity as do projective drawings, yet they seem to be more often included within the personality assessment battery because the psychologist somehow "has more control" over the interview. I would suggest that this logic is faulty, but not in order to recommend that the clinical

## 106 II. PROJECTIVE METHODS

interview be purged from our assessment procedures. Clearly, the clinical interview *is* an important, almost irreplaceable, procedure. Projective drawings, in certain cases, may be just as important; they cannot be categorically dismissed from the assessment battery, and their use should be at least considered for referrals that require face-to-face testing with the referred child.

Although space considerations preclude a description of every projective drawing technique—its administration and scoring procedures, its interpretive approaches, and its integration into the comprehensive personality assessment battery—a few generalizations are important. First, projective drawings can be analyzed from both structural and content perspectives. Specific structural characteristics of a drawing (e.g., actions of and between drawn figures; characteristics of specific figures; the position, distance between, and barriers between figures;

the drawing style; and symbols or objects present in the drawing) are interpreted depending on the psychological orientation being used for analysis. These interpretations are evaluated both *within* drawings and *across* drawings, such that recurrent themes and issues are identified. These recurrent themes, and not the random and isolated ones, are integrated into the broader diagnostic picture of the referred child, which includes all other assessments and pieces of data, all multisetting evaluations, and all multisource observations. These themes are what generate the projective hypotheses for further study and validation.

Overall, however, given the research reviewed earlier, there appears to be little, if any, consistent empirical support for a structural approach to projective drawing analysis and interpretation. Those structural indicators that have some interpretive support are generally idiosyncratic to the studies in which they are cited, and they do not, figuratively, account for enough variance in explaining a child's difficulties to warrant a structural analysis for all other indicators. For those indicators that have little interpretive support, while additional research can proceed, it should focus on clusters or composites of indicators that can contribute to a functional analysis of a child's affect or behavior that will then link to effective intervention.

At present, then, the structural analysis of projective drawings is discouraged.

The content analysis is related to the second important generalization: the importance of an Inquiry process during the projective drawing administration. Projective drawings are administered in two phases.

The Performance phase involves providing the child with the materials needed to complete the drawings and with the actual directions needed to proceed. After the drawing is completed, the Inquiry phase begins, using a series of questions aimed at (1) clarifying the objects and actions in the actual drawing and (2) eliciting a broader description or story about the drawing, which expands into a semistructured interview focusing on the child's underlying decisions and cognitions in choosing to draw what was completed. The child's responses to the inquiry-phase questions are used in the content analysis.

The content analysis has the same goal as the structural analysis: to generate hypotheses

that contribute to a comprehensive understanding of the referred child. Although the source of the data differs from the structural analysis, the content analysis should also evaluate themes within and across projective drawings and integrate only recurrent themes in the broader assessment and interpretation process. In some ways, the content analysis can be conceptualized as similar to that done with the thematic techniques (e.g., as in the Thematic Apperception Test). That is, questions about the projective drawing can be put into story format, and important details about the characters in the pictures can be ascertained. This is not necessary; what is necessary is that the inquiry process be used to clarify and expand the information in the projective drawing. This process decreases the potential for misinterpretation and increases the potential for eliciting significant cognitions and issues that are inherent in the material of the projective drawing. The third important generalization relates to training and expertise with projective drawings. The administration, scoring, and interpretation of projective drawings take a great deal of training, practice, experience, and supervision. One does not simply read "the book" or take "the course" to become "expert" in projective drawing use. Instead,

5. Evaluation of Projective Drawings 107

the effective utilization of projective drawings is an ongoing endeavor. It would be improper to require some specific amount of training or supervision before one could use projective drawings independently. In a sense, the only criterion of readiness is the same criterion on which projective drawing interpretation should be evaluated as valid: whether or not the projective drawings elicit valid hypotheses that contribute significantly to a greater understanding of the referred child and to a subsequently successful intervention program. If a psychologist is able to use projective drawings consistently in this way under supervision, then he or she should be considered ready to use them independently. All this, however, appears far too specific to projective drawing techniques. Clearly, if projective drawings are a potential part of the comprehensive personality assessment process, then psychologists need to be expert in the *entire process* before practicing it independently. Personality assessment is both a content- and a process-oriented

skill. It is a skill that develops over time and experience. It is a practice that requires ongoing supervision and continuing education to best serve referred children in the settings in which they live and interact.

### **CASE STUDY: INTEGRATING AND INTERPRETING PROJECTIVE DRAWINGS**

This section presents a case study that demonstrates the interpretation of projective drawings and their potential integration into a comprehensive personality assessment battery.

A “stream-of-consciousness” format is used throughout this discussion to show the psychologist’s thinking processes as numerous personality assessment tools and techniques were integrated and a complex case was fully analyzed. The projective drawings in this case were analyzed primarily as data providing behavioral and social-emotional hypotheses that would need to be confirmed using more objective and direct means.

However, they did sample areas of personality and cognitive-behavioral functioning that other assessment mechanisms might evaluate only tangentially or in different ways. Finally, a multitrait, multisetting, multimethod assessment approach was used in this case study. The number of multiple assessments, however, was limited to only those needed for reliable and valid evaluation of the various traits, settings, and methods unique to the referred child and situation. This number will vary from case to case and from psychologist to psychologist; with every case, however, there is a point of diminishing returns where another assessment tool or interview adds nothing new to the data or the analysis of the referred problem.

#### **Referral and Background**

David, at the time of referral and evaluation, was between 15 years, 4 months and 15 years, 8 months old. An athletically built white adolescent, David was placed during this assessment period in a full-time residential program for students with learning disabilities (LDs), many of whom had associated behavioral and social-emotional difficulties. Indeed, David has recently been kicked out of school after numerous fights with other students and a serious confrontation with his high school principal, who eventually suspended him. In an initial interview, David spent a good deal of time bragging about how tough his high school was and made allusions to his possible drug use. He complained that he was weak in

spelling, English, and reading; that he could do better in math, but that he did not like to hand in assignments or study for exams; and that he did best in classes in which he could listen to the course material rather than take a lot of class notes.

A developmental interview with David's mother indicated that David was an unhappy baby, that he cried constantly during his first months of life, and that he never stayed in his playpen for a long period. The mother characterized him at 9 months as a "child who constantly defied authority—when he learned to walk, he would not stop grabbing things, even when he was told." She noted that he walked comparatively early, and that "he always amazed everybody" with what he could do. For example, she noted that he would carry on a full conversation at a young age.

Specific to their disciplinary styles, the mother noted that she and her husband never agreed on how best to discipline David. She would typically spank him, carry a specific punishment through for a period of time, and then give David back the privileges lost. The father would usually threaten David but not carry through. Early on, however, the mother noted that David was quite facile at manipulating her and his father. This manipulation had continued to the present day, and the mother fully acknowledged that she was tired of dealing with David, that she was ineffective in controlling and structuring him, and that the residential placement was probably best for David *and* her.

Academically, David did appear to present a classic LD pattern. His problems began in second grade, when he was tested for reading and associated problems and considered "dyslexic." David's mother recalled many frustrations in trying to help David to learn how to read, especially because a significant number of behavioral problems were already surfacing, at home and in school, as a result of David's frustration with his academic failure. In school, David was exhibiting a lot of out-of-seat behavior, daydreaming, argumentativeness, and anger. He was eventually placed in an LD program for part of the year. At the same time, however, David also was considered lazy, and his parents would put him on restriction at home because he was not doing

## 108 II. PROJECTIVE METHODS

his work in school. After the special education placement, David's LD teacher actually blamed David and his behavior for a miscarriage that she experienced during the school year.

In fourth and fifth grades, David went on to a private school, where he was provided with special training and support as a student with LD. In fact, during one of the summer programs at that school, David's mother recalled a "phenomenal" teacher who was particularly successful with him. It was during this summer that he made the most progress, academically and emotionally. David's mother bemoaned the fact that he had never had as talented a teacher since that summer, and that David had never felt so positive about his academic accomplishments or future. Ultimately, David's family relocated to another state and he returned to the public schools for his middle school and early high school years. During those latter years, David's mother noted that she lost control of him, that he was running with a crowd of underachievers who were involved with drugs and continually skipped school, and that the school administration was unresponsive to both educational and behavioral needs. David's mother saw the current program at the residential LD institution as David's "last chance"; she felt that his failure in that program would leave him as a dropout with no educational foundation to get an appropriate job in the work force.

As for his family life, David had a younger sister, but she did not seem to have any relevance to his referral problem or history. More significantly, David's parents experienced a number of marital problems as he was growing up. In fact, they divorced when David was 10 years old but were separated only 3 months; they eventually remarried only 9 months after the divorce.

David's mother noted that her husband's problems with alcohol were a significant reason for the divorce. She also noted that David blamed himself for the divorce and that he was very "lost" during it. Nonetheless, no one in the family was involved in any psychological counseling or support during the time immediately before or after the divorce and remarriage. In fact, the only counseling the family ever experienced occurred for only 1½ months and centered around David's LD problems when he was

in third grade. David's mother noted that the psychologist involved was negative and tried to blame her and her husband for many of the problems they were experiencing. The psychologist's "blame" notwithstanding, it did appear that David's parents had unrealistic expectations for his behavior even when he was 9 months old (e.g., describing his behavior as "defying authority"), and that their inability to discipline him consistently had left David with little external structure, few explicit behavioral and interpersonal expectations, and no real consequences for either positive or inappropriate behavior. David's parents did not fully understand his LD, nor were they prepared to deal with its academic and social-emotional characteristics. Finally, all these situations had existed for such a long time that most of the familial behavioral patterns were ingrained and intractable—both for David's parents with respect to parenting and for David with respect to his interactions both at home and in school. At the time of the interview, it appeared that a residential program was appropriate for David, but that both he and his parents needed specific counseling support to change their patterns of behaviors and expectations so that David could eventually and successfully be reintegrated into the family and community setting. The assessment goal, then, was to gain a full understanding of David's intrapersonal, interpersonal, academic, social-emotional, and familial behavior and self-concept, so that appropriate and comprehensive therapeutic directions could be developed and implemented.

#### **Initial Assessments**

The first part of this ongoing assessment process began when David was evaluated for entrance into the residential placement at 15 years, 4 months of age. At that time, a battery of intellectual (the Wechsler Intelligence Scale for Children—Revised [WISC-R]), academic achievement (the Wide Range Achievement Test, the Detroit Tests of Learning Aptitude, the Gates–MacGinitie Reading Tests, the Gilmore Oral Reading Test, the Metropolitan Achievement Tests, the Schonell Spelling Lists), processing (the Illinois Test of Psycholinguistic Abilities, the Frostig Visual Perceptual Tests), and personality tests (the Personality Inventory for Children [PIC], the Millon Adolescent Personality

Inventory [MAPI]) were given.

On the WISC-R, David received a Verbal Scale IQ in the average to above-average range, a Performance Scale IQ in the average range, and a Full Scale IQ in the average range ( $100 \pm 6$ ). On the Verbal subtests, he demonstrated consistent verbal skills—conceptually, expressively, with respect to academic information, and in arithmetic. Significantly, David achieved a scaled score of 13 on the Comprehension subtest, indicating that he understood societal expectations and appropriate behavior, even though he might not choose or might be unable to behaviorally meet those expectations.

On the Performance subtests, David did extremely well on the Picture Completion and Picture Arrangement subtests, less well and average on the Block Design and Object Assembly subtests, and extremely poorly on the Coding subtest. It appeared that David's strong verbal skills were helping him to "talk himself through" and score well on these first two subtests. Thus, David might be able to control his behavior and aggressive actions through cognitive, mediational processes or through a "thinkaloud" approach if motivated to do so. His low Coding subtest (scaled score, 4) and his low Digit Span subtest (scale score, 4) suggested some significant distractibility, probably associated with his LD, and/or some significant anxiety and emotional lability. Actually, both of these hypotheses were found to be accurate. David scored in the at-risk range on the Gordon Diagnostic System, an objective attention and vigilance task that assesses for attention-deficit disorders. And the personality tests and interview with David's mother suggested that he had significant behavioral and emotional difficulties, especially when frustrated and stressed.

In the academic achievement areas, David's spelling achievement was measured at approximately the third-grade level, his oral reading at the fourth- to early fifthgrade level, and his math skills at the fifth to early sixth-grade level. Significantly, his oral reading comprehension was at grade level or above when he was able to decode the specific words in the reading passage. In the processing area, David's auditory memory was almost age-appropriate when the stimulus material was concrete and relevant to him, and his ability to process and follow

oral directions was excellent. But when the auditory material became abstract and/or novel, David had significantly more memory problems. David's visual memory and visual processing abilities, in contrast, were well developed. In fact, the visual modality in general appeared to be David's stronger processing area.

When this psychometric information was compared with his intellectual test results, it appeared that David was still manifesting an LD pattern: specific difficulties in reading, auditory memory, auditory processing, and spelling, and a potential attention deficit disorder. As for his social-emotional status, David's history of LD and its concomitant pattern of academic and social failure had clearly influenced his personality development, and in some cases his LD

#### 110 II. PROJECTIVE METHODS

might be influencing how he saw and interpreted circumstances and events in his past and present worlds. Although his social-emotional status might be evaluated separately, it could not be analyzed without considering his academic history and its influence both on his self-concept and on the behaviors and expectations of his parents. This was not an attempt to explain all of David's behavior as a part of his LD; rather, it was an attempt to recognize the complexity of his behavior and the need for a complex analysis of his difficulties and a multifaceted intervention program.

David's social-emotional status initially was evaluated with the PIC, completed by his mother, and the MAPI, completed by himself. David's personality profile on the PIC (see Figure 5.1) indicated severely delinquent and hyperactive tendencies that were apparent to such a degree that significant antisocial, out-of-control, and irrational behavior could be expected. In addition, David was seen as an adolescent (1) who lacked appropriate interpersonal and social skills and the ability to deal with conflict-producing situations in the environment; (2) who had significantly withdrawn from his environment and who might exhibit periods of depression along with agitated, acting-out behavior; (3) who felt a great deal of anxiety and distrust in his world; and (4) who continued to have significant intellectual and academic achievement problems in the midst of all the social-emotional turmoil. All but three of the PIC's clinical

scales reached significance, suggesting a disturbed, emotionally labile individual. Two of those nonsignificant scales related to David's overall development and his lack of somatic concerns. The other scale was the Family Relations scale, a scale that assesses the relationship between family and parental variables and the child's psychological difficulties. This result suggests (1) that many of David's problems were thought to be specific to him, rather than related to family problems or issues; (2) that many of David's problems influenced characteristics that did exist within the family unit; and (3) that David's absence from the home because of his placement in the residential program was having a positive effect on the atmosphere and intrafamily relationships at home.

The MAPI (see Figure 5.2) reinforced many of the results from the PIC. On this scale, David was described as both fearful and apprehensive about his relationships with others, and as resentful and critical of those who had not supported him in the past. His profile reflected a person who was sensitive as to how others evaluated him as a person and predicted that David would tend to withdraw and distance himself from close involvements to avoid further rejections or disappointments. The MAPI further indicated that David revealed feelings of low self-worth, expressing a minimal degree of self-acceptance and a great deal of difficulty with academic achievement and success. The family problems and David's perceived rejection by his parents were also evident in the profile, as were his indifference to others and his lack of empathy for others and their welfare. In total, the MAPI results predicted that David would have limited success in therapy and that he would probably withhold his innermost feelings, primarily because he would not believe that anyone could really care about him in the end.

### **Continuing Assessments**

After the initial assessments, David was accepted by the residential program and placed in an individualized LD curriculum and a dorm with his own room. Over the next 4 months, he was behaviorally monitored and discussed at monthly meetings of the clinical staff. During the first of these staff meetings, David was described as hyperactive, sneaky, and nontrusting, and as

continually breaking the no-smoking rules. It was noted that he did serve consequences when they were applied, but that he had put his head through a wall of his dorm room, had made additional holes there with some metal rods, liked to stare people down, and was making no progress in his academic program. At the second meeting, this pattern was described as continuing, with additional concerns about his manipulative behavior and his constant challenging of both school and residential staff. By the third meeting, David's belligerent attitude, lack of respect for adults, inability to work independently, and inability to be trusted during off-campus activities resulted in a call for

5. Evaluation of Projective Drawings 111

#### 112 II. PROJECTIVE METHODS

**FIGURE 5.1.** David's Personality Assessment Profile on the PIC.

additional diagnostic assessments to determine whether he was appropriate for the program or not. At this point, a more projectively oriented battery was completed to supplement the data already collected during the intake process. This battery consisted of the Rorschach, some incomplete-sentence blanks, the H-T-P, the Kinetic Drawing System (the K-F-D and K-S-D), and four sessions of diagnostic counseling and clinical interviews.

David's Rorschach evidenced a significantly high number of total responses; significantly high percentages of large blot area responses (D%) and Animal responses (A%); and significantly low percentages of small blot area responses (Dd%), high-quality Form responses (F+%), Animal Movement responses (FM), and Human responses (H%) as compared to same-age norms from Ogdon (1977, 1982). In addition, his responses qualitatively focused on aggressive activities, monsters, and "scary" scenes fairly often. The interpretation of the Rorschach suggested an above-average intellectual potential, anxiety, a suppression of emotions and spontaneity as a way to protect against rejection, social isolation, and an absence of empathy for others. The analysis also indicated that David would probably be a poor therapy candidate because of his inability to trust others and to overcome his feelings of vulnerability and cautiousness.

5. Evaluation of Projective Drawings 113

**FIGURE 5.2.** David's personality assessment profile on the MAPI.

David's incomplete-sentence blank responses

revealed the following issues: his desire to smoke on the grounds of the campus, his dislike for authority figures who told him what to do, his poor peer interpersonal relationships, his dislike for school and the residential school in particular, and his secret admiration for his father and his desire to graduate from high school to please his father. Some of the more pertinent sentences were as follows (the stems are in roman, David's responses in *italics*):

People *suck—they are sometimes not the nicest things in the world.*

In the lower grades *I was always a very bad student.*

Other kids *annoy me.*

The future *holds a lot for me—working with my father, marriage, owning a car.*

I am best when *I'm good.*

I hate *school.*

I wish *I were at home right now.*

My father *is cool.*

I secretly *admire my father.*

My greatest worry is *that I graduate from high school.*

After the incomplete-sentence blanks, David was asked what he would do with three wishes. He wanted (1) "to be with Sharon [his girlfriend] whenever I want," (2) "to graduate from high school and go to college for a civil engineering degree," and (3) to be "rich off of life—and I don't mean just have money."

David's H-T-P drawings are shown in Figures 5.3, 5.4, and 5.5, and his Kinetic Drawing System drawings in Figures 5.6 and 5.7. As each drawing was completed, a series of inquiry questions was asked to describe and clarify the objects and actions within the drawing. (These are indicated by [Q]: but are not given in full.) David's verbatim responses to the questions were as follows:

*House:* [Q]: "This is the house I lived in for around 2 to 3 years when I was about 7 to 10 years old." [Q]: "This is a real house, a sketch of my old house in Tennessee." [Q]: "It is 7 years old." [Q]: "Me, my parents, and my sister live in the house." [Q]: "The house is in with a group of houses."

## 114 II. PROJECTIVE METHODS

**FIGURE 5.3.** House drawing (on the H-T-P) from David's comprehensive personality assessment battery.

**FIGURE 5.4.** Tree drawing (on the H-T-P) from David's comprehensive personality assessment battery. [Q]: "It's thinking about running into the

woods and going up to this huge cave—like an airplane hanger.”

*Tree:* [Q]: “It’s a treehouse in Tennessee—the same place as my house drawing.” [Q]: “It’s a huge oak tree—4 feet in diameter.”

[Q]: “Me and my friends built the treehouse. I have good memories of this treehouse.

It was in the backyard of my house on about 1 1/2 acres of land.”

*Person:* [Q]: “This is just something that I drew—it’s a nothing, just a body—a ‘he-she-it.’” [Q]: “It’s anywhere from 5 to 90 years old.” [Q]: “It lives alone—in the middle of the Alps. It’s a hermit—it doesn’t like society—it doesn’t get along with crowds.” [Q]: “In the Alps, it hunts, fishes—the usual.” [Q]: “It’s thinking, ‘Where’s my clothes? I’m freezing.’” [Q]: “It’s feeling cold—and lonesome. Even hermits have to show some feeling. It has a dog—but it’s been dead for six years.”

*K-F-D:* “This is a happy picture—everyone is having a good time. It is a picture of us in a river near our house in Florida. We go water skiing all the time. There is my father driving, my mother sitting behind him watching my sister ski, me sitting, and my sister skiing and saying ‘Wheeee!!’ My father is saying, ‘David, get me another beer.’ I’m saying, ‘Why do you want another beer? You just had one.’”

*K-S-D:* “This is my classroom with kids who are just doing their homework. There’s the teacher. She thinks she can control everything—she thinks she’s the dominant power. But that’s not true—we just follow her for fear of being suspended. I respect her ‘cause she’s the teacher—you should respect adults.”

From a content perspective, a number of social-emotional and behavioral hypotheses were generated as hypotheses that would need later, objective validation. Integrating the qualitative/content information for each projective drawing resulted in the following analyses:

*House:* It was interesting that David chose to draw a house that he had lived in for 2–3 years in Tennessee, when he was about 7–10 years of age. Structurally, the house looked like a typical house, except that it was hard to tell whether the driveway led to a door or a garage (one of these was missing, in any case). Chronologically, this was the period when David’s family was all together, although the parents were having

marital problems that led to their divorce when David was 10. Immediately after the divorce, David and his family had to move from this house in Tennessee to another state. Thus, the fairly typical structure of the house might suggest that David was secure with the nuclear family together. His comment about what the house was thinking (“It’s thinking about running into the woods and going up to this huge cave—like an airplane hanger”) might suggest that his

#### 5. Evaluation of Projective Drawings 115

**FIGURE 5.5.** Person drawing (on the H-T-P) from David’s comprehensive personality assessment battery.

**FIGURE 5.6.** K-F-D from David’s comprehensive personality assessment battery. parents’ fighting made him want to escape from the house. It also might suggest that he wanted to escape blame for his parents’ fighting, given his mother’s interview comment that David blamed himself for his parents’ divorce.

*Tree:* The tree drawing helped the examiner to continue the analysis of David’s “Tennessee” years, because he chose to draw a tree that was in the backyard of his house in Tennessee. The tree had a treehouse that was built by David and his friends, and David had good memories about this treehouse. The tree appeared to be strong, perhaps indicating that David used his peers as a support group while his parents were leading up to their divorce. However, there was some shading, a shaky baseline grounding the tree, and a barrenlooking branch, collectively indicating that this peer group did not provide David sufficient support to compensate him fully (in a social-emotional sense) for the turmoil at home. In addition, these characteristics may suggest that David might have perceived this peer support as being stronger than in fact it was. The generalized barrenness of the tree also might reflect David’s current emotional coldness, his social and interpersonal aloofness, and his lack of empathy toward others. This coldness might be accentuating his memories of Tennessee as “the last time I was really happy,” thereby allowing him to ignore his current difficulties through fantasy and to blame his current situation on his parents’ divorce and his move from his home in Tennessee.

*Person:* David’s person drawing provided insight into how he was really feeling and into his current self-concept. First of all, the

figure was “an it.” David did not feel as if he had an identity at the present time. He had conflict all around him—with his family, at the residential school, with his peer group, in the classroom. Furthermore, he had lost all sense of control over his own life: Everyone was making decisions for him, and he was caught in a downward spiral in which his attempts to reassert his control were causing the adults in his life to make more restrictive and more controlling decisions. Thus, David wanted to be all alone (“a hermit”), but this was a forced isolation. The hermit (David) did not like “society”—that is, having all the adults forcing him to conform to their societal rules and expectations—and he did not like having all the adults (both his parents and those at the residential school) working together to apply consistent expectations and consequences (the hermit “doesn’t get along with crowds”). It was a forced isolation also in the sense that David described the hermit as “lonesome—even hermits have to show some feeling.” His final comment about having a dog that had been dead for 6 years was very sad. It suggested that David really wanted to interact and share his feelings with others and that he wanted to be accepted, but the overall tone of the drawing indicated that this was a deep, underlying aspiration that would not soon be fulfilled. Finally, it was interesting that 6 years before drawing this picture, David was living in Tennessee and his parents were approaching their divorce. Perhaps, David had never emotionally recovered from that traumatic event.

*K-F-D:* David’s family drawing included all the members of his family, but the figures were incomplete and poorly formed and could not be differentiated from one another. This suggested that David felt a significant amount of conflict among the members of his family—conflict that he focused around his father. Indeed, the potential issue of and conflict around David’s father’s alcohol abuse was readily apparent: David had the father in the drawing asking for another beer and himself rejecting the request, saying, “You just had one.” Significantly, the mother in the drawing was depicted as somewhat passive, and as an individual

## 116 II. PROJECTIVE METHODS

**FIGURE 5.7.** K-S-D from David’s comprehensive personality assessment battery.

who often got in between David and his father to act as a buffer or intermediary. Finally, David's sister appeared to be the psychologically safest and healthiest one in the family—she was allowed to water ski behind the boat, free from the conflict within the boat, and was able to have fun.

It was interesting that David described the picture as a happy picture and as one taking place in Florida. No longer was David drawing pictures about himself or his family in Tennessee. He was now focused on his more recent experiences in Florida, and he still perceived his family as his father, mother, and sister despite his being placed in the residential program. The description of the picture as happy, despite the subtle indications of conflict and confrontation, might suggest that David perceived any family interaction that had only limited conflict as "happy." This, then, might provide a baseline as to the amount of interpersonal conflict actually in the home setting, and might suggest that David had not recently experienced a truly supportive family environment, at least while living in Florida.

On a more positive note, David did include himself within the family unit. He had returned from "the Alps," although his drawing did reflect a sense of isolation within the family unit. It might be that despite its faults, David felt most supported within his family, and that his perceptions of rejection and isolation were related more to peers, other adults, and school and community settings.

*K-S-D:* David's reaction to and need for control in the school setting were clearly apparent from his school drawing. His choice of the bird's-eye view, looking omnisciently over the classroom, with the teacher's head drawn more heavily than those of the students, suggested his need to be even more powerful than the teacher and to be more of an "overseer" or authority. And the fact that he did not identify himself in the picture almost suggested that the teacher could not force him to be at one of the desks, the way the other students were depicted. David verbally expressed this power theme and need when he stated that the teacher thought she could control everything in the classroom and that, from his perspective, she really could not. He then identified the powers that might control him most: the fear of being suspended (although he did

not appear to fear that in reality), and the more moralistic power or reason that students should respect teachers and adults because they *are* adults. (Once again, this was a power that did not seem to curtail David's resistance to authority figures.)

One interesting question not answered by the school drawing was the reason why David seemed to harbor these power issues. On the one hand, it might be that David's acting out and controlling behavior countered his feelings of powerlessness with respect to his LD and his inability to achieve academically in school. That is, he needed to feel in control of something, and his negative behavior and ability to anger and frustrate others were the things he could best control. Or, on the other hand, it might be that David had always exhibited this negative pattern of behavior, and that the behavior had affected his academic failure and his poor peer interrelationships and acceptance. Regardless of the answer, it was clear that the issues of power and conflict were well embedded in David's perceptions of school and family, and that these perceptions had to be considered in any intervention program that might hope to address David's social-emotional, behavioral, and academic needs.

### A Final Integration

Objective tests such as the PIC and the MAPI, and behaviorally oriented measures such as behavior rating scales and behavioral observations, are clearly important to the personality assessment process; they provide reliable and valid samples of the student's functioning intrapersonally, interpersonally, ecologically, and across multiple settings. From a diagnostic perspective, David was correctly placed in the residential setting, according to these objective and behavioral assessments. In addition, the psychoeducational assessments relative to his intellectual and academic functioning were critical in developing an appropriate educational program with reasonable demands and expectations.

From a therapeutic perspective, however, the addition of the projective tests, and the drawing tests in particular, provided a more in-depth cognitive-behavioral assessment that contributed to a fuller understanding of David's attitudes, attributions, expectations, and perceptions. Who would have expected

David to draw a figure in the Alps feeling lonesome and reflecting that it didn't like society and didn't get along with crowds, and what personality assessment measure could have elicited a feeling that descriptive and intense other than the projective drawings? Who would have predicted that David would have addressed his need for power and control so explicitly in response to the K-S-D inquiry? Clearly, no one. But these cognitions were what completed the diagnostic picture within a multitrait, multisetting, multimethod assessment approach. Furthermore, these cognitive beliefs and perceptions might be most instrumental in creating an intervention program that addressed both David's cognitions and his behaviors. Integrating the projective drawings with all the other assessments done with and for David resulted in a comprehensive personality picture. David had a long history of personality and adjustment difficulties, beginning with an interaction between his own behavior and his parents' perceptions of him being a child who "defied authority" at 9 months, and extending through a diagnosis of hyperactivity and numerous failure experiences at home and at school; parents with inconsistent disciplinary styles; a parental divorce and subsequent remarriage; a suspension from school; and a move to a residential school because of his behavior and active resistance. Qualitatively, David had serious feelings of negative selfconcept, deep feelings of inadequacy and vulnerability, and a great need to be accepted by others. Behaviorally, however, he had poor impulse control, and he lacked the social and conflict resolution skills to interact appropriately with peers. Thus, he was caught in a vicious cycle: He was rejected continually by peers and adults; he withdrew from them in anger and self-protection; he increasingly resented his rejection and isolation; and then he tried to interact again, only motivated by revenge, anger, and a self-fulfilling prophesy for another social failure. This vicious cycle had played itself out continually over the past several years, such that the behavioral pattern was now occurring many times per day, and David's behavior was more and more out of control. David now presented himself as someone with serious conduct disorder and delinquency problems and behaviors, and as one with little internal self-control or trust

in anyone who could help him to assert that self-control. David was not a good therapy prospect at the time, and unfortunately his parents were not either: They did not want to acknowledge their part in the broad, ecological problem, or their need to change their own behavioral styles and cognitive beliefs so that David could be fully reintegrated into the family system.

Therapeutically, the picture of David emerging from this assessment was quite distressing. It was, however, accurate. Within 2 months of the final diagnostic assessments, David's behavior was so far out of control that he was asked to leave the residential school. At last contact, David's parents were looking for a residential psychiatric setting for David, still disavowing their part and responsibility in the presenting problems and the needed comprehensive therapy and intervention process.

#### **SUMMARY**

This chapter has attempted to provide a comprehensive picture of projective drawing approaches: their historical and theoretical development, their psychometric and clinical strengths and weaknesses, and their use in the context of a comprehensive personality assessment battery. A case study has also been presented to exemplify the use of projective drawings in the comprehensive personality assessment process and their potential contributions to that process. There is no doubt that projective drawing techniques—indeed, projective techniques as a whole—continue to be controversial and questioned (Knoff, 1991). However, there is also no doubt that they continue to be discussed in graduate training, employed in the field, and integrated into clinical practice. Relative to the latter use, projective drawings need more sophisticated research attention and evaluation. However, they should be evaluated as part of the entire personality assessment process, not in an isolated and out-of-context way, and their use should be strategic and well considered. Projective drawings are

#### **118 II. PROJECTIVE METHODS**

not needed in most assessments of a referred child, but when they are used, they should be used correctly and with an eye toward an accurate understanding of the referred child and an appropriate and effective intervention and service delivery program. This is the bottom line for all personality assessment: not that we complete our tests with just a

better description or understanding of a child but that we implement intervention programs that have a positive, lasting impact on the children, families, and systems who are referred to us as needing emotional, behavioral, and other support.

## ACKNOWLEDGMENTS

I gratefully acknowledge the assistance of Dr. William Carlyon in helping to review the projective drawing literature, and for his precision and dedication in helping to develop the appendices. I also am grateful for the assistance of Carrie Finch and Amanda Denecke in helping to update the research in projective drawings for the revision of this chapter for the second edition of this volume.

## REFERENCES

- Abell, S. C., Horkheimer, R., & Nguyen, S. E. (1998). Intellectual evaluations of adolescents via human figure drawings: An empirical comparison of two methods. *Journal of Clinical Psychology, 54*, 811–815.
- Abell, S. C., Von Briesen, P. D., & Watz, L. S. (1996). Intellectual evaluations of children using human figure drawings: An empirical investigation of two methods. *Journal of Clinical Psychology, 52*, 67–74.
- Adler, P. T. (1970). Evaluation of the figure drawing technique: Reliability, factorial structure, and diagnostic usefulness. *Journal of Consulting and Clinical Psychology, 35*, 52–57.
- Aikman, K. G., Belter, R. W., & Finch, A. J. (1992). Human figure drawings: Validity in assessing intellectual level and academic achievement. *Journal of Clinical Psychology, 48*, 114–120.
- Albee, G. S., & Hamlin, R. M. (1949). An investigation of the reliability and validity of judgments of adjustment inferred from drawings. *Journal of Clinical Psychology, 5*, 389–392.
- Anderson, B., & Rallis, K. (1981). Relationship between Bender errors, emotional indicators and performance on Bender recall. *Perceptual and Motor Skills, 53*, 497–498.
- Baroff, G. S. (1957). Bender–Gestalt visuo-motor function in mental deficiency. *American Journal of Mental Deficiency, 61*, 753–760.
- Beck, H. S. (1959). A comparison of convulsive organics, non-conclusive organics, and non-organic public school children. *American Journal of Mental Deficiency, 63*, 866–875.
- Bender, L. (1938). *A visual motor gestalt test and its clinical use* (American Orthopsychiatric Association Research Monograph No. 3). New York: American Orthopsychiatric Association.
- Bensberg, G. (1952). Performance of brain-injured and familial mental defectives on the Bender–Gestalt Test. *Journal of Consulting Psychology, 16*, 61–64.
- Berman, S., & Laffal, J. (1953). Body type and figure drawing. *Journal of Clinical Psychology, 9*, 368–370.
- Bieliauskas, V. J. (1960). Sexual identification in children's drawings of human figure. *Journal of Clinical Psychology, 16*, 42–44.
- Billingslea, F. Y. (1948). The Bender–Gestalt: An objective scoring method and validating data. *Journal of Clinical Psychology, 4*, 1–27.
- Blaha, J., Fawaz, N., & Wallbrown, F. (1979). Information processing components of Koppitz errors on the BVMGT. *Journal of Clinical Psychology, 35*, 784–790.
- Bodwin, R. F., & Bruck, M. (1960). The adaptation

and validation of the Draw-A-Person test as a measure of self concept. *Journal of Clinical Psychology*, 16, 427–429.

Bradfield, R. H. (1964). The predictive validity of children's drawings. *California Journal of Educational Research*, 15, 166–174.

Breen, M. J., & Butler, L. (1983). Applicability of Bender–Gestalt emotional indicators for emotionally disturbed and nonreferred students. *Psychological Reports*, 52, 569–570.

Britain, S. D. (1970). Effect of manipulation of children's affect on their family-drawings. *Journal of Projective Techniques and Personality Assessment*, 34, 234–237.

Buck, J. N. (1948). The H-T-P technique, a qualitative and quantitative method. *Journal of Clinical Psychology*, 4, 317–396.

Buck, J. N. (1970). *The House–Tree–Person technique: Revised manual*. Los Angeles, CA: Western Psychological Services.

Burns, R. C. (1987). *Kinetic House–Tree–Person Drawings (K-H-T-P): An interpretive manual*. New York: Brunner/Mazel.

Burns, R. C. (1990). *A guide to family-centered circle drawings (F-C-C-D) with symbol probes and visual free association*. New York: Brunner/Mazel.

Burns, R. C., & Kaufman, S. H. (1970). *Kinetic Family Drawings (K-F-D): An introduction to understanding children through kinetic drawings*. New York: Brunner/Mazel.

Burns, R. C., & Kaufman, S. H. (1972). *Actions, styles, and symbols in Kinetic Family Drawings (K-F-D)*. New York: Brunner/Mazel.

Burton, A., & Sjoberg, B. (1964). The diagnostic validity of human figure drawings in schizophrenia. *Journal of Psychology*, 57, 3–18.

Byrd, E. (1956). The clinical validity of the Bender–Gestalt Test with children: A developmental comparison of children in need of psychotherapy and children judged well-adjusted. *Journal of Projective Techniques*, 20, 127–136.

## APPENDIX 5.1. A

### Methodological Summary and Analysis of Projective Interpretations for Selected Drawing Techniques

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>

Studies Prior to 1990

Adler (1970) DESC 216 N Psychiatric patients of Subjected 32 scoring Yielded one large factor — — P 2

varying diagnoses, criteria to factor analysis (Maturity of Body Image

18–65 yr old Concept) and three small

factors; diagnostic

categories were not

differentiated by factors

Albee & EPF 10 N Individuals with normal HFDs rated by 15 Ratings for clinicians split — — P 0

Hamlin to severe psychological clinicians in pairwise into two groups correlated NP

(1949) problems comparisons; clinicians .955 (reliability); rank-order

were asked to pick better- correlation between rank

adjusted individual from by case records and rank

each pair by ratings on HFD = .624

(validity)

Beck (1959) EPF 843 Y 805 normal children, 5–6 House drawing of H-T-P: Signif. diff. between MH — M? NP 1

yr old; 25 organic MH recognizable or not, children and normals in

children, 6–19 yr old; 13 bizarre appearance, details recognizabilities and omisionorganic

MH children present/or absent sions; organic, nonorganic

MH children were not diff.;

5- and 6-yr-old normals

signif. diff. on omissions

Berman & EPF 88 N Neuropsychiatric male Subjects body-typed (endo- Correlations between Body type — P 0  
Laffal patients morphic, mesomorphic, drawing and body types rating  $r =$   
(1953) ectomorphic) and were signif.—patients did .73  
compared on whether self- tend to draw themselves  
drawing matched actual as they were  
body type or not

Bieliauskas EPF 1,000 N Normal school children, Tested hypothesis that Males tended to draw — — NP 2  
(1960) 5–14 yr old; 50 male there would be no signif. males, females drew  
and 50 female for each diff. between boys and females; tendency  
age girls at various age levels increased with age  
in drawing figure as male  
or female

127

Bodwin & DESC 60 N Children 10–17 yr old; 30 Compared DAP scores High correlation between — — P 0  
Bruck (1960) male and 30 female based on 13 characteristics ratings and scales  
thought to be associated  
with self-concept to clinical  
ratings of self-concept  
based on interviews

Bradfield EPF 50 Y Chronic schizophrenic Percentage of total height Signif. diff. ( $p < .01$ ) — — P 1  
(1964) women represented by the head NP

Bradfield EPF 85 Y Children identified by Compared groups on size, No signif. diff. — — P 0  
(1964) teachers as acting out, location, and degree of im- NP  
withdrawn, overachievers, plied movement in DAPs  
underachievers, normals  
( $n = 17$  in each group)

Britain EPF 64 Y Middle-class nursery Four groups compared on Signif. diff. among groups 90% agree- R P 2  
(1970) school children family drawings: control, on such things as sequence ment  
control free play, play ther- of self figure, colors, area  
apy, play session designed of family figures, elaborato  
reduce self-esteem ation; predictable by emotional-  
evocative level  
induced

Burton & EPF 98 Y 49 schizophrenic females, Judges sorted random Trained clinicians were Interjudge — NP 1  
Sjoberg 49 normal females HFDs into two stacks to able to discriminate  $r = .79$   
(1964) determine hit rate and between schizophrenics  
identify their criteria for and normals, but no clear  
selection; also tested within agreement on characterismethod,  
anatomical dis- tics used to distinguish  
tortions, and a checklist of  
67 items

Carlson, DESC 59 N Psychiatric patients; 28 Factor analysis of 14 Two factors, Body Dis- Interrater — P 1  
Quinlan, schizophrenics, 11 specific features of HFD turbance (BD) and Sexual  $r = .79$

Tucker, & personality-disordered, 8 along with global ratings Elaboration (SE); BD signif.

Harrow neurotic, 13 mixed of sophistication and correlated with artistic

(1973) artistic skill; correlated skill and sophistication; SE

results with other measures correlated with pathological

of body image disturbance thinking

and psychopathology

(continues)

## 128 APPENDIX 5.1. Continued

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sub>a</sub>

Cassel, EXP 130 N White applicants for Compared H-T-Ps of appli- Same comparisons statisti-  $r > .90$  on — P 1

Johnson, & employment cants where examiner re- cally but not practically 16 pairs

Burns mained in room with those signif.

(1958) where examiner was not

present during drawing

Chase (1941) EPF 150 Y 50 schizophrenic, 50 Compared on Goodenough Schizophrenics had signif. — — P 1

hebephrenics, 50 normal drawings lower scores than normals;

male adults other variables, such as age,

level of education, duration

of psychosis, and mental age

did not differentiate

Craddick EPF 46 N Fifth-grade students and Compared DAP and self- Signif. relationship of size 95% agree- — P 0  
(1963) (23 pairs) college sophomores portrait on size, no. of and grades ( $p < .05$ ); more ment NP  
same-sex drawings, females drew opposite sex  
position on page. no. of on DAP (ANOVAs); most  
subjects' correct pairing maintained same position  
of pictures by judges on two drawings

Craddick EXP 86 N College males Subjects asked to draw a "Crazy" drawings signif. NA — P 0  
(1964) person, draw themselves, larger than other two,  
and draw a person the way suggesting expansiveness as  
someone who is "crazy" a perceived characteristic  
would of craziness

Craddick & DESC 200 N Male alcoholics Compared size of male Male drawings signif. NA NA P 1  
Leipold vs. female figure drawing smaller ( $t = 3.4, p < .01$ )  
(1968)

Datta & DESC 939 N Head Start children: 487 Sex differentiation of DAP Girls drew more sex- 49 out of 50 R? NP 2  
Drake (1968) males, 452 females aged 3 figure differentiated DAP, but agreements  
yr to 6 yr, 11 mo dependent on sex of  
examiner

Delatte DESC 38 NA 38 females, 16–18 yr old Compared femininity Small but signif. correlation Interjudge — P 0  
(1985) ratings of HFDs with self- (.31) between femininity of (on 38  
esteem on Rosenberg Self- HFD figures and self- drawings)  
Esteem Scale esteem  $r = .97$

129

Delatte & DESC 76 NA Male and female high Explored relationship be- No signif. correlations for NA — P 0  
Hendrickson school seniors tween size (height, width, females; signif. correlation  
(1982) area) of HFD and Rosen- between self-esteem and  
berg Self-Esteem Scale HFD width and area for  
scores males (only .35 for width);  
some restriction of range  
(mostly high scores) in selfesteem  
scores

DeMartino DESC 100 N MH males, mean age 19.5 DAPs analyzed on 39 char- Most drew male first; more — — NP 1  
(1954) yrs acteristics and which sex than 75% had mouth  
drawn first open, front view, feet/  
shoes, large head, nose,  
arms, generally poor proportions,  
standing; fewer  
than 25% had mouth  
closed, teeth, arms perpendicular  
to body

DeMartino EPF 74 N 37 homosexual MHs; 37 DAPs analyzed as above Homosexuals had signif. — — NP 0  
(1954) straight MHs more eyelashes and high  
heels

Dunleavy, DESC 141 NA Kindergarten students in Tested Koppitz scores of 42% of "nonready" chil- — R? P 2  
Hansen, & six randomly selected HFD in predicting school dren identified; 10% false  
Szasz (1981) schools representing three readiness of Kindergarten positives, so HFD was a  
SES levels children (measured by reasonable predictor  
Metropolitan Readiness  
Test)

Exner (1962) EXP 80 Y Psychoneurotics, charac- Compared groups on line Character-disturbed group — — P 0  
ter-disturbed, normals, pressure, sketchiness, used lighter pressure,  
and group experiencing shading movement, pro- more sketchiness, more  
induced fear conditions ( $n$  files, buttons, feet, hold- shading than the rest;  
= 20 in each group); attri- ing object, using bottom psychoneurotics used  
butes not clearly defined edge of page as baseline more unbroken lines.

Altogether, 6 of 10 variables  
differentiated pathological  
groups from each  
other and other groups;  
nothing really conclusive,  
though (*continues*)

130

## APPENDIX 5.1. *Continued*

Type Number  
of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Statistics reliability<sub>a</sub>  
 Fellows & EPF 278 N Enrollment of parochial Compared drawing and DCT scores not signif. Only DCT — P 1  
 Cerbus elementary grade school Drawing Completion Test diff. for males vs. females reported;  
 (1969) (DCT); six variables of up to age 12; signif. diff.  $r = .89$   
 H-T-P between males and between males and  
 females females on sex of figure  
 drawn first on HIP;  
 strongest correlations at  
 ages 11, 12, and 13  
 Fiedler & EPF 46 N 15 “improved” psy- Used characteristics of Unimproved patients re- 92% agree- — NA 0  
 Siegel (1949) chootherapy clients, 19 Free Drawing Test to ceived signif. lower ment  
 “unimproved” clients; all predict success or nonsuc- (poorer) scores on criteria  
 male veterans cess of therapy for formation of head  
 Fisher DESC 1,154 N White male adolescents Degree of nudity in figure Subjects drawing female — NA NP 1  
 (1961) jailed for delinquent drawings first had signif. more  
 behavior nudity; low numbers of  
 drawings contained nudity  
 overall  
 Fisher DESC 1,000 N Male felons Nudity in DAP, sex of Drew male figure first — NP 2  
 (1968) figure drawn first less often than reported  
 for normals in literature;  
 twice as many adolescents  
 indicated some nudity  
 compared to adults  
 A. P. Gold- EXP 39 Y Male and female atten- Analyzed pre- and post- No diff. in line pressure — NP 0  
 stein & dants in state mental HFDs for control and ex- (no. of carbons imprinted)  
 Rawn hospital perimental groups for in- or figure size; other quali-  
 (1957) creased aggression level tative diff. were signif.  
 by announcing they must (symbolic representation  
 work longer hours for of aggression)  
 same pay

### 131

H. S. Gold- EXP 23 Y Normal males Shading differences be- Signif. more shading Interrater — NP 0  
 stein & tween subjects’ HFDs following high-stress films  $r = .90$   
 Faterson following high- and low-  
 (1969) stress films  
 Goodman DESC 8 Y Obese vs. nonobese Rank-order correlations Signif. no. of high correla- NA R NP 1  
 & Kotkov women ( $n = 4$  in each between judges’ ordering tions, but also many  
 (1953) group) of drawing and ordering nonexistent correlations  
 based on scoring criteria  
 Graham EXP 23 N Graduate students: 12 Compared subjects’ initial Very few changes from — P 0  
 (1956) males, 11 female figure drawings and their first to second drawing  
 drawings following lectures  
 emphasizing negative  
 characteristics inferred  
 from certain  
 aspects of figure  
 Granick & EPF 571 N Male and female un- Compared sex drawn first Most drew own sex first, — P 1  
 Smith (1953) dergraduates on DAP to masculinity- but males signif. more  
 femininity scale of MMPI than females; no relation  
 between MMPI scores  
 and sex drawn first  
 Gravitz DESC 2,000 N Normal adults; 1,088 Drawing of same-sex, 76% drew same sex, 21% NA R Percentages 2  
 (1966) males, 912 females opposite-sex, or un- opposite, 3% undiff.; only  
 differentiated-sex figures more than twice as many  
 women drew opposite sex  
 Gravitz DESC 800 N Normal adults, 20–30 yr Sex of drawing No differences between NA R NP 2  
 (1967) old; half single, half married people and sinmarried  
 gles on whether drawing  
 was same-sex  
 Gravitz DESC 200 N 100 male and 100 female General characteristics of Males, 85% same-sex, NA — P 1  
 (1968) job applicants HFDs compared 15% opposite-sex; females,  
 67% same-sex; no diff. in  
 mean heights of figures  
 (continues)

### 132 APPENDIX 5.1. *Continued*

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>

Gravitz EPF 200 N Male and female adult job Compared groups on High-masculinity males, NA — NP 1

(1969a) applicants divided into drawing of same-sex 85% signif.; lowfour

groups: males with figures masculinity males, 78%;

high Masculinity scores both female groups were

on the MMPI; males with identical

low scores; females with

high scores; females with

low scores

Gravitz EPF 469 N 328 men, 141 women; Compared HFDs of mar- Males drew same sex sig- NA — NP 1

(1969b) 40–60 yrs old ried people and singles nif. more than females,

for percentage of same-sex but no diff. between male

drawings or female married people

or singles

Gravitz EPF 1,000 N 500 normal male adults, Compared on drawing of No diff. on fully clothed By joint R P 2

(1971) 500 normal female adults; opposite or same sex, de- same-sex figures; males agreement NP

20–50 yrs old, job appli- gree of nudity drew more fully clothed

cants females; males and females

no diff. on opposite- and

same-sex nude figures

Gray & EXP 88 Y College students: HSE took personality LSE group drew smaller Interjudge R P 1

Pepitone 25 high self-esteem battery and received a re- figures, but not signif.; agreement

(1964) (HSE), port stating that they made LSE had signif. more iso- 88–94%

25 low self-esteem (LSE), unusually favorable scores; lated and smaller figures;

38 controls LSE got opposite report; HSE more similar to concontrols

didn't get report trols

before doing HFD; groups

compared on figure size,

placement on page, emotional

tone, activity level

Green, EPF 55 Y 30 “feminine” boys, 4–10 Compared groups on sex Feminine boys drew girl NA M NP 1

Fuller, & yrs old (judged to like drawn first in DAP; sub- first (57%); controls drew

Rutley dressing in girls' clothes); jects matched on age, sex boy first (76%)

(1972) 25 control normals of children in family, and

marital status and SES of

parents

### 133

Griffith & EPF 76 Y 18 male mental patients Compared eye–ear em- Signif. more of eye–ear — — NP 0

Peyman selected on overemphasis phasis to whether clinical- group had ideas of

(1959) of eyes and ears on DAP; ly judged “ideas of refer- reference

58 controls ence” were present

Hamilton DESC 55 NA Male and female kinder- Compared HFD sample ANOVAs revealed signif. 3 judges' R P 1

(1984) garten through third-grade protocols across grades for diff. in growth of self- agreement

children in bilingual presence of indicators of concept during kindergar- by discuseducation

programs (three self-concept ten and remained stable sion

groups;  $n$ 's = 16, 17, through third grade

and 22)

Hammer EPF 40 Y Normal controls and eu- Looked at H-T-Ps before Signif. diffs. suggested — — NP 0

(1953a) genically sterilized males and after operation (con- more genital symbolism

( $n$  = 20 in each group) trols had other types of and feelings of castration

surgery) in sterilized group

Hammer EPF 400 N 148 black children and Clinical judgments of Blacks got higher overall Among — P 1

(1953b) 252 white children in first H-T-P on 6-point scale adjustment scores; whites three

through eighth grades from 0 (well adjusted) to 6 and blacks got closer judges,  $r$  =

(psychotic); whites and together as age increased .90

blacks compared

Hammer EPF 64 N All sex offenders: 31 adult Compared H-T-Ps: age of Pedophiles drew signif. NA — P 0

(1954a) rapists, 33 pedophiles tree, dead or alive, age of younger trees; pedophiles

people, male or female drew signif. older females

figures

Handler & EXP 57 Y 21 nonstressed control un- Compared DAPs for Hoyt– Signif. diff. Percentage — NP 1

Reyher dergraduates; 36 stressed Baron scoring of anxiety of agree-

(1964) undergraduates indices; stressed subjects ment bewere  
hooked up to a tween two  
polygraph in a small, judges  
dimly lighted room with ranged from  
experimenter looking over 67% to  
their shoulders 100%  
Handler & DESC 96 N Male college students Compared drawings of Auto yielded lowest mea- 67%–97% R P 1  
Reyher male, female, and auto- sures of anxiety; female agreement NP  
(1966) mobile on GSR and yielded highest measure  
Hoyt–Baron scoring in- of anxiety; low but signif.  
dicators of anxiety correlations between GSR  
and 10 of 23 graphic  
indicators (*continues*)

### 134 APPENDIX 5.1. *Continued*

Type Number  
of of Control Scoring General-  
Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izabeth<sup>a</sup>  
Heberlein & EPF 160 N Four groups of college Compared groups on Greater need for aggres- NA R P 1  
Marcuse females: those drawing Aggression, Heterosexual- sion in females who drew  
(1963) same-sex figures on two ity, and Abasement scales males  
DAPs; opposite-sex figures of Edwards Personal Prefon  
two DAPs; opposite- erence Schedule  
then same-sex DAP  
figures; same- then opposite-  
sex DAP figures  
Hiler & EPF 60 Y 30 adolescent psychiatric Compared on criteria Valid criteria for patholo- — R NA 1  
Nesvig patients; 30 normal scores of DAP by six psy- gy were bizarre, distorted,  
(1965) adolescents chologists and eight non- incomplete, and transparent  
psychologists to determine characteristics; criteria for  
valid criteria for differen- normality were happy extiation  
pression and no pathological  
characteristics present; nonpsychologists  
discriminated  
just as well as psychologists  
Holtz, EPF 146 N American college students; K-F-D: measured length, Signif. number of females — — P 0  
Moran, & 50 male, 96 female activity, strength of figures; drew self closer to mother  
Brannigan which parent figure was figure; signif. size diff. for  
(1986) closest to self figure females: father biggest, next  
mother, then self; only men  
signif. diff. between figures for  
strength/activity: self highest,  
then father, then mother  
Holzberg & EPF 108 Y Control group, 78 female Comparisons between Signif. diffs. on 27 Scored only M? NP 1  
Wexler student nurses; ex- groups on checklist of 174 variables items on  
(1950) perimental group, 38 drawing variables which  
schizophrenic women judges  
agreed  
Hoyt & EPF 112 N Female psychiatric Subjects divided into two Placement and size of Mean abso- — NP 1  
Baron (1959) patients groups according to high drawing signif. related to lute differor  
low Manifest Anxiety MAS score but not eight ence  
Scale (MAS) other indicators between

### 135

Jensen EXP 175 N College students randomly DAP administered by No diff. for male or NA R NP 2  
(1985) assigned into mixed male either male or female ex- female subjects according  
and female groups aminer; groups compared to sex of examiner when  
by sex on sex of first DAP administered in  
drawing mixed-sex groups  
Johnson DESC 103 N College students Compared on IPAT Anxi- Signif. relationship: more — — NP 1  
(1971) ety scale scores and anxi- upper left-hand placement  
ety indicator of upper left- as IPAT score higher  
hand placement of DAP  
Jolles DESC 8,500 N Children in Illinois public Drawing of same-sex Younger children drew NA R? NP 2  
(1952a) schools, 5–12 yrs old person (H-T-P) opposite sex more than  
older; girls drew opposite  
sex more often than boys

Jolles DESC 2,701 N Same as Jolles (1952a) Phallic-looking trees on More common in younger — R? NP 2  
 (1952b) H-T-P children; became more  
 psychosexually signif. with  
 age; more common in  
 females

Jolles & DESC 2,083 N Same as Jolles (1952a) Horizontal placement of Supported Buck's hypoth- NA R? P 2  
 Beck (1953) drawing as indicator of esis that psychological  
 intellectual control over center is to the left of  
 affect geometric center; normal  
 range varied with age

Jordan CS 1 N Child 9 yrs old with Analyzed HFD and Bender- Case study: Drawings NA NA NA 0  
 (1970) cerebellar disorder Gestalt Test drawing "floating in space"

Judson & DESC 240 N Psychology patients Foliage present or absent Signif. more bare trees Only 1 dis- NA NP 1  
 MacCasland (mixed diagnoses) on trees of H-T-P draw- drawn in winter by agreement  
 (1960) ings drawn over the four females but not males in 240  
 seasons of the year protocols

Kamano DESC 45 N Institutionalized schizo- HFDs rated by subjects Signif. more rated draw- NA NA P 1  
 (1960) phrenic women as to whether they were ing most like actual self;  
 most like: ideal self, supported idea that HFD  
 actual self, least-liked self; is a perception of drawer's  
 these compared to self- self  
 ratings on semantic differential  
 scales (*continues*)

### 136 APPENDIX 5.1. *Continued*

Type Number  
 of of Control Scoring General-  
 Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sub>a</sub>

Koppitz EPF 161 N 100 "good students," 61 Analyzed HFDs on 30 Five EIs were signif. more — — NP 1  
 (1966a) "poor students"; first and Koppitz emotional in- often present in poorsecond  
 grade (by Metro- dicators (EIs) student group, including  
 politan Achievement Test) poor integration of parts,  
 slanting

Koppitz EPF 152 Y 76 children from guidance Compared HFDs on 30 11 EIs signif. more fre- 95% agree- M NP 1  
 (1966b) clinic; 76 normal school EIs quent in clinic group ment on 25  
 children protocols

Koppitz EPF 62 N Guidance clinic children; Compared HFDs on 30 Asymmetry of limbs, pres- — M NP 1  
 (1966c) 31 identified as aggres- EIs ence of teeth, long arms, big  
 sive, 31 identified as shy hands were present signif.  
 more often in aggressive  
 group; hands cut off, no  
 mouth were more frequent  
 in shy group

Kotkov & EPF 55 (pilot); N 61 obese women, 40 nor- Compared groups on 43 32 signif. diff. measures in — M NP 2  
 Goodman 56 (experi- mal-weight women; all scoring items on the DAP pilot group led to seven  
 (1953) ment) matched on age, educa- combined signs that diftion,  
 IQ, marital status, ferentiated obese from  
 and employment status normal women's drawings;  
 primarily related to obese  
 using more of page

Kurtzberg, EPF 125 Y "Normal" inmates; DAP; female drawn first, Signif. more addicts drew NA — NP 1  
 Cavior, & inmates addicted to female larger than male female first; addicts drew  
 Lipton opiates females signif. larger than  
 (1966) males

Laird (1962) EPF 303 N 132 male introductory Compared HFDs on sex Percentage who drew NA — NP 1  
 psychology students; 100 of figure drawn first same-sex figure first: normale  
 alcoholics, mean age mals, 94.7%; psychiatrics,  
 45 yrs, mean education 84.5%; alcoholics, 81.0%;  
 11th grade; 71 male psy- (latter two groups signif.  
 chiatric patients, mean diff. from normals but not  
 age 42 yrs, mean educa- from each other)  
 tion 9th grade

### 137

Lakin (1956) EPF 49 N 25 third-grade children; Compared groups on area Supported hypothesis: — — NP 0  
 24 elderly residents of a used, figure height, and Drawings by aged more  
 home for aged centeredness of DAP as constricted, shorter, and

indicators of self-concept less centered  
and body image

Lehner & DESC 421 N College students; 229 Subjects' age compared to Men assigned slightly NA — Percentage 1  
Gunderson males and 192 females age assigned to DAP older ages to male and only  
(1948) female DAPs as own age

became greater; for  
females, the function was  
curvilinear—age of drawn  
figure increased with actual  
age until actual age of  
35, then decreased

Lewinsohn EPF 100 N Four groups of psychiatric Compared groups on Depressed patients had NA — P 1  
(1964) patients rated as de- height of DAPs signif. shorter figures

pressed male or female,  
or nondepressed male or  
female, by physicians  
( $n = 25$  in each group)

Lyons (1955) EPF 50 N Last 50 people the author On H-T-Ps, asked sub- Signif. correlation be- NA — NP 0  
tested at work jects to imagine the tree tween “scar” height and  
had been struck by light- age when “worst thing”  
ning and place an “X” occurred; not signif. for  
where this might have “best thing”

occurred; compared  
height of mark on tree to  
relative age at time when  
“worst thing that ever  
happened to you” occurred  
and “best thing”

Mabry CS 1 N Patient with malignant DAP Case study NA NA NA 0  
(1964) brain tumor

Marzolf & DESC 850 N College students List of 73 H-T-P charac- 29 items were signif. diff. Median and — NP 1  
Kirchner teristics: compared males across sex agreement

(1970) and females, and diff. be-  $r = 91.8$   
tween first and second  
drawing (*continues*)

### 138 APPENDIX 5.1. *Continued*

Type Number  
of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sub>a</sub>

Marzolf & DESC 760 N College males and females Analyzed presence or ab- Some signif. but low cor- Interjudge R P 2

Kirchner sence of 108 drawing relations; 347 signif. com- agreement

(1972) characteristics and com- parisons out of 3,672 cor-  $> 90\%$

pared to 16PF scores relations (17 16PF traits

× 108 drawing characteristics

× sex)

McHugh DESC 626 N Male and female students, Analyzed whether same Overall tendency to draw NA — NP 1

(1963) first through sixth grades sex was drawn first and same sex first; females

size of drawings drew female figure larger

than males drew male

figure

McHugh EPF 108 N Four groups ( $r = 27$  Groups compared on 23 Children with neurotic — — NP 1

(1966) each): two diagnosed with variables on HFD traits drew first figure sigadjustment

reaction of nificantly shorter and both

childhood (neurotic traits figures farther from botor

conduct-disturbed); two tom of page; neurotic

diagnosed with adjustment boys drew opposite sex

reaction of adolescence first more frequently than

(neurotic traits or con- duct-disordered childduct-

disturbed); both sets dren

matched on age and IQ

McPhee & EPF 264 Y 102 ED children (male Compared groups by sex No sex diff.; signif. diff. Five — P 0

Wegner and female), 162 normals; on stylization of K-F-Ds between ED and normals judges;

(1976) no ages reported on stylization  $r = 66-1.00$

Melikian & DESC 162 N 137 male and 25 female Percentage drawing same- Women started signif. more NA — NP 0

Wahab Moslem, African-born sex figure first on DAP often than males on opposite

(1969) college students sex (females, 48%; males,

18%); similar to literature  
with American samples  
Meyer, CS 22 N People undergoing a num- Case study comparisons Various conclusions; many NA NA NA 1  
Brown, & ber of surgical procedures on pre- and postoperative postoperative changes in  
Levine HIPs drawings noted  
(1955)

139

Michal- EPF 50 Y 25 individuals with normal H-T-P comparisons on six Only *line quality* differen- — — NP 0  
Smith EEG; 25 with abnormal variables from Buck's tiated

(1953) EEG scoring system

Modell DESC 28 Y 28 mental patients: 13 re- Subjective scoring of An illustrative case study Tested by — NA 0  
(1951) covered, 8 unimproved, 7 HFD on "body image is presented in full; recov- \_2 for signif.

uncertain maturation" and "sexual ered group showed diff. diff.

maturation" in patients in body image maturation

returning from a regres- and sexual maturation

sed state

Modell & CS 32 N Medical patients with Compared features of Made conclusions about NA NA NA 1

Potter hypertension, peptic ul- HFDs for diff. types of various personality charac-

(1949) cers, or bronchial asthma patients; qualitative de- teristics

scriptions and case study

provided for each type

presented

Mogar DESC 123 N Male psychiatric patients Looked at relationship be- No results with MA, thus Interrater  $r$  — NP 1

(1962) tween Manifest Anxiety supporting previous re- = .84-1.00

(MA) scores and Hoyt- search; several RCT vari-

Baron anxiety indicators ables had signif. correlaon

DAP, and same anxi- tion with DAP anxiety inety

indicators with Ror- dicators

schach Content Text

(RCT) scores

Moll (1962) DESC 269 N Normal college students Foilage or not on H-T-P Signif. no. of bare trees in Agreement NA NP 1

trees; drawings done by fall and winter drawings reached by

subjects during all four two judges

seasons in all but

three cases

Mostkoff & DESC 50 NA 25 male, 25 female stu- Determined interrater and Signif. agreement on nine Interrater — NP 1

Lazarus dents qualifying for Title test-retest reliability of an variables: self in picture, agreement

(1983) I; second-fifth grade objective scoring system evasions, arm extensions, 97%

for K-F-D elevated figures, rotated

figures, omission of body

parts (self, other), barriers,

drawings on back of

page (out of 20 variables) (*continues*)

140

#### APPENDIX 5.1. *Continued*

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>

Nathan EPF 72 Y 36 chronically obese chil- Compared on Good- Obese sample signif. — M P 1

(1973) dren, 36 controls; 7, 10, enough-Harris scoring of more global and unand

13 yrs old, matched HFD looking for diff. in differentiated than conon

sex, IQ, SES detail and sex differentia- trols; related to selfton

image/body image

Oas (1984) EPF 214 Y 100 adolescent psychiatric Tested Bender-Gestalt Discriminant-function — M P 2

patients, 114 "normal" Test and HFD ability to analyses of Bender-

adolescents from regular discriminate between im- Gestalt and HFD varand

special education pulsive and nonimpulsive iables indicated high

classes adolescents (Matching relation; HFD variables

Familiar Figures Test, Be- slightly better at dishavioral

Checklist, Im- criminating impulsives

pulsive Behavior Checklist from nonimpulsives in

used to discriminate each group; 79% of school

groups) sample correctly classified;

93% of hospitalized

sample correctly classified

Otterbacher, DESC 40 NA Male and female LD stu- Investigated relation be- All variables (age, nystag- Interrater — P 1  
 Haley, dents, 59–146 months old tween HFD and postro- mus, IQ, sex) yielded signif.;  $r = .94$   
 Abbott, & tary nystagmus age and nystagmus were  
 Watson as predictive of HFD  
 (1984) performance without IQ  
 and sex added into  
 equation; nystagmus was  
 still predictive with age  
 partialled out  
 Paine, EPF 24 Y 12 pediatric oncology Compared groups on Cancer patients' drawings NA M P 0  
 Alves, & patients, 12 general Koppitz HFDs according signif. smaller than those  
 Tubino pediatric surgery patients to size of drawing of surgery patient/student  
 (1985) and students group; implied increased  
 anxiety

141

Pollitt, CS 7 N Males with priapism Case studies used Bender– Concluded that a “blind” NA NA NA 1  
 Hirsch, & (prolonged erection) Gestalt Test and DAP investigator would find no  
 Money clear signs in drawings to  
 (1964) distinguish this group  
 Prout & DESC 100 NA Normal male and female Examined relation be- Modest but signif. correla- — — P 1  
 Celmer fifth-grade students tween indicators of emo- tions between achievement  
 (1984) tional conflict and nega- (SRA achievement test) and  
 tive affect on K-S-D and 6 of 10 variables, 26% of  
 academic achievement variance in achievement  
 accounted for by K-S-D  
 variables in stepwise  
 multiple regression  
 Resnikoff & EPF 75 N 25 organic brain-impaired, Compared HFDs on 17 Organics signif. diff. from Two scores, — P 1  
 Tomblen 25 schizophrenic, 25 indicators other two groups on five 91% agree- NP  
 (1956) neurotic individuals to six indicators: weak ment  
 synthesis, parts misplaced,  
 shrunken appendages, etc.  
 Rosen & DESC 98 N Male college students en- Nudity in DAP 64% completely nude, NA — Percentage 0  
 Boe (1968) rolled in weight-lifting 48% with penis; discussed only  
 class as very unusual finding  
 Saracho DESC 480 NA Random samples: 240 first Assessed interjudge, test- High reliabilities found: Interjudge R P 2  
 (1984) graders, 240 third graders retest, and split-half re- Embedded Figures, .90s;  $r = .91$  for  
 liability; tested relation- body concept, .50s; DAP  
 ship between Goodenough Goodenough, 80s; high  
 DAP and Children's Em- correlation with articulabedded  
 Figures Test and tion of Body Concept  
 Articulation of Body Con- Scale (.90s); low correlacept  
 Scale (measures of tion with Embedded Figfield  
 dependence and in- ures Test; concluded DAP  
 dependence) was good measure of field  
 dependence–independence  
 Schubert DESC 22 N Army enlisted men Compared three adminis- Revealed signif. linear Interrater — P 0  
 (1969) trations of DAP (male and trend toward poorer quali-  $r = .80$ s–  
 female) on a DAP quality ty; indicated a motivation .90s  
 scale (drawings once a deficit that must be considweek)  
 ered in studies where  
 more than one administration  
 is involved (*continues*)

142

#### APPENDIX 5.1. *Continued*

Type Number  
 of of Control Scoring General-  
 Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>  
 Short- EPF 15 N 8 boys, 7 girls; 47–57 DAP using Ayres and Reid Draw-A-Man and Draw- Interrater R? NP 0  
 DeGraff, mo old guidelines and estimated A-Woman scores signif.  $r$ s ranged  
 Slansky, & IQ scores (from WPPSI) correlated; Ayres and Reid from .93  
 Diamond scores signif. related to longer to .97  
 (1989) DAPT scoring system; IQ  
 did not signif. correlate to  
 any drawings

Sobel & EPF 40 Y 20 institutionalized male Compared K-F-Ds on 16 Only 3 of 16 traits showed R NP 1  
Sobel (1976) delinquents, 14–17 yrs traits signif. diff. between  
old; 20 normals from a groups; questionable ability  
public school, 15–17 yrs to diagnose delinquency  
old

Stawar & EPF 52 N Adolescent psychiatric Participants grouped into MMPI D and Sc scales signif. — NP 1  
Stawar inpatients between 13 overlapping diagnostic correlated with diagnostic  
(1989) and 19 yr old; 54% groups of disruptive, groups; no signif. results with  
female, 90% white, depressive, anxiety, and any KFD indicators  
10% black thought disorders; KFDs  
scored across 24 variables;  
select MMPI scales used  
(D, Pt, Pd, Sc)

Strumpfner EPF 81 N Psychotic inpatients; 45 DAP compared by age and Most variables showed sig- Interrater — P 1  
(1983) male, 36 female length of time since di- nif. negative correlations  $r = .85$   
agnosis; DAP variables in- with chronicity (length of  
cluded quality, height, sex illness); the longer subjects  
differentiation had been sick, the poorer  
the performance

Vane & EPF 662 N Kindergarten children di- Compared Goodenough Poor group showed signif. NA R? NP 2  
Eisen (1962) vided into three groups by Draw-A-Man on four signs more signs compared to  
teacher adjustment ratings and combinations other two groups  
of good, fair, or poor behavior  
on a 9-item scale (*continued*)

#### 143

Vroegh EPF 151 N Preschoolers rated most Sex of DAP No diff. based on degree of — NP 1  
(1970) masculine and most femi- gender; signif. diff. on permine,  
and least masculine centage drawing same sex  
and least feminine, by by biological sex (males,  
teachers 76%; females, 58%); only  
25% of figures sexdifferentiated  
at all

Weider & EPF 438 N 210 boys, 228 girls; 7–12 Compared HFD: Subject Girls drew own sex more NA R? NP 2  
Noller yrs old sex  $\times$  sex of first figure; often; girls drew own sex  
(1950) subject sex  $\times$  larger figure larger more often; younger  
drawn; location on page  $\times$  children placed drawing  
age; subject sex  $\times$  IQ  $\times$  closer to upper left quadfull  
face or profile rant more often

Weider & EPF 153 N Children 8–11 yrs old For HFD: Divided sub- Girls drew own sex first — R? NP 2  
Noller jects into upper, middle, more than boys; boys drew  
(1953) lower SES; compared no. own sex larger; boys  
of characteristics drawn, increased no. of charactersize  
of figures, same or istics drawn as SES inopposite  
sex drawn first; creased, especially in  
overt drawing compared same-sex figure; girls did  
to covert things learned too, but no signif. diff.  
from interview but not in depending on sex of figure;  
drawing responses in interview  
did not differ from drawing

Wildman EPF 60 N Patients from psychiatric Judges rated patient rec- More than twice as many On ratings — NP 0  
(1963) department of state hos- ords on degree of para- patients who had drawn of paranoia,  
pital divided in two noid pathology (little, joints were rated highly interrater  
groups according to moderate, high); com- paranoid compared to no  $r = .80$   
whether or not their pared across jointed and joint patients; more than  
knee/arm joints were nonjointed groups twice as many who didn't  
present in H-T-Ps draw joints were rated  
( $n = 30$  in each group) low compared to joint  
patients

Wisotsky DESC 490 N Black and white male in- Compared percentage Whole group drew male NA R NP 2  
(1959) carcerated alcoholics who drew male figure first figure first signif. more  
between whole group and often than normals; no  
normative group from diff. between blacks and  
literature; also between whites in sample  
blacks and whites (*continues*)

#### 144 APPENDIX 5.1. *Continued*

Type Number  
of of Control Scoring General-  
Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>  
Woods & DESC 138 N Eighth-grade students Looked at relationship be- Found signif. relationship; Reported — NP 0  
Cook (1954) tween detailing in draw- questioned use of detail- with \_2 siging  
of hands in HFD and ing in hands as a measure nificance  
level of proficiency in of personality variables only  
drawing  
Worden CS 1 N 11-yr-old male; recovered Administered DAP and Author used results to — NA NA 0  
(1985) from viral encephalitis KFD pre/post (4 mos) indicate usefulness of tools  
relative to therapy for in tracking therapeutic  
oppositional behavior changes  
Wysocki & EPF 100 Y 50 crippled school chil- Machover DAPs com- Eight variables showed — P 1  
Whitney dren (due to polio, cere- pared on 15 variables signif. diff.—large head,  
(1965) bral palsy, clubfoot, dis- opposite sex, large figure,  
located hips), 50 noncrip- placement, shading, prespled  
school children; all sure, paper rotation, and  
6–11 yrs old an area of insult  
Yates, EPF 34 Y 17 females referred to Developed clinical rating Only two signif. diff. (di- Combined M P 1  
Beutler, & psychiatric clinic as vic- scale of indicators of mensions of impulse con- ratings of  
Crago tims of incest; 17 referred potentially disturbed trol and quality of repres- two raters  
(1985) for other reasons functioning while compar- sive defenses); signif. diff.  
ing groups in *variability* of  
hypersexualization  
Studies from 1990 to 2000  
Abell, Von EPF 199 N Children, ages 5–15, Evaluated the Goodenough– Both drawing systems signif. Inter- Archive P 1  
Briesen, & whose drawings Harris and Koppitz HFD correlated with IQ tests; the rater *r*:  
Watz were obtained from scoring systems' ability Goodenough–Harris had a Koppitz:  
(1996) the archives of a to assess children's intellec- signif. higher correlation with .93;  
University Psychological tual abilities via the WISC-R Performance Scale G-H: .88  
Services Center WISC-R or Stanford–Binet than Koppitz  
Abell, EPF 200 N Males, ages 14–15, Raw scores on the Goodenough–Harris scores Inter- Archive P 1  
Horkheimer, whose drawings Goodenough–Harris and signif. lower than Verbal, rater *r*:  
& Nguyen were obtained from H-T-P were converted to Performance, Full Scale IQs; H-T-P:  
(1998) the archives of a standard scores and G-H and H-T-P about .91;  
University Psychological compared to WISC-R IQ equally correlated with IQ; G-H: .88  
145 Services Center scores G-H  
and H-T-P  
scores signif.  
correlated with each other  
Aikman, EPF 216 N Child and adolescent Assessed HFDs with IQ, Sign. but low correlations NA — P 0  
Belter, & psychiatric inpatients academic achievement, between Goodenough–  
Finch (ages 6–28 yr) and the BGT Harris HFD scoring and  
(1992) IQ/achievement; 58% of  
Ss were misclassified when  
HFD used to predict IQ;  
BGT did not improve  
accuracy of IQ prediction  
Groth- EPF 40 N 9 male and 31 female Assessed the concurrent Neither HFD or H-T-P Inter- — P 1  
Marnat & undergraduates, ages validity of HFDs and H-T-P scores were signif. indicators rater *rs* of  
Roberts 18–47 as measures of self-esteem of either measure of self- .88  
(1998) with the Coopersmith esteem; age, gender, and  
and Tennessee self concept artistic ability accounted Internal  
scales for 21–22% of variance *rs* ranging  
in HFD/H-T-P scores from .50 to  
.76  
Hackbarth, EPF 60 children, Y 30 sexually abused Evaluated KFDs using Comparison scored higher Kendall's W M P 1  
Murphy, & 60 mothers children (25 girls, 5 “Like to Live in Family” on scoring, indicating more indicated  
McQuary boys); rating procedure to positive adjustment as did high  
(1991) investigate discriminant comparison mothers; interrater  
30 unidentified validity mothers of sexually abused agreement  
comparison school children scored higher than between  
children their children; no signif. diff. raters  
between comparison mothers  
and children  
Handler & Lit. NA NA Primarily children and Reviewed research relative Interrater reliability good to See results NA NA 2

Habenicht Review adolescents to reliability, normative, excellent in all studies; test– (1994) validity outcomes as well retest reliability not as good; as cultural norms and KFD generally failed to implications for discriminate between groups personality assessment of people and very few signif. diff. found in drawings of well-adjusted and children with pathology; authors recommend more research using composites of indicators rather than single scoring variables (*continues*)

146

#### APPENDIX 5.1. *Continued*

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sub>a</sub>

Hibbard & EPF 134 Y 94 females and 40 males, Study compared HFDs via No sign. diff. in the presence Test–retest M P 0

Hartman ages 5–8 years Koppitz’s Emotional of individual indicators were ranged from

(1990) indicators to determine observed; abused children’s 84% to

65 children were alleged discriminant validity drawings appeared to show 100% ( $M =$

sexual abuse victims from more anxiety than the 94%) on

a clinic or therapeutic comparison children indicators

program; 64 were from first

comparison children from to second

general pediatric clinics drawing

Robins, EPF 32 N 14 females and 18 males, Study compared HFDs Goodenough–Harris scoring Interrater — P 0

Blatt, & 16 to 29 yrs old, who were and Robins Balance-Tilt of the HFDs from Time 1 to  $r = .95$

Ford hospitalized in a private, Scale after participants Time 2 were highly

(1992) intensive, inpatient were hospitalized for 6 correlated, independent of

treatment center for weeks and then 15 the gender of the figure

serious disturbances months after drawn; data indicated a main

effect of time with later

drawings more fully articulated

and differentiated

Taylor, CS 3 N 2 males, 1 female, 16–18 Analyzed KFD and Authors report that — — — 0

Kymissis, & yrs old attending a day Prospective KFD drawings additional information on

Pressman hospital for adolescents (the latter of participants’ the thoughts and feelings

(1998) with psychiatric and families in 10 yrs) of the adolescents was

substance abuse problems gained using the drawings

147

Tharinger EPF 52 Y Clinical group: 41 girls, Compared children’s DAPs Individual emotional Interrater Comp- P 1

& Startk 11 boys (86% white) in via Koppitz system on 30 indicators for both tests  $r$  for arison

(1990) grades 4–7 with mood, emotional indicators and failed to differentiate DAP = group

anxiety, or combined KFDs via Reynold’s system children with internalizing .92; for

mood/anxiety disorders on 37 indicators; disorders vs. normal KFD =

MANOVA used to controls; when indicators .80

Control group: 13 children evaluate scoring system’s summed, no sign. diff.

from same schools ability to discriminate among the four groups

presenting with no between the broad

DSM-III symptomatology diagnostic groups

West (1998) Meta- 12 Y Meta-analysis of 12 Used studies where effect Results indicated that Study by Study by P 1

analysis studies studies, published between sizes could be calculated; projective tests could study study

1986 and 1996, assessing projective reviewed discriminate between

the efficacy of projective included Rorschach, Hand children who were sexually

techniques to discriminate Test, TAT, KFD, HFD, abused and those who

between sexually and H-T-P, others were not abused sexually

nonsexually abused

children

Wu, Rogers, EPF 127 N 30 male, 97 female Assessed one week Using Buck (1948) scoring See Results — P 0

& Searight college students test–retest reliability of system, no sign. diff.

(1991) “House” drawing of the between two groups on

H-T-P with half of the 21 scoring indicators;

participants asked to draw sameness scores reached

a different house during over 50% agreement for

the second drawing 18 indicators and 75%  
for 5 indicators

*Note.* EPF, ex post facto design; DESC, descriptive “design”; EXP, experimental design; CS, case study; LD, learning-disabled; ED, emotionally  
disturbed; MH, mentally handicapped;

SES, socioeconomic status; HFD, Human Figure Drawing; DAP, Draw-A-Person Test; H-T-P, House-Tree-Person Test; K-F-D, Kinetic Family  
Drawing; K-S-D, Kinetic School

Drawing; MMPI, Minnesota Multiphasic Personality Inventory; GSR, Galvanic skin response; IPAT, Institute for Personality and Ability  
Testing; 16PF, Sixteen Personality Factor Test;

SRA, Science Research Associates; NA, not available; ANOVA, analysis of variance; (?), results questioned; R, random sample; M, matched  
sample; R?, possible random sample; M?,

possible matched sample; P, parametric statistics used; NP, nonparametric statistics used.

Generalizability rated on a scale of 2 (good), 1 (fair), or 0 (limited or none).

148

## APPENDIX 5.2. A Methodological Summary and Analysis of Projective Interpretations for the Bender–Gestalt Test (BGT)

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>

Studies Prior to 1990

Anderson & DESC 100 NA School children referred Explored relationship be- No signif. correlation — — NP 1

Rallis (1981) for ED or behavioral tween BGT recall scores found; children with sigproblems

in metropolitan and no. of emotional in- nif. number of emotional

school district dicators indicators did not recall

fewer BGT figures

Baroff EPF 76 N Twins institutionalized as Compared results on Pas- Similarities suggested — — P 0

(1957) mental defectives cal–Suttell scoring with feasibility of establishing

(endogenous) results from previous clinical norms for

research endogenous retardates

Bensberg EPF 322 N 161 organic brain damage Bender scoring of BGT Familial group signif. — M P 2

(1952) MH, 161 familial etiology more accurate in repro NP

MH; matched on MA and duction; reversals, parts

CA, mean MA about 6 repeated, and lines inyrs,

6 mos stead of dots signif. more

frequent in brain-injured

Blaha, DESC 74 N Male and female black Examined relationship 16% of variance in BGT — — P 1

Fawaz, & first graders among BGT errors, BGT performance due to pre-

Wallbrown time, DAP scores, DAP processing and central

(1979) time, Matching Familiar processing with general

Figures errors & time, IQ partialled out; conand

Slosson IQ ceptual tempo accounted

for little variance

Billingslea EPF 150 Y 100 male psychoneurotic Compared groups on 38 Equivocal results NA R? NP 0

(1948) patients, 50 normal males factors in scoring and in-

(all soldiers) terdrawing reliability

Breen & EPF 59 Y 30 students diagnosed Compared groups on 12 No signif. diff. — — P 1

Butler ED; 29 normal controls Koppitz emotional in-

(1983) (ages 7–11 yr) dicators on BGT

149

Byrd (1956) DESC 400 Y 200 children in need of 15 interpretation factors About half of factors — M? NP 1

psychotherapy; 200 chil- developed by Hutt on signif. at each age level;

dren judged well ad- BGT groups compared by pattern differed for these

justed; 8–16 yr old age level factors

Chorost, EPF 68 N Children under 18 who Compared BGT rotations Correct prediction of — — NA 1

Spivack, & had EEG and a scorable to whether EEG was nor- abnormal EEG by rota-

Levine BGT mal or abnormal tions in 65% of cases; not

(1959) a great increase over

other clinical signs

Curnutt EPF 50 Y 25 males in Alcoholics An- Compared BGT scored Signif. diff. between mean — M P 0

(1953) onymous (AA); 25 control using Pascal–Suttell scores; signif. indicators

males matched on age, method (scored by author, included rotations, count-

SES, occupation, educa- who knew which group ing dots; higher scores

tion each case was in) (poorer performance) in

AA group

Fabian DESC 692 Y 106 children with a vari- Comparisons, descriptions Established developmental — — NA 1

(1945) ety of psychological prob- of BGT rotations and nature of rotation of horilems

who were all “re- reversals zontally oriented figures

tarded" in reading; 586 to the vertical; persistent normal school-age children - "verticalization" concluded to be a clinical sign of problems

Fuller & EPF 270 Y Normals, ED, schizo- Rotation of BGT Signif. diff. between each — — P 2

Chagnon phrenic ( $n = 90$  in each drawings pairwise group comparison; also diff. depending on original orientation of figure, or figure-ground

Gavales & EXP 80 Y 40 undergrads with high Subjects divided into ex- Signif. greater diff. in size — R P 1

Millon Taylor Manifest Anxiety experimental (anxiety- in high-TMA induced- (1960) (TMA) scores; 40 un- inducing tasks) and con- anxiety group undergrads with the lowest TMA scores (out of 195) and recall BGT figures compared (*continued*)

## 150 APPENDIX 5.2. *Continued*

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Statistics reliability<sup>a</sup>

Gobetz EPF 393 Y 108 neurotic, 285 normal Two BGTs from each sub- Result was a final scoring — — NP 1

(1953) 118 male adult veterans (validation study) controls compared using discriminated consistently

64 neurotic, 54 normal objective scoring system between groups on both

adult male nonveterans with 82 general categories test and retest; test-retest

(cross-validation study) and 312 specific signs reliabilities in the high

.60s did not corroborate

Hutt's "neurotic syndrome";

BGT recommended

as a screening

device and supplement to

other instruments, not for

elaborate interpretation

Goldberg EPF 45 Y White male children: 15 Looking for diff. among Signif. diff. between nor- — M P 0

(1956-1957) schizophrenic, 15 MH, 15 groups on BGT according to males and other two groups;

normals; ages 11-16 to Pascal-Suttell scoring concluded that objective

scoring in conjunction

with qualitative analysis

effective for clinical

diagnosis

Goldfried DESC 80 N College undergrads; 40 Rated BGT drawings on Generally negative findings - NA — NP 1

& Ingling male, 40 female descriptive semantic scales; no "universal" sym-

(1964) differential scales (e.g., semantic meanings among

kind, strong, fast) individuals

Goodstein, EXP 114 N 54 male and 60 female Exp. 1—effect of serial Designs differed signif. on 86.3% — P 1

Spielberger, undergrads position on BGT design ease of recall; recall re- 91.8%

Williams, & recall; Exp. 2—effect of later to both difficulty agreement

Dalstrom where more difficult design- and serial position; no

(1955) signs placed in presentation order - diff. between males and

presentation order on recall females

Griffith & DESC 1,000 N Matched neuropsychiatric Compared no. of BGT Categories with signif. — NA NA 2

Taylor (1960) patients; mean age 35.5 rotations to diagnostic more rotations were

years category chronic brain syndrome

and mental deficiency

## 151

Griffith & EPF 213 Y 56 neuropsychiatric patients - Compared no. of BGT Signif. fewer rotations — — NP 1

Taylor (1961) patients; 157 control rotations for regular when blank paper preneuropsychiatric

patients admission and paper presented lengthwise instead

presented lengthwise of vertically

Guertin DESC 100 N Male and female mental 41 BGT scoring character- Six factors: propensity for — — P 0

(1952) patients: organic brain- istics from Billingslea curvilinear movement,

injured, nonpsychotic, and method factor-analyzed poor reality contact, schizophrenic

less execution, constriction,

poor spatial contiguity,

unidentified

Guertin DESC 100 N Male and female schizo- 46 BGT scoring variables Six factors: unstable clo- — — P 0

(1954a) phrenics and nonschi- factor-analyzed sure, curvilinear distortions, propensity of curvilinearity, fragmentation, irreg. propensity of curvilinear movement, experimental dependence; suggested using clusters of variables in scoring rather than single variables  
 Guertin DESC 27 N Organic brain-diseased 100 items as above factor- Three factors: curvilinear — — P 0  
 (1954b) males analyzed distortions, spatial disability, construction and feelings of inadequacy  
 Guertin DESC 37 N Male schizophrenics BGT scored on 100 items, Four types of schizo- — — P 0  
 (1954c) factor-analyzed phrenic performance factored out  
 Guertin DESC 30 N Male schizophrenics 100 items as above factor- Four types of schizo- — — P 0  
 (1955) analyzed phrenic performance factored out  
 Hain EPF 101 Y Patients in Compared groups on a High scores did differ- — — NP 1  
 (1964) neuropsychiatric ward; 20 scoring system designed tiate, but low scores did brain-damaged, 38 psy- to differentiate brain dam- not predict lack of brain chiatric, and 20 controls aged group from others damage not diagnosed in either category (*continues*)

152

#### APPENDIX 5.2. *Continued*

Type Number  
 of of Control Scoring General-  
 Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>  
 Halpin EPF 30 N 15 brain-damaged chil- Compared no. of rotations No signif. diff. 96% agree- M P 1  
 (1955) dren; 15 familial retar- on BGT ment NP  
 dates; ages 7–13 yrs  
 matched on CA, MA, IQ  
 Hammer EPF 40 Y 20 males undergoing Numerous variables Many variables showed — — NP 1  
 (1954b) sterilization; 20 males un- selected for each drawing signif. diffs.  
 dergoing other surgeries indicative of castration  
 anxiety before and after  
 surgery  
 Hanvik & EPF 44 Y 20 patients with lesions in Compared groups on no. Dominant and nondomi- — — P 0  
 Andersen dominant hemisphere; 24 of BGT figures recalled nant groups not signif.  
 (1950) patients with lesions in from memory and no. of different, but signif. more  
 nondominant hemisphere; rotations rotations than control  
 control group of patients group  
 with low back pain  
 Helgert DESC 120 Y 60 severely ED children; Compared traditional and Interrater reliability for (?) — P 1  
 (1985) 60 “normals”—referred digital computer methods computer, .99+; interrater  
 not placed of scoring BGT ter reliability for traditional,  
 .70s; small but signif.  
 correlations between  
 methods; computer  
 method accurately placed  
 61.7% of subjects; traditional  
 method accurately  
 placed 70% of subjects  
 Hellkamp EPF 180 Y Psychiatric patients, Investigated ability of Signif. correlation be- — M P 2  
 & Hogan organic and nonorganic Hutt BGT scoring system tween IQ and BGT errors  
 (1985) (functional) etiology to diff. organic from func- for either group; no actional  
 group as a function ceptable percentages of  
 of IQ accuracy scores for groups  
 across IQ ranges

153

Hinkle EPF 60 Y 30 institutionalized male Compared BGT error Signif. diff. in error scores — M P 1  
 (1983) delinquents, 30 nondelin- scores; tested a cutoff cri- between groups; cutoff  
 quents terion for discriminating criterion yielded 35% false

delinq. from nondelinq. positives and 47% false adolescents negatives; indicated need for further research

Holmes & DESC 76 NA Male and female college Investigated amount of Not a consistent factor; NA — NP 1  
Stephens students edging on BGT. Memory very little edging found; (1984) for Design, and DAP signif. *Q* indicated *lack* of consistency

Hutt & EPF 180 Y Carefully screened normal Compared on Hutt Signif. diff. in mean scores; Interrater — NP 1  
Monheit male students (13–16 yrs adaptation of BGT scoring the configuration score (1985) old) and children di- for disturbed adolescents agnosed ED signif. differentiated groups

Jernigan EPF 779 N Patients with central ner- Rotation of BGT figures Three separate studies: — — NP 2  
(1967) 206 vous system, psychiatric, compared on type of Rotators were older, had P 247 physical, and unknown patient, education, direc- lower IQ, had less educaproblems tion of rotation tion; 45% rotated in clockwise direction

Johnson EPF 50 N 25 psychiatric patients Constriction defined using *t* test yielded signif. diff.; NA — P 0  
(1973) with and 25 without con- less than half of page to constriction indicated dextricted BGT figures, re- complete figures; com- pression, but low rate of spectively pared groups on MMPI occurrence detracted from Depression scores its usefulness

Koppitz EPF 51 Y Elem. students (first Group comparison on 7 of Signif. discrimination in 93% agree- — NP 1  
(1958) through fourth grades) di- 20 scoring categories that cross-validation group ment on 14 vided into good and poor tested signif. on a first protocols achievement groups (read- sample and cross-validated ing, writing, spelling) on second sample

Koppitz DESC 1,055 N School children; 5 to 10½ Blind scoring using Kop- Normative table with — R? NA 2  
(1960) yrs old, kindergarten- pitz approach mean scores SDs by age fourth grade groups and grade

Koppitz EPF 384 Y 103 brain-damaged elem. Compared groups on Ben- Signif. more “poor” BGT — M NP 1  
(1962) students; 5–10 yrs old; der scoring system scores in brain-damaged 281 normals group ( $p < .001$ ) across ages (*continues*)

154

## APPENDIX 5.2. *Continued*

Type Number  
of of Control Scoring General-  
Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>

Lachmann EPF 120 Y 40 reading-retarded elem. Compared groups on BGT Reading-disabled group — M NP 2  
(1960) students; 40 behavior dis- scores (Pascal–Suttell) tended to produce more ordered normal readers, distortions, but diff. not 40 controls (no problems); quite signif. all normal IQ, matched on age

McCormick DESC 40 NA Matched ED adolescents Examined ability of emo- Correlations between — NA P 1  
& Brannigan 12–17 yrs old from day tional indicators on BGT BGT and Devereaux in- (1984) treatment and residential to discriminate behavioral dicated some signif. corcenters indicators of Devereaux relations for individual Adolescent Behavior signs; overall BGT with- Rating Scale drawal and anxiety signs did not signif. correlate; caution indicated in using BGT to draw implications about behavior

Moore & DESC 452 Y 150 LD, 189 educationally Compared normals to Nonhandicapped com- — — P 1  
Zarske disadvantaged, 113 non- 1974 Koppitz and SOMPA pared favorably with Kop- (1984) handicapped Navajo chil- norms for BGT; compared pits norms and SOMPA dren; ages 6–11 yrs BGT scores across groups norms; nonhandicapped by age and sex and LD groups showed signif. diff., with nonhandicapped having lower error rates across age and sex

Mosher & EPF 262 Y 142 brain-damaged, hospi- Compared ability of two Each discriminated statis- On 30 pro- — P 1  
Smith (1965) talized veterans (BD); BGT scoring systems tically, but not at clinical- tocols, in- 120 nonbrain-damaged (Peek–Quast and Hain) in ly useful level terscorer

controls (NBD) discriminating BD from  $r = .84-.86$   
NBD patients

155

Neale & DESC 195 NA Kindergarten students fol- Explored relationship be- BGT performance was — R P 2  
McKay lowed through fourth tween kindergarten BGT predictive of overall  
(1985a) grade performance and later school achievement; major  
reading and writing per- error categories (e.g., cirformance  
(Neale analysis cles for dots, distortion)  
of reading abilities and more predictive than total  
developed handwriting errors; only predictive of  
analysis system) one reading or writing  
variable (reading comprehension  
at end of first  
grade)

Neale & DESC 200 NA School children 5–7 yrs BGT (Koppitz) errors de- Reported percentage of For three R P 2  
McKay old scribed; item difficulty each error type across scores,  $r =$   
(1985b) analyzed figures; interjudge agree- .90s  
ment across error categories  
in low .90s for total  
error scores ( $r$ 's ranged  
from .71 to .98); angulation  
errors were the most  
frequent

Peek (1953) EPF 150 Y 75 neuropsychiatric Compared groups on fre- 27 variables were signif. — — NP 0  
patients who had certain quency of list of 40 ex- diff. between groups  
characteristics on BGT ternal characteristics of  
Figure 5; 75 randomly behavior and personality  
selected patients from and complaints  
same population

Quast EPF 100 N 50 suspected brain- Compared groups on 17 10 of 17 signif. dif- — M? NP 1  
(1961) damaged psychiatric attributes of BGT draw- ferentiated groups; low inpatients,  
50 ED psychiat- ings tercorrelations between  
ric patients; all 10–12 yrs these 10 differentiating  
old attributes

Schulberg DESC 106 N 15 neurotics; 41 functional Wanted to see what con- Not much difference by NA — P 0  
& Tolor psychotics; 15 acute notative meaning psy- patient type  
(1962) organic psychotics; 45 chiatric patients attached  
with character disorder to BGT drawings using  
semantic differential scales (*continues*)

156

## APPENDIX 5.2. *Continued*

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sup>a</sup>

Story (1960) EPF 60 Y Control, white male elem. Various characteristics of Most characteristics showed Interscorer R P 1  
and second. school BGT drawings signif. diff. including more  $r = .99$  NP

teachers; exp., male in- counting of dots,  
patient alcoholics ( $n = 30$  nonintersections, “liquid”  
in each group) responses on Design 6

Suczek & DESC 48 N Matched college students Subjects asked to free Compilation of more fre- Disagree- NA NA 1

Klopfer (male and female) associate to BGT figures quent symbolic associa- ment re-  
(1952) tions to each of the fig- solved by  
ures discussion

Taylor, DESC 652 NA Statewide school children, Subjected elements of Variance associated with — R P 2

Kaufman, ages 5–11 yrs Koppitz scoring of BGT to age for total sample was

& Partanio multiple regression by age only 35% (for 9- to 11-yr

(1984) olds, 3%); concluded BGT

scores not valid developmental

indicator beyond

age 8

Thomas DESC 66 NA Children with low scores Assessed relationship be- BGT errors positively — — P 0

(1984) on horse puzzle of WISC-R tween ratio of horse score correlated to “horse ratio”

Object Assembly and other three Object (.35)

Assembly puzzles and

BGT errors

Tolor (1957) DESC 50 N Matched male and female Had patients match Ror- Tried to establish asso- — NA NP 0  
neuropsychiatric patients schach blots with BGT cational stimulus value of  
(Air Force pop.) design that best repre- the BGT figures; found  
sented or stood for each that some produced signif.  
blot; rated them on type better associative reof  
response, or quality sponses than others

### 157

Tolor (1958) EPF 54 N Three groups ( $n = 18$  Compared groups on abil- Some signif. diff. but not NA M? NP 0  
each): Character-behavior- ity to reproduce BGT de- of clinical value  
disordered, schizophrenic, signs from memory after  
organic brain-injured; all standard testing  
matched on IQ and age

Tolor (1960) DESC 68 N College undergrads; 41 Ratings of BGT designs 79 out of 180 \_2 com- NA R NP 2  
males, 27 females on descriptive semantic parisons of item ratings  
differential scales between pairs of figures  
were signif.; connotative  
meanings for each were  
different; no sex differences

Trahah & EPF 93 N Male and female students, Investigated relation be- No signif. correlations Determined R? P 2  
Stricklin kindergarten- tween 15 proposed in- by \_2 (?) NP  
(1979) fourth grade dicators (emotional) on  
BGT and acting-out behavior  
in class as rated by  
teachers

Wagner & EPF 50 Y 25 children with organic Clinicians (MDs, PhDs, No signif. diff. between — — NP 1  
Murray brain damage; 25 normal MAs) made diagnoses professional groups' high  
(1969) controls (brain-damaged or not) degree of correct diagbased  
on BGT and design noses  
reproduction test

Wallbrown DESC 84 NA Matched reading-disabled Examined test-retest Pearson correlation was Judges re- NA P 1  
& Fremont children referred from reliability of Koppitz BGT .83 for total error scores solved dif-  
(1980) one school district scoring (12- to 24-day in- ferences by  
terval) discussion

Wiener EPF 822 N 822 children from a Johns Compared BGT scores to Some variables were sig- For 50 pro- M P 2  
(1966) Hopkins study on pre- neurological data to see nif.: ability to copy angles tocols, immaturity,  
ages 8–10 yrs; whether BGT could pre- and curves, tendency to terrater  
matched sample (race, diet minimal brain dam- produce distortions  $r = .92$   
sex, SES) age for premature and  
full-term groups (*continues*)

### 158 APPENDIX 5.2. *Continued*

Type Number

of of Control Scoring General-

Authors design subjects (Y/N) Population sampled Variables Results reliability Sampling Stats izability<sub>a</sub>

Wright & DESC 86 NA 55 boys, 31 girls; ages Compared VMI and Kop- Partialing out variance — — P 1  
DeMers 6–11 yrs; referred for pitz scoring of BGT on assoc. with IQ yielded signif.

(1982) learning or adjustment ability to predict academic correlations between VMI

problems achievement (WRAT) and WRAT Spelling and

BGT and WRAT Reading

and Spelling; absolute

magnitude was very small;

added little to prediction

beyond measure of general

ability (WISC-R)

Zolik (1968) EPF 86 Y Two groups: delinquents Compared on BGT scored Scores signif. diff. (de- — M? P 1

and nondelinquents by Pascal-Suttell method linquents higher) on all

matched on age and IQ but Design 8; each item

(roughly) also further analyzed for

specific characteristics

Studies from 1990 to 2000

Maloney & EPF 240 N Subjects (ages 15–65 yr) Assessed each subject Variables correlated with NP — P 0

Wagner separated into five age with the BGT, WAIS, age-related changes; most

(1991) intervals Hand Test, Rorschach of signif. variance attributed

to intellectual factors

Nyfield & EPF 46 N Male adolescents in a Investigated BGT errors Confirmed major hypothesis; Interrater  $r_s$  M: P 1

Patalano residential treatment that were valid or invalid also found that BGT errors for BGT Ss split

(1998) center for behavioral indicators of organicity of organicity had signif. ranged into 2 disorders hypothesizing that these correlations with other like from .88 groups errors were separate and measures (Block Design, to .94 distinct from Koppitz's Bender Recall); BGT Developmental scoring emotional indicators did not errors; also collected sign. correlate with WISC-III Block Design, Connors, nor did BGT Connors Parent Rating organicity indicators Scale, Bender Recall data

*Note.* EPF, *ex post facto* design; DESC, descriptive "design"; EXP, experimental design; LD, learning-disabled; ED, emotionally disturbed; MH, mentally handicapped; MA, mental age; CA, chronological age; DAP, Draw-A-Person Test; SES, socioeconomic status; MMPI, Minnesota Multiphasic Personality Inventory; SOMPA, System of Multicultural Pluralistic Assessment; WISC-R, Wechsler Intelligence Scale for Children—Revised; VMI, Test of Visual Motor Integration; WRAT, Wide Range Achievement Test.

Generalizability rated on a scale of 2 (good), 1 (fair), or 0 (limited to or none).

The sentence completion technique grew out of one of the earlier approaches to psychological assessment, the word association technique. Using partial sentence stems to stimulate the verbal production of a client rather than single words, A. F. Payne (1928) devised the first sentence completion instrument. He hoped to overcome the limitations of simple word association, which were primarily the cultural and structural biases being experienced with that procedure. Although these biases no doubt affect the sentence completion technique also, the sentence completion is far more capable of providing a proper degree of set for the responses to be elicited than is word association or any other projective procedure. This makes it practical as an assessment instrument for investigators who wish to study personal attitudes or adjustment related to specific situations.

#### BRIEF HISTORY AND DISCUSSION OF THE TECHNIQUE

The sentence completion has had many uses since Payne's initial creation, ranging from its use as a measurement of adjustment in certain settings, including school settings, to measuring the effects of conditions upon performance, effects of interventions upon mental health, and the lasting effects of trauma over time. By midcentury, the sentence completion had become a popular measurement approach in countries across the globe, including Japan, China, Germany, Denmark, India, Taiwan, France, and others (see, e.g., Agesen, Brun, & Skovgaard, 1964; Derichs, 1977). Clearly its most popular international use, including that of the United States, has appeared to be the assessment of attitudes and degree of adjustment of persons to specific situations, conditions, and settings (see, e.g., studies of Costin & Eiserer, 1949; Irving, 1967; Sanford, Adkins, Miller, & Cobb, 1943; Wilson,

1949).

Since 1976, there have been 864 studies reported in the psychological literature using the sentence completion. These studies cannot be reviewed here as that is not the purpose of this chapter. Nevertheless, it may be useful to know at least how the sentence completion has been most recently employed. For this purpose, the author briefly

159

## 6

### The Sentence Completion as a Tool for Assessing Emotional Disturbance

RUTH ADLOF HAAK

analyzes the functions of the last 100 studies involving the sentence completion that have been reported for this past decade (the 1990s) in Psychinfo (American Psychological Association, 1967–2000). (See Table 6.1.)

It is rather a surprising trend to see that the sentence completion has become an important research instrument in the burgeoning fields of neuropsychology and linguistics.

This, of course, exceeds its past history as primarily an instrument measuring personal adjustment. Even in its use as a measure of personal functioning, the sentence completion is now often used as an outcome measure. Those familiar with the sentence completion will realize the difficulties in this use—the lack of reliability and validity for many sentence completion measures, the highly clinical nature of establishing clear outcome targets expected on the sentence completion, and so forth. And yet, apparently due to its extreme degree of face validity (it is hard to interpret “I . . . hate school” as anything but rejection of the school experience), the sentence completion seems to be making itself useful more and more in a wide range of studies.

The topics under “personal functioning” that most studies of that type appear to be measuring by use of the sentence completion are illustrated in Table 6.2.

Only three studies in this whole decade appear to be centered directly on clinical analysis itself in the use of the sentence completion in the assessment of clients. This does not indicate that the sentence completion

has become solely a research instrument (in spite of the surprising degree to which this is happening). Clinical psychologists are “out there,” evaluating clients on a daily basis, writing reports, and meeting in conferences; they seldom have or take the time to report their activities to scholarly journals. It is a safe guess that the clinical use of the sentence completion still far outnumbers the research uses of it; neverthe-

## 160 II. PROJECTIVE METHODS

### **TABLE 6.1. Uses of the Sentence Completion Reported in the Last 100 Studies for 1990–2000**

Studies of the sentence completion technique itself 12%  
 Neuropsychological studies using the sentence completion 14%  
 Studies of language using the sentence completion 18%  
 Studies of personal functioning 56%  
 Exploratory 17%  
 Correlational 25%  
 Outcome 14%

### **TABLE 6.2. Personality Topics Investigated 1990–2000 Using 56 Sentence Completion Studies**

Exploratory  
 Personal development by age  
 Personal development of drug users  
 Personal development of Ss with eating disorders  
 Stages of ego development  
 Ego development of twins reared apart  
 Life satisfaction by the elderly  
 Mother–child relationships  
 Motives of psychology majors  
 Styles of interaction of nurse therapists  
 Patterns of maturity by age  
 Correlational  
 Social identity to personal history  
 Trust to factors in adolescence  
 Body image to dieting  
 Ego development to a number of factors  
 Fear of war to mental health  
 Goal setting to personality traits  
 Abnormal eating to family relations  
 Conflict level to stages of adolescence  
 Time perception in women to social roles  
 Cognitive complexity to functioning as principals  
 Reference patterns to depression  
 Internal self representations to reality  
 Irrational beliefs to impression management  
 Outcome studies  
 Effects of facilitated communication in autism  
 Effects of incest on ego development  
 Effects of special education on gifted maladjusted  
 Effects of language programs on Romany children

Effects of posttraumatic stress on  
meaninglessness  
Effect on values of large group awareness  
training  
Effects of psychodrama on depressed subjects  
Effects of using trained fifth graders as  
mentors  
Effects of instruction upon measures of ego  
less, it must be noted that the sentence completion  
is being accepted more and more as  
a reliable measure of outcomes in spite of its  
lack of standardization.  
A number of forms of the sentence completion  
technique are in use today. In many  
studies, the form is not reported and is obviously  
one invented for the occasion. A  
particularly impressive standardized sentence  
completion instrument has appeared  
during this time, Loevinger's (1998) University  
of Washington Sentence Completion,  
which is involved in about a dozen of the  
studies reported in Table 6.2. Its scientific  
properties—reliability, validity, interpretation—  
have been studied far more than usual.  
It is centered on the concept of ego development.  
Another, the Hart Sentence  
Completion Test (Hart, 1986), also has  
more than the usual degree of psychometric  
sophistication. Hart describes the measure  
as an "effective social-emotional screening  
instrument" but notes that it is robust  
enough to predict behavior and to suggest  
appropriate educational intervention strategies.  
Other sentence completions are available  
today, of course, some for specific purposes.  
Those interested in efforts that have been  
made to standardize the sentence completion  
responses may be interested in reviewing  
an earlier version of this chapter (Haak,  
1990). It is as much the case now as a  
decade ago that the sentence completion  
technique could profit from a much higher  
degree of study than it has received in both  
standardization and clinical interpretation.  
Nevertheless, the procedure remains so face  
valid, flexible, and useful, its results so disclosing,  
that both clinicians and researchers  
continue to use it extensively.  
Is the sentence completion primarily just  
a form of structured interview? Some suggest  
that this is the case (Dean, 1984; Hart,  
Kehle, & Davies, 1964; Knoff, 1963). Nevertheless,  
the sentence completion covers a  
broader range of subjects faster than the interview,  
and with more uniformity, especially  
when time and comprehensiveness are  
important.

Sentence completion remains a highly useful clinical instrument 70 years after it was invented. It may not always be a good thermometer, but it is nearly always a useful x-ray.

## SELECTING A SENTENCE COMPLETION INSTRUMENT

### Components of a Satisfactory Sentence Completion for Assessing Emotional Status

Although the sentence completion format has seldom been subjected to the desirable reliability and validity studies, there are ways to judge a more dependable and useful measure to be used in establishing the emotional status of a subject. A satisfactory sentence completion for this purpose will usually contain the following features:

1. The sentence completion form needs to elicit responses that can contribute to a general judgment about mental health versus mental disturbance. It also needs to elicit responses that tap the major categories of mental disturbance, such as anxiety and depression. There could hardly be a case of potential emotional disturbance where these judgments would not be necessary.
2. The sentence completion form needs to yield information of a unique or personal nature. It should provide evidence of specific concerns and allow for the discovery of the mental organization and motivational organization of the individual client. This imposition of structure upon structure is part of the richness of the sentence completion's possibilities.
3. The sentence completion form needs to be biased toward the population and the questions regarding that population, which the examiner needs to assess. The instrument is not a "fishing trip" for pathology. It needs to be geared to the population and the context in question.
4. The sentence completion needs to provide a sufficient number of the types of stems in question so that some degree of confidence can be put in their interpretation. For example, to have only one stem geared to achievement in a school population would be pointless, for the response would be undependable. Six stems about achievement scattered throughout the protocol, on the other hand, might provide much more reliable information. As in all psychological testing, one response cannot be viewed as reliable, although a single response may at times be highly pathognomonic

and meaningful.

#### 6. Sentence Completion as a Tool 161

5. The sentence completion should be appropriate for the age served. Adolescents are “put off” by childish language. Younger children cannot “hang on” to a long stem (neither can children with reduced mental ability or severely depressed children).

#### The Haak Sentence Completion Forms

Haak developed a sentence completion instrument (in its first version) in conjunction with the then Learning Disabilities Center of the University of Texas at Austin in 1973. The final version of this instrument is the one that is used for discussion and illustration in this chapter. Two levels of the current forms exist: the “elementary,” for children approximately below the eighth grade, and the “secondary,” for students above that level. These titles are rough guidelines, however, as mental ability and maturity can affect which form the examiner chooses to use. As far as downward extension goes, most normal 5-year-olds can respond to the elementary form and many younger children can do so.

The original version of the Haak Sentence Completion was based partly on the earlier work of the late Robert Peck, who provided some of the original stems and the critical insight to organize the stems into content areas to improve reliability of interpretation.

The first version of the Haaks was also influenced by the late Dr. Fern Williams, first director of the Learning Disabilities Center. It was she who requested the writer’s help to develop a relatively short procedure for screening the emotional problems of children brought to the Center for learning problems. It was she who also contributed some of the categories of the protocol, particularly the “Openness to Help” category; and it was she who first named the instrument the “Haak.” Like other sentence completion forms, this instrument has never been adequately studied, though three doctoral dissertations have addressed it (Baggerly, 1999; Baker, 1988; Wells, 2000). The Haaks are not claimed to be model forms of the sentence completion technique, although they do meet the requirements discussed earlier for an adequate sentence completion instrument. They are widely used in the Texas area and elsewhere (Haak, 1996, 2000). (Psychologists interested in using the Haaks may obtain permission by writing

the author. See Appendix 6.1.) Even the most useful sentence completion can be invalidated by its delivery. If the instrument is read to children, it should be read in the most neutral manner and tone possible. One should read the stems as quickly as the child can respond to them without, of course, creating a frantic pace. There is one exception to this habit: When a child shows by his or her behavior that he or she is thinking about a response before giving it, one should exercise restraint and simply wait the child out. Such responses are usually telling. Sometimes a child waits so long to give a response, and works it over in his or her head so long, that he or she loses the stem. The stem may be repeated, but if the child chooses to avoid it, the examiner should move on (discreetly marking such a response for his or her own benefit in interpretation later).

It is best not stop and discuss responses. Some query of responses may be carried on when the protocol is finished, but even this is likely to create suspicion and distrust if there are remaining instruments to complete. The child who is being examined will take his or her cues from the examiner; therefore, a calm, pleasant, but neutral administration is best. The more such an attitude can be maintained, the more confiding the child is likely to be. If the examiner behaves nervously about any items on the sentence completion, the child will notice. This is also true if the examiner reacts to the child's responses. An effective examiner should be as nonreactive to "I like . . . axe murderers" as to "I like . . . chocolate cake."

#### **USING THE SENTENCE COMPLETION AS PART OF A TEST BATTERY**

The sentence completion should normally be used as part of a comprehensive battery of tests if the emotional status of a subject is being evaluated. Hypotheses generated by the sentence completion can be explored for support or nonsupport in the other, more objective data (and vice versa). The assumption is that a rounded battery of tests will generate data that can be integrated into a meaningful, organized picture of the client's

#### **162 II. PROJECTIVE METHODS**

functioning. This picture, of course, must invariably rest on some theoretical base. The sentence completion is particularly suited for use with a theoretical orientation to

the assessment, which depends on general systems theory.

In systems theory, the client—the schoolage child—constitutes a small system composed of important subsystems: physical, intellectual, emotional, temperamental, and attitudinal, at the least. These subsystems interact in ways that support or interfere with each other to produce the behavior of the child in question. A child, or anyone else, whose major subsystems function together to support normal daily functioning, growth, and development can be said to be healthy.

When a child is healthy, however, that judgment infers more than just that the child him- or herself is doing well, and this is because human beings are open systems, affected by their surroundings. The child operates within a larger context of systems—home, school, and community and, with increasing age, national and global systems. It is difficult for even a healthy child to function normally if he or she must live within a dysfunctional home, school, or neighborhood system. The effects of such negative factors are felt by the child.

It may be the case, of course, that at times the supportive systems around the child are functionally adequate for most normal children, but the child in question has serious problems due to the imbalance within his or her own system caused by disabilities (inadequate subsystems). Disabilities distort the smooth growth and development of the child, and that fact makes demands on the environment for compensatory interventions and treatments to improve or restore the inadequate subsystem(s). In the best scenario, these demands are met and the child is restored to normal functioning. However, the more common case occurs when disabilities are found to exist in both the child and the environment.

When disabilities exist in both the child and his or her environment, in its mildest form this is certainly understandable—it is wearing on parents and teachers to make the daily compensations that must be made for a handicapped child. Such compensations are often expensive in terms of both time and money. In its most virulent form, the disabilities in the environment may actually cause or definitely exacerbate the disabilities in the child. This is the complicated situation that the assessor most often faces

when conducting an emotional evaluation. One must understand how the dysfunctional aspects of a child's system itself interact with possible causes or contributors in systems at large and then make appropriate recommendations for remediation, alleviation, or compensation of these dysfunctions wherever they exist. Such is the role of the comprehensive assessment; and it is rather a daunting task.

When an examiner is first given a referral, he or she must consider how to approach the examination task. Some highly obvious disabilities such as physical, visual, or auditory problems have their own specialized pathways to be followed in assessment. However, a child referred for emotional problems is not so uniformly approached because one is sure of neither the problems nor their sources. Personally, I prefer to approach such an examination by gathering and considering the data in a most objective to least objective track—standardized intellectual and achievement data first, health history, developmental history, history of previous treatments (if any), referral information, parent and teacher information and checklists next, and, finally, the personal examination of the child—checklists, sentence completion, drawings, and projective stories. (Here we speak of nonstandardized clinical data originating with the subject as “subjective” and other data as “objective,” recognizing that much of the “objective” data [i.e., parent and teacher information] are hardly objective at all in the true sense of the word.)

When beginning with the more objective data, a framework—intellectual abilities, physical information, history of treatment, and opinions of teachers and parents—is created to which the dimensions of the more subjective data can be related. Sometimes the objective explains the subjective, and sometimes vice versa. A comprehensive examination for emotional disturbance should result in a tightly woven assessment in which issues of major concern are supported by data from multiple sources in the assessment.

#### 6. Sentence Completion as a Tool 163

School is the work of the child; the intellectual abilities of the child are the tools he or she has to do the work. That is why it is critical to have a full battery of results from the Wechsler Intelligence Scale for Children—III (Wechsler, 1991), the Stanford–

Binet Intelligence Scale IV (Thorndike, Hagen, & Sattler, 1986), or the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983). Also, as important as the verbal IQ and performance IQ are, an analysis of the individual subscores of the test is equally informative. Anyone not comfortable with analyzing these subscores can find a useful assist from several measures of organization of these subscores—for example, the Individual Ability Profile (Dean, 1983)—which casts these subscores into more neuropsychologically meaningful and useful dimensions.

History is, of course, critical. The developmental accomplishments (talking, walking, etc.) are little standardized tests within themselves. Is this a child who met these tests successfully? The history tells us about past illnesses, accidents, and accomplishments. The history tells us about the child's environment. Nowadays, for example, one wishes to understand how many adjustments the child has been called on to make (e.g., in relating to father or mother figures in the home). What are the stresses with which the child has had to cope, and how are they reflected in the child's more subjective productions?

In gathering parent and teacher information, it is common today to use, in addition to history questionnaires, some form of behavioral checklists. The checklists with perhaps the most acceptable reliability and validity are the Achenbach Child Behavior Checklists (Achenbach, 1991), although others have now been developed that meet reliability and validity standards. Nevertheless, it should be remembered that however much behavior scales may have been standardized, they are not necessarily reliable or valid in the *individual case*. The instrument may be reliable in the general population, but the person filling it out may not be. This is not altogether an unfortunate thing. One learns methods of evaluating the evaluators—a process that provides highly useful information about the major individuals with whom the child interacts. For example, when teacher reports and parent reports agree, it is seldom the case that either one of them is inaccurate in reporting behavior. When they do not agree, there are several possibilities. One possibility is that the child behaves differently in the two settings represented by the two reporters. Some assessors

do not give full credence to this possibility. It is likely that the school is the problem when this happens. Many children escape from school to a happier, more relaxed home life; few children escape from a disturbed home into school and function well there. Some children *do* function differently in different settings, and when they do, this ability argues against a diagnosis of generalized emotional disturbance.

There are several other possibilities when teacher and parent behavioral reports disagree. Either of these reporters may be reflecting a present high level of concern about the child in the reporter's setting. In this case, the concerned reporter will elevate most scores. The profile of the child based on such data will be elevated across the board, violating the factor structure that provided the profile. This is not liable to be a realistic report of the child, but it is an indication of the reporter's level of concern, either with the child or from other sources of stress.

Reporters may also have definite feelings about the child's eventual diagnosis, which they know is forthcoming. If they personally believe the child is disturbed, one way to ensure this diagnosis is to elevate scores. The reverse, too, is also true: the underreporter, whose scores create a flat profile, may be expressing his or her idea about whether a diagnosis of disturbance is appropriate. Or, he or she may be registering a hostile attitude toward the examination itself ("There is nothing wrong with my family. . . . Leave us alone . . . etc.") When either the teacher or the parent reports in a manner that generates a differentiated profile of results, and the other reporter produces either a flat profile or a generally elevated profile of scores, it is usually the case that the reporter who produced the differentiated profile is more accurate.

A behavior observation in the classroom can also be a valuable piece of information. Some teachers are truly expert in performing these observations, providing a concrete sample of a child's behavior that answers many of the questions raised by the checklist reports. The most valid procedure for obtaining behavioral observations is to use one teacher who regularly makes these observations. If this is not possible, teachers can perhaps do observations in each other's

## 164 II. PROJECTIVE METHODS

classes. It is nearly impossible for the head teacher in a classroom both to teach and to do a truly detailed, concrete behavioral observation at the same time. The examiner, of course, may be required to make these observations. The most common objection to this procedure is that the teacher may protest the resulting observation, saying that the child did not behave as he or she usually does. This may be true, as the examiner is a complete stranger to the classroom and may have a generally dampening effect on all behavior in the room while he or she is there. Unless the examiner is specifically invited to see the behavior of a certain student (this is not unusual with a teacher who is self-confident and concerned about a specific student), using a single teacher or corps of teachers trained to make concrete observations appears to produce the most reliable data.

Once all the objective data (discussed earlier) have been gathered and examined, one is prepared to go ahead with the more “emotional” part of the exam. One has a set: One knows the capacities of the child, what has happened historically to the child, the environmental impact on the child, and something of the views of that child as seen by the parent and the teacher. It is time to go behind the curtain and see what effect all these variables have had on the child, for, after all, *emotions are reactions*. It is time to know what the emotions are reacting to. The child will tell us that.

#### **USING THE SENTENCE COMPLETION TO ASSESS EMOTIONAL DISTURBANCE**

The focus of the remainder of this chapter is on the clinical interpretation of the sentence completion technique itself in evaluating school-age children for the condition of emotional disturbance. The intention here is to offer assistance to the user in analyzing the rich data usually obtained with this instrument toward that end. The sentence completion is a particularly useful subjective instrument because it casts a wide net in a short period for issues of immediate concern to the assessor: relationships, work attitudes, view of self, ambitions, sources of emotional support or pain, the child’s own view of the educational difficulties, and others. The sentence completion is a semiprojective instrument: It allows the projection of the child to come into play, but it focuses on reality issues important to the assessment.

Its face validity makes it more believable than more subjective approaches to professional persons who are not clinical psychologists.

A complete emotional evaluation usually includes, however, the more subjective and projective instruments, particularly drawings and stories made up in response to stimuli such as pictures. As the examiner moves from more objective to less objective data, one loses the comforts of standardization and normalization but gains in richness of data. This is, of course, where clinical expertise becomes necessary.

### **Establishing “Emotional Disturbance”**

To establish the simple-appearing matter of whether or not the child is “emotionally disturbed” is the goal that must be kept in mind in working through all the child’s productions.

This task is analogous to requiring a physician to certify someone as sick or well. Every professional person who deals in clinical settings knows that such arbitrary dichotomies are usually false. Some persons may be clearly sick and others clearly well, but most will fall into some degree of what is often called “the worried well,” neither free of concerns nor totally dysfunctional—and this applies also to children (who were not officially capable of being classified as depressed until 1970!).

In addition to the fact that most functioning persons have some degree of identifiable problem dynamics, there are further difficulties that appear with children. Children are always going through predictable changes and stresses, if only those associated with stages of development. Each of these stages presents its known challenges, which are probably the result of temporary imbalances in physical and psychological growth and maturity. In Western culture, these seemingly

6. Sentence Completion as a Tool 165  
inherent problems associated with aging have been variously conceptualized as “childhood stages” (Ilg & Ames, 1960; Spock, 1946), “psychosocial stages” (Freud, 1949), developmental tasks (Havighurst, 1949), and “psychosocial crises” (Erickson, 1959), all resting on some fundamental stages of neurological growth and development (Reinis & Goldman, 1980), cognitive accomplishments (Piaget & Inhelder, 1958), and the simpler matter of “on time” developmental milestones (age of sitting, standing, walking, talking, etc.). No psychological examiner should attempt to evaluate children

without a sound knowledge of all these developmental relationships. Another complication of assessing children is that they are far less stable and predictable than adults. Assessors may have fairly constant notions of what capabilities to expect in a 30-year-old adult with an eighth-grade education with a Verbal IQ of 80 and a Performance IQ of 105. What to expect in a child with similar scores is far less clear for a number of reasons: (1) all test scores are less stable in children; (2) children's performances are more affected by environmental conditions than those of adults; (3) children have not formed stable work habits; (4) children's performances on verbal measures are affected by their state of verbal development; and (5) children's performances are affected by the degree of pressure or structure they feel from significant adults, especially on items they feel to be revealing or intrusive.

Finally, assessing "emotional disturbance" as a global condition of a child involves to a great extent, as we discussed previously, a judgment about the general adequacy of the environment in which the child finds him- or herself. Every practicing school psychologist has had to label a child "emotionally disturbed" when this disturbance was clearly the result and only the result of observable environmental deficiencies. There is rarely an analogous problem in adult psychology or psychiatry. Not all emotionally disturbed children fit this description, but a troubling number do.

### **Systems of Emotional Classification**

There are two major systems by which children today are classified as "emotionally disturbed": (1) the criteria of Public Law 94-142 (since partially revised as IDEA, but these criteria have not changed) and (2) the criteria set out in the fourth edition of *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV; American Psychiatric Association, 1994). The federal criteria (Individuals with Disabilities Education Act [IDEA]) must be met for the child to be classified as "seriously emotionally disturbed" and receive appropriate services in the public school. The federal criteria are based on the work of Eli Bower (1969) and are as follows:

- (i) The term ("emotional disturbance") means a condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree, which adversely

affects educational performance:

- (A) An inability to learn which cannot be explained by intellectual, sensory or health factors
- (B) An inability to build or maintain satisfactory interpersonal relationships with peers and teachers
- (C) Inappropriate types of behavior or feelings under normal circumstances
- (D) A general pervasive mood of unhappiness or depression; or
- (E) A tendency to develop physical symptoms or fears associated with personal or school problems.

(ii) The term includes children who are schizophrenic.

The term does not include children who are socially maladjusted, unless it is determined that they are also seriously emotionally disturbed. (Education for All Handicapped Children Act, 1975, §300.5)  
The matter of what constitutes "social maladjustment" has never been satisfactorily determined, although it has been the center of much discussion. Also, recent legal and governmental opinion leads us to believe that abnormal feelings (the affection component) must be part of the reason for the diagnosis of "emotional disturbance," not merely abnormal behavior per se. Apparently, the condition of emotional disturbance was never intended to cover problem behavior without the affective component.

In short, for a child to receive many services today, especially services provided by special education in the public schools, some professional must finally make a judgment that a child is emotionally disturbed

#### 166 II. PROJECTIVE METHODS

(the "sick" or "well" question). The reliability problem inherent in using such a label across the whole of the United States by variously trained personnel using various systems of judgment simply boggles the mind. Unlike other categories of disability, especially the physical ones which have clear-cut and measurable criteria, the disability of emotional disturbance remains a matter for professional judgment alone. The only available "reliability" for this situation at present is to ensure that such judgments are made by fully trained and experienced professional personnel without conflicts of interest, using a range of instruments which provide as much information as possible about the condition at issue.

The professional who assesses a child must also determine at what point on the continuum of emotional disturbance he or

she is to declare that a condition has reached the stage defined by federal law as “serious emotional disturbance.” After the matters of having the condition for a “long time” and having the condition to a considerable “degree” have been met, that judgment (a sufficient *degree* of emotional disturbance in the school setting) is supposed to be made on the basis of “educational need.”

The effects of the proposed emotional disturbance on the child’s ability to learn, either academic material or the normal level of interactional (behavioral) skills, must be clear and serious. Often that point on the continuum is hard to determine and subject to argument. Many schools want to consider only “academic” need, yet the rule clearly states that retardation in “developmental” skill constitutes an educational need. And what of the bright child whose performance is significantly below expectation for his or her ability level but still average or better when compared to the whole population? “Educational need” is not always easy to determine. But it must be the final hurdle an assessor clears before judging a child to be emotionally disturbed.

#### INTERPRETING THE SENTENCE COMPLETION

If it is the case that the sentence completion can make a major contribution to the decision of whether or not a child is emotionally disturbed, it is also the case that a critical element in acquiring the interpretive knowledge necessary to be comfortable with the sentence completion is the matter of knowing what to expect in the child’s response.

“Norming” the sentence completion responses is inextricably meshed with the reliability and validity questions and knowledge of developmental stages noted earlier.

The highly experienced examiner carries such norms in his or her head, but that is not helpful to other examiners, especially those who are just beginning their careers.

In one attempt to generate some such “norms,” Peck and Haak (1973) studied the responses of 600 normal children ages 10 and 14 to the sentence completion and found some surprising results. For example, 75% of the children would actively problem-solve if they saw others doing it successfully in school, but only 50% would ask for appropriate help. Also, only half would attempt to resolve a problem with a friend

or comply with an (adult) request, and only 33% would try to do something active about a class bully or try to deal with parental anger. In short, coping skills in normal preadolescents were not remarkably positive. It would seem from such results that one should be hesitant to overvalue such deficits in an individual child being evaluated for pathology.

All types of norms for children need to be understood by the assessor, including subcultural norms. But particularly important are developmental norms and language developmental norms. For example, it is fairly typical of a 5-year-old to state that his or her favorite school activity is "playing"; this response given by an 8-year-old is a red flag for developmental delay. Unfortunately, the ability to make many of such interpretations comes from experience.

The sentence completion contains both behavioral and affective information seen from the viewpoint of the child. The behavioral reports of the child can be checked against the "reality" of the behavioral reports of the assessor and parents, teachers, and students. The affective report can be checked against the "reality" of the parents', teachers', and students' perception of how they think the child feels (an area in which great discrepancies often exist). Some

6. Sentence Completion as a Tool 167

of the comparative judgments that can usually be made with assistance from the sentence completion data are as follows.

#### The "Rule-Outs" Intellectual Difficulties

A clinician experienced with the sentence completion can often predict a child's IQ score within a few points from the sentence completion data. This can be done primarily on the basis of fullness or poverty of language usage, articulateness, concreteness of language, complexity of language, perseveration in language, imagination, general level of expressiveness, empathy, openness versus defensiveness, expressed social intelligence, and qualifications of responses. In particular, qualification of responses almost never occurs without good intellectual potential ("I am generally nice but sometimes I am not").

Certainly no emotional evaluation should be conducted without hard, recent data regarding intellectual potential and achievement status. But the sentence completion

can act as a functional check on other intellectual data, especially in the sense of indicating more potential than the child is presently able to demonstrate on testing.

(Beware, however, of overstating this case; a great deal of damage is done by careless clinicians who lead parents to believe their child is “much brighter than the test scores indicate.” Such statements sometimes lead to years of parental pressure on a child who is not having his or her needs met.)

It is also not enough to have simple intellectual “test scores” without considering which test was used, the subscores involved, and the pattern of subscores reported by the test. (We have discussed this issue previously.)

A child with a great deal of within-child variance in important mental abilities, as represented by the subscores on the aforementioned tests, may be suffering from learning disabilities, other neuropsychological dysfunctions, or a mismatch between the abilities of the child and the curriculum he or she is required to master. Such a child can experience much stress and discomfort with school.

It is also common to find children who have masked intellectual difficulties for years behind a front of exceptional social competence. This particular adaptation often breaks down at the high school level, when the student’s social competence can no longer compensate for the increasing degree of difficulty of educational expectations.

The charming little girl who has watered the pot plants all the way from first grade to high school is suddenly engaging in acting-out behaviors. Be suspicious that she has real difficulties in meeting new and higher expectations.

Will these academic stresses be reflected in the child’s sentence completions? Almost always. The child complains that school and certain subjects are hard; the teacher does not try to help him or her; there is no use to study; the tests are tricky; peers are rude; the smart kids get to do everything. The self-concept often reflects feelings of stupidity and hopelessness. Increased alienation from peers and school is reflected in many stems. Depression may be expressed in some stems.

A child experiencing intellectual difficulties may also be experiencing emotional disturbance, of course; however, if the emotional symptoms being displayed can be

directly traced to the intellectual deficits and the stress that the deficits place on the child, then it is improper to label the child “emotionally disturbed.” In such a case, the emotional symptoms are secondary effects of the primary diagnosis, and the child requires other academic remedies than emotional remedies, though emotional support for a time may be helpful indeed.

#### Attention-Deficit Disorder (Attention-Deficit/Hyperactivity Disorder)

Symptoms of attention-deficit disorder (attention-deficit/hyperactivity disorder; ADD/ADHD) can be easily elicited by the sentence completion. In fact, the child with one of these disorders will commonly and openly complain about his or her own attentional problems (“I have a hard time paying attention”; “I often think about other things”). The child will also often complain that others shout at him or her, yell too much, lose their tempers at him or her, are too impatient, become exasperated, and are otherwise frustrated with the child (they are!). The child often seeks isolation, freedom, a high degree of personal movement (the life of a bird, motorcycles, high-speed cars) (i.e., being alone, in control, and away from the “bugging” of others).

The child with ADD/ADHD typically demonstrates a “thin-skinned” attitude and feels that other people are always picking on him or her. The child often feels acutely unfairly treated. No one understands the child. He or she nonetheless displays a paucity of empathy and understanding of others and sees nothing from any viewpoint other than his or her own. The empathetic and reflective processes are usually markedly deficient. This does not mean that the child does not give some positive social responses; indeed, this is often the case. But positive social attitudes are positive in the conceptual sense only and often do not translate into concrete social skills. (Again, beware: We speak of the typical case. Because of other compensating factors, an individual child with ADD/ADHD may be quite socially competent.)

The child with ADD/ADHD often projects few plans, including plans of action. This of course may relate to deficient frontal lobe development. This lack of planning is often a critical factor. Responses to the stems that refer to future plans are often

vague, undifferentiated, and obviously not previously considered. And, there is another matter connected with this paucity of projection: The child does not appear to feel “responsible” for impulsive actions that he or she did not mentally consider in the first place. (Indeed, the history of common law agrees with this concept of premeditation and responsibility.) Such children complain bitterly about other people’s reactions to their impulsive misbehavior. And the other people rail (constantly!) about the child’s lack of “responsibility.” Between the two points of view rests a “moral” battleground. These issues are often reflected in the sentence completions and in the referral originally received. (“My father thinks I am irresponsible.” “My teacher thinks I am stupid.” “I wish teachers would leave me alone.” “My friends always expect me to get in trouble first.” “Adults blame me for everything.” “I am always blamed.”) As previously stated, social intelligence is often not lacking in the sentence completion data of an ADD/ADHD child. Such a child will express positive social attitudes that adults will scarcely believe (and often do not). Rather than being immune to the effects of others, such children seem hypersensitive to these effects. They are often in acute states of affective starvation, with perhaps the mother the only compensating factor in the picture (“My mother loves me anyway”). Their behavior, of course, is what puts them in this position. Attentional deficit leads to emotional double binds for children, and many of their difficulties are expressed in the sentence completion data.

How does one differentiate whether the symptoms displayed in the sentence completions of a child with possible ADD/ADHD are attentionally or emotionally based, particularly when the child has not been previously diagnosed as having an attention deficit? This is a difficult and ongoing question even for research. The current answer appears to be that when almost all the troublesome sentence completion responses coalesce into what is known of the common secondary symptoms associated with the deficit (examples have been given previously), one can clearly feel comfortable in assuming that ADD/ADHD rather than something else is the basic cause of the child’s disturbing behaviors. At that point, when

one has clustered the emotionally loaded sentence completion responses of the child with true ADD/ADHD, few such responses will be left out of the picture painted previously. Of course, a child can be both attentionally disordered and emotionally disturbed. In that case, the examiner will see the usual coalescence of secondary symptoms associated with ADD/ADHD, but there will also be a large number of negative, emotionally loaded responses left over that do not neatly fit this picture and do not fit another symptom picture, such as depression. (One caution needs to be observed in this regard, however, as a mild level of depression is common in ADD/ADHD children.) Nevertheless, strong levels of emotional conditions, including depression, should not be viewed as deriving solely from attentional deficits. It is poor practice to classify such a child as ADD/ADHD and ignore the second or third symptom picture because one has already gotten the child "qualified." Such a child will perhaps receive services for one condition and fail to thrive because other disabilities have been ignored.

### **Stress**

One of the major rule-out decisions facing the clinician who must decide the emotional status of a child is a decision regarding the contribution of stress to the child's symptom picture. Sometimes this decision is difficult. The clinician needs to consider the number and intensities of stressors upon the child, their recency, and their duration. Important is the child's reaction(s) to the stressor(s). A child for whom multiple and strong stresses in the immediate past are found cannot safely be fudged to be emotionally disturbed unless the emotional disturbance preceded the period of the stressors or is in itself extreme. The "rodeo rule" is particularly apt here: One must evaluate the rider by first evaluating the horse. Evaluating the "horse" (in this case the stressors, of course) is not always simple. Persons causing the stresses will often be tempted to underplay their role in upsetting the child. Parents looking for an excuse for their child's condition that does not point back to any kind of family variable will overevaluate minor "outside" stressors that any healthy child should be able to handle. Teachers may exaggerate home conditions for the same reason. Yet, difficult as it may

be, it is usually productive to try to get an accurate picture of the “horse” before proceeding to evaluate the strength of the rider.

(A discussion of how to judge the accuracy of parental and teacher data appeared earlier.)

And, of course, life being what it is, the causes of the child’s stresses may be beyond the ability of any human being to change (death of a parent, terminal illness in the family, etc.). Obviously, this latter event is blatant to anyone and does not require a complicated assessment to determine that a severe stress exists.

Some typical responses of children under heavy stress include the following: the child will report physical problems, often more than anyone else has reported for him or her (adults are surprised to learn the child says he experiences daily stomachaches in the morning, for example). The physical conditions of headache and stomachache may be reported as occurring together. In this case—and in any other appearing potentially serious—the clinician is justified in asking the parents to obtain a physical examination of the child. Although stress and anxiety can cause co-occurrence of symptoms, so can subclinical seizures, migraine headaches, and other serious conditions. Although it is the case that most often the child will be found to have no physical problems, this judgment can only be made by a person trained to do so and should not be assumed.

The content of the stressed child’s sentence completions can reflect a low level of energy (“I am tired most of the time”; “I cannot get my lessons”). There is seldom the focused hostility of the acting-out child, for the student is often blaming him- or herself for any problems at this stage. There may be consequent confusion and sparse, somewhat noninformative and defensive responses.

But, in any case, the most important feature of the sentence completion responses in the case of an overstressed child *who is not emotionally disturbed* is that usually his or her responses are “better” than one would have expected from knowledge of the stressors (and often from knowledge of overt behavior). This is important.

There will be some preserved positive social attitudes, even though spotty; preserved though damaged belief in oneself (“The one thing people need to understand about me is that I am really trying . . .”; “I am really a good person”); articulated though not necessarily

followed-out coping strategies; some belief that things will get better; and some hope for the future with scraps, at least, of future plans. The examiner's reaction to such a child is often something like, "This is a pretty good kid!" The examiner finds him- or herself "rooting" for the child. These reactions in the examiner are an important source of data. That one wishes to rescue such a child is telling—for this is a child who usually needs to be "rescued." Sad to say, when multiple and/or strong stressors persist over time, they will almost inevitably erode the functional capacities of the stressed child. Finally, that sad state is reached in which, even though all the causes of the child disturbance can be traced directly to the environment, the child is actually emotionally disturbed. His or her functional capacity is scarred as a consequence of the damaging conditions. To call the child emotionally disturbed—a term many still see as pejorative—seems unfair to the child. To fail to call the child emotionally disturbed, which he or she has finally become, may be unrealistic and, more important, may deny the child critical services. And, of course, it is dishonest. Most often this diagnostic call in strong cases of stress will be resolved by the presence of depression, which clearly qualifies the child for services.

Fortunately, when these cases are discovered earlier and at milder levels, interventions can be made to circumvent the further effects of the unreasonable stressors. A proper report that outlines the problems and suggests practical interventions is due the child, his or her teachers, and his or her family. Just because the child does not "qualify" for services in special education does not mean that proper services should not be sought in the regular education programs of the school and in the community.

Specifying a time to reevaluate the condition of the child (if only informally) can also be helpful, as it serves as a surveillance function or goad for others to seek proper services. The assessor becomes the child's informal advocate in such cases.

Of course, we have not considered the case in which the child's stresses are emanating from activities that are against the law. Clearly the examiner is duty-bound to report such events to the proper authorities. In that case, it is often best to inform the

## 170 II. PROJECTIVE METHODS

school's major administrator of what has been found and the fact it will be reported to the proper authorities. School at times may be the child's best protector—sadly, at times his or her only protector.

### The "Rule-Ins"

#### Depression

A depressed child almost never completely "escapes" the sentence completion. Evidence for depression may be found in the following instances, at the least:

1. *Outright references to depression, flatly stating that the child is depressed and sad.* (It is rather remarkable how a child can often say this and have it totally ignored.) Other language the child may use to indicate depression includes tiredness, hopelessness, noninterest in the daily affairs of school or play, a desire for isolation, and a need to escape the school situation.
2. *Outright reference to self-destruction.* These references can also be vague, such as responding to questions about future plans with an indication the child will no longer be around.
3. *Responses that confuse anger and sadness* ("When I am angry, I cry"; and "When I am sad, I am mad").
4. *A response to either of the following stems that indicates sadness:* "I am nearly always . . ." "I am never . . ." The denial form of this particular stem often allows the child to express a depression he or she otherwise defends against.
5. *Quite a number of expressions of anger combined with a number of expressions of positive social and moral attitudes* (depressed persons are more apt to be socialized than not).
6. *A self-concept almost altogether negative except, at times, for the person's own view of him- or herself* (i.e., he or she may report that others see him or her negatively, but he or she holds on to a positive though perhaps defensive self view).
7. *Homicidal ideation.* This is not rare and in recent years, with the advance of violent video games, movies, and TV, appears to be more easily expressed and less socially inhibited than previously. (This means, also, that its expression is harder to judge.) Such ideation can be accompanied by concrete and vivid images. Anger may be more obvious in

some younger children, who lack the controls to subvert their depression into adult type, self-depreciative forms of this emotion.

8. *The expression of a conscious level of nonaffection and nonsupport* from parental or adult figures, often with a yearning for this support.

9. *Premature sexual and romantic ideas.* A young child thinking about getting married, for example, has usually given up on obtaining any immediate source of affection.

10. *A general paucity of expressiveness or high degrees of defensiveness in com-*

6. Sentence Completion as a Tool 171  
*pleting the sentence completion stems;* protests over having to complete the instrument or obvious upset in doing so; demands to stop the examination; obvious and building agitation; anger directed toward the examiner (“My biggest problem is you”). These behaviors in completing the instrument are an indication of the low tolerance of depression.

11. *Intolerance of noise and stimulation* (“Everyone yells”).

12. *A strong desire for activities combining escape, aloneness, and self-sufficiency,* sometimes but not always involving the element of speed (“Nothing feels as good as being alone.” “I would make a good pilot.” “If I were a king, I would go skiing on a high magic mountain—the tallest in the world.” “I often daydream about life on a little island.”).

Of course, the quantity and intensity of such expressions as the aforementioned will determine to what degree one suspects depression in a child. But the sentence completion is a productive place to look for depression, for its partial structuring of the responses in the sentences stems is an aid to expression in the low-energy, depressed youngster.

### Anxiety

The presence of strong anxiety in a child may exhibit itself in the sentence completion as follows:

1. *A large number of statements expressing fear and apprehension.* One needs to be careful in evaluating fear, however, for some fears are known to be associated with certain developmental states. Also, certain fears are related to real-life concerns, and in

such cases, these statements are more indicative of stress than of an emotional disturbance of anxiety. Children also have fears blown out of proportion by television, even well-meaning TV ads. For example, it is common now to find children who fear that their parents are eminently going to die because they smoke cigarettes. The number and intensity of fears can be helpful in evaluating whether fears are indicative of a generalized anxiety state.

2. *Unexpected expressions of anxiety in response to otherwise innocuous stems.*

When a child responds anxiously to a stem designed to elicit anxiety, that is one thing; when the same child gives an anxious response to a stem probing for other issues, that is another. (A couple of examples: "If I had all the books in the world, I would stack them up around me and hide inside" or . . . use them to throw at burglars.")

3. *A number of statements that express vulnerability.* Anxious students do not feel safe. One indication of vulnerability does not constitute anxiety; however, several normally do.

4. *Premature responding to an item* (an instance of overreaction). Anxious students often cannot wait for the examiner to finish the sentence stem (if it is being read aloud) or do not read the whole stem carefully themselves. It is valuable for the examiner to notice the content of a stem that draws this type of overreaction. What develops with such a sentence completion may be logical in thought but simply incorrect in grammatical structure ("I often think that she is leaving").

5. *Overproduction of responses.* The overproduction of the anxious (including the compulsive) student is mere verbal running on and not the same as the long, complex, and, more important, qualified responses of the articulate, thoughtful student. The anxious student may repeat him- or herself over and over; go off on a pointless tangent in a linear, compulsive manner; or simply run on and on, unable to put a stop to his or her verbalization. This may happen particularly in response to a stem elicited to probe for anxiety. Such behavior is indicative, of course, of anxiety.

6. *Lack of defensiveness.* Most children are open in their responses to the sentence completion, and a few terse sentences are not necessarily significant. A few stems simply

may not interest the child or allow for little development. The following example—“When my father gives me a lot of work, he gives me a lot of work”—may be quite meaningful in relation to other stems suggesting hostility or problems with authority, but if no such responses exist elsewhere, the single, terse response should not be overevaluated. Nevertheless, the protocols of anxious children are usually void of

## 172 II. PROJECTIVE METHODS

defensiveness and replete with overresponding because control over their responses (or anything else) is beyond their power.

7. *Worry*. Anxious children indicate in many responses that they are worried, even in response to stems often seen as pleasant. The psychologist working with an anxious child will not lack for reports of agitated behavior in the child. But the sentence completion information will help the examiner to determine the sources of this agitation and the extent of it. The absence of many secondary symptoms of ADD/ADHD, the ruling out of major stresses, the ruling out of physical problems, the presence in the sentence completion of many responses typical of anxiety (as discussed previously)—these leave a picture of a child experiencing an emotional anxiety state. It is also common to find an anxiety state in a child suffering from other emotional problems, especially depression.

### Thought Disturbance

Thought disturbance is certainly an indication of disturbance, and a rather uncommon, severe one in a child at that. Children may truncate their responses, summarize their responses, respond with partial data, suppress their responses, convert their responses, and do all manner of things to try to avoid fully expressing how they actually feel or what they think. These maneuvers are generally unsuccessful if the testing instrument is of sufficient length. But thought disturbance is generally frankly expressed: the thought content is just outright disturbed. And it is usually expressed in a manner that is at odds with the context of the stimulation.

Indications of a thought disturbance produced by the sentence completion include the expression of illogical ideas and the expressions of these and other ideas in grammatically abnormal, fractionated sentence structure. One highly illogical thought does not constitute thought disturbance, but it should not be ignored either. One such red

flag may alert the examiner to look for thought disturbance at a more subtle level in other expressions. Thought disturbance also has an intrusive quality. This can be overt, in the interruption of an otherwise appropriate response; or it can be inferred from a response incongruent with the sentence stems (“My father wanted I was a chicken in the Easter play.” “If I were a king, I would buy a whole truckload of toys. I would invite all the children. And I would murder the queen. We would all get together in the park and have a good time.” “In my family, I am he cane to see us.”). Thought disturbance may be differentiated from the premature responding of the anxious student. Even though the anxious student’s response may not be grammatically smooth, the responses will usually be in the same ballpark as the stem. In thought disturbance, the response may have no obvious relation to the stem (though it may have an illogical personal relationship, which at times the examiner can trace). The use of highly convoluted grammar itself is often a sign of thought disturbance. Grammatical improprieties are rare in children, even uneducated ones. The grammatical mistakes of uneducated children are fairly well known and mean nothing psychologically (“I ain’t finished”; “He done it”). But actual fracturing of the basic grammatical structure should not be excused on the basis of lack of education. It may indicate brain damage, severe language dysfunction, neuropsychological conditions, psychosis, or a combination of any of them. The danger in assessing thought disturbance is the tendency of the examiner not to “make anything of it” with a response that makes no sense. There is the temptation to ignore a sentence or two like this (“Oh well, kids say a lot of nonsense things”). Although one is certainly not going to arrive at a diagnosis of thought disturbance upon the basis of a couple of responses, the opposite danger—to disregard that possibility because one cannot easily explain such sentences—is an equally dangerous behavior. The child in whom the possibility of thought disturbance is raised should have follow-up assessments, including the basic matter of whether the child’s auditory processes are intact. Often such a child can be quite disturbed but sedate and withdrawn, and no one appreciates the depth of the child’s suffering. It is such a

child who may be “caught” by the sentence completion if the child’s responses are taken with appropriate seriousness. Such a child seriously needs help.

In a rare case, a child who concocts a system of nonsense responses may not be

#### 6. Sentence Completion as a Tool 173

thought disturbed but under some intense psychological pressure. The child is frightened to be giving responses to the threatening items in the sentence completion but sees no way out of it. One such protocol, for example, was given by a child involved in a bitter divorce and custody suit. He was being pummeled by both sides for his loyalty.

In responding to the sentence completion, the child substituted animal figures for family members, actions for other actions, and so forth. For example, one of his responses was “I love to dance with a horse.”

The examiner counseled with the boy and obtained a second, more straightforward protocol. The two could be laid side to side to see the inventive system the child had employed.

Obviously, no thought-disturbed child could have done this, a task requiring superior intelligence, imagination, quick invention, and good short-term memory.

Defensiveness must always be considered when a child gives nonsense responses to sentence stems. Even more, the reasons for the defensiveness need to be determined. But the fractionated, illogical responses of thought disturbance derive from, though probably include, more than the need to be defensive.

#### Other “Rule-Ins”

Depression, anxiety, and thought disturbance are major contributors to a final diagnosis of emotional disturbance. If the child is being assessed in school for possible services, the evidence for these conditions will still need to be subsumed under one of the categories of the definition of “emotional disturbance” (see previous discussion for a definition of “emotional disturbance” under federal law). In this system, depression has a category of its own, but anxiety, thought disturbance, and other extreme psychological conditions do not so easily fit under one of the remaining categories of that definition, namely, nonperformance for which no other cause can be found; inability to relate to teachers or peers; abnormal behavior under normal conditions; and psychosomatic behaviors and fears, plus, of course, schizophrenia—

and specifically not “social maladjustment (see previous federal definition for exact wording for these categories). In my experience, a state of high anxiety or extreme anger most often qualifies as abnormal behavior under normal conditions. Thought disturbance, fairly rare, may be the cause of any qualifying condition or may not qualify a student at all depending on the investigation of the causes for this condition. This leaves “the inability to relate to teachers or peers” and “psychosomatic behavior and fears” as two conditions we have not discussed. The reason for this lack of explication is that the evidence for these two categories is usually so overwhelmingly obvious, both in the sentence completion and in other instruments obtained. Data for the relational category usually come in with the referral, loud and clear, and are often the very basis of others’ concerns. It must be remembered that experienced teachers who are concerned about a child’s relationships have a great deal of normative data on which they make this judgment—usually more than psychologists. It is sometimes the case that the complaint is based on a teacher–pupil mismatch, but not often. The sentence completion is particularly helpful with assessing the *quality* of the child’s relationship. Blatantly expressed hostility in stems that measure relationships often characterize the child about whom others complain. Rarer but still indicating a severe relational problem is the totally withdrawn child who expresses this by short, unknowing types of responses. Such a child cannot relate how he or she feels about others or how they feel about him or her. Such a child often responds with “I don’t know.” The selfconcept is an empty shell. This behavior in the extreme may be a component of autism or Asperger’s syndrome, but it will more often be a consequence of emotional neglect or abuse. The child has simply not had ordinary relational opportunities. In other cases, when a child is referred for constant acting-out behavior, the assessor may find that the child has much more motivation to get along with people than others suspect; his or her motives are much more positive than people will believe. Absent are sentence completions that indicate strong negative feelings or ideas about others. Critical is that the sentence stems that often elicit *emotional* responses in the child with disturbed

relationships simply are responded to with other or milder content. Such a child may have disturbances in his or her behavior

#### 174 II. PROJECTIVE METHODS

but is not *emotionally* disturbed. The usual requirement for such a child is concrete assistance in acquiring skills needed to get along with peers and adults, sometimes a change in conditions and often short-term counseling. The teacher should be encouraged to explicate the concrete behaviors that are expected and not to assume the child knows them. Such a child is an ideal candidate for mentoring because he or she is often the product of an unstructured, fairly chaotic environment. One to one assistance is perfect for such a child.

The psychosomatic category is also quite obvious. The child usually checks off a list of physical problems, if asked. He or she typically openly complains in the sentence completion of physical symptoms or fears. Often adults working with the child have not taken these reports seriously, or the child may exhibit these symptoms so much that everyone is thoroughly tired of it. Parents may resent the suggestion that there is something physically wrong with their child that they have not attended to. Nevertheless, an assessor cannot dismiss reported physical symptoms. In most cases, real physical disorders will indeed be ruled out. But in some cases, overlooked and sometimes serious problems will be discovered. Headaches and stomachaches together, for example, seldom denote nothing, as we have previously discussed. Only a physical examination can answer physical questions, and one needs to be done to complete a responsible assessment that contains physical complaints, if these problems have not already been addressed and ruled out. When a physical basis for the problems is not established, there is then the possibility that the child qualifies as emotionally disturbed. Usually this diagnosis has been made simpler by the discovery of a strong anxiety state in the child.

#### **A Negative but Unclear**

##### **Signal: Defensiveness**

Whereas the presence of ADD/ADHD and stress are rule-outs (ADD/ADHD has its own category, and stress is not yet disturbance) and anxiety, depression, and thought disturbance, as well as severe relationship difficulties and psychosomatic behaviors,

are usually rule-ins for “emotional disturbance,” the presence of a high degree of defensiveness in the sentence completion protocols of children is a less clear and discrete signal. For the most part, the sentence completion protocols of children are generally nondefensive. Where there is the appearance of a high degree of defensiveness (e.g., stress), further examination needs to be pursued. Defensiveness will manifest itself as sparse, short, nonengaged, mundane kinds of responses. The child will be hesitant, both in overt and in verbal behavior, to engage in the task. His or her responses will be as unrevealing as possible.

Sometimes the sentence completion format, however, brings forth a rich production of responses in an otherwise defensive student. A student who is remarkably hostile in behavior toward others, including psychological examiners, may flow forth in the sentence completion with the most blatant statements: exact plans for homicide, hatred of individuals, retaliatory schemes, injuries, fantasies, and bitter feelings. This degree of revelation can also occur in the positive sense: A student who appears defensive in his or her interactions with other people will reveal a high level of conceptual social intelligence, plans for changing the world to a better place, personal dreams and plans for the future, and even a detached perspective on those with whom he or she is having difficulties. The results of the sentence completion are hard to predict, and this is one of the features that makes it valuable (and interesting!).

But there are students who react defensively to the sentence completion procedure itself. After all, the sentence completion format is straightforward, not disguised. Students know exactly what is wanted. (That is why they often take the opportunity to tell the examiner of their need for help—hoping something will come of their remarks.) But no matter how defensively the student reacts to the sentence completion, he or she cannot get around telling the examiner something. Not telling is also telling with the sentence completion.

An analysis of the items to which the student is responding defensively can be instructive; further investigation into the daily life conditions of the student is often thereby warranted (why do they feel it so necessary to hide information on the defended sub-

## 6. Sentence Completion as a Tool 175

ject?). And, again, as with all conditions, the apparent “defensiveness” may rest on other than purely psychological conditions: lowered verbal ability, inferior intellectual ability, depression, shyness, lack of trust of the examiner, and fear of what will be done with the test results. Also important but often overlooked is the fact that some subcultures do not approve of a person’s discussing intimate and family matters with strangers, and such children can be made extremely uncomfortable by the sentence completion. There are also the cases in which children have been warned by their parents, who may have given reluctant permission for the testing to be done, that they are not to discuss certain matters away from home. These children are so busy trying to remember all the categories they are not to discuss, they simplify matters by attempting not to disclose anything. It needs to be remembered that children have few rights. An adult brought into a strange room by a strange person who asks all kinds of personal questions and gives one all kinds of tests without so much as the person’s own permission: How would an adult react to such a situation? Yet this is precisely what happens to children. It is up to the examiner to be straightforward, emotionally neutral, and pleasantly businesslike. It helps as far as possible to treat the child as an equal. The examiner should tell the child plainly that he or she is a person who works with children having problems in school. He or she is there to see whether together they can figure out why the child is having problems. The child certainly *knows* he or she is having trouble in school—it is no surprise. Inviting the child into a process that may alleviate that awful fact usually sets a positive tone. It is useful to let children express their own views of the problem (views that often turn out to be fairly accurate!), and if they ask for an opportunity to find out what the tests told about them, the examiner should be sure to set up such a session and follow through. Finally, if after all of the examiner’s best efforts are made to obtain a meaningful sentence completion protocol and the results of that effort are sparse, he or she may switch to a more projective approach. It is sometimes the case that a child who produces little content on the sentence completion will produce effusive projective stories. Such

children are frightened by the straightforwardness of the sentence completion and consider that imagination, as they perceive what is involved in the projectives, is less readable. For others, the unstructured nature of projectives means that they will guard even more against providing any revealing information.

### **Individuality: Beyond the Rules**

The sentence completion, as we have discussed previously, can help to rule in or rule out a diagnosis of emotional disturbance. But it can do more. The sentence completion can provide data that will help to reflect the individual structure of the child. If people did not show some uniformity in their behavior, we could never predict them or develop classification systems of disturbances and illnesses, but it is also the case that we cannot completely predict the behavior of anyone. Each individual is uniquely made up of far more factors than we can measure and of qualities we fail to measure. The behavior of a person at any moment is a most complicated equation, made up of countless multiplicands for which we have no measure. Although a psychological assessment can certainly not take in all possible factors, the sentence completion can help to make the picture we generate of the student more accurate and more useful by noticing at least some of the following elements in the data.

### **Individual Qualities**

The individuality of the client will emerge in the sentence completions in the form of humor, imaginativeness, insight, tolerance, forgiveness, ambition, drive, social intelligence, talent, and many other qualities. Clearly, we cannot measure every student for all these and other positive and productive factors, although a number of checklist-type instruments do set out models of psychological functioning that attempt to measure the components variables in that model. Some of these instruments can be useful, particularly if the instrument is selected to measure the student for what is already somewhat obvious as a personality propensity. In opposition to the effort to try to measure everything, however, is simply the matter of noticing what is plainly there. Like cream,

### **176 II. PROJECTIVE METHODS**

strong talents and personality factors will rise to the top (though often no more than as potentials). The key is for the alert assessor not to ignore signs, not to consider them

“blips” on the screen while hunting for emotional monsters.

In the sentence completion, children often display realistic, surprisingly appropriate plans for their own future that take into account their particular abilities. Some children divulge needs and dreams they are too shy to have articulated anywhere else. Often these dreams could be pursued if children could be reinforced in their plans or appropriate support could be obtained. If a child’s strong desire is to learn to play the violin, for example, a violin can probably be found somewhere. Not all class clowns are acting out of pathology; there are some true comedians in the group, children with acute social perception. Why are they not in the drama club? And the boy with disabilities whose lifetime dream is to play football: Can he be a student assistant to the team (probably so, if the examiner approaches the coach and works it out). The individual characteristics of the student are potent elements that should be consciously used in any successful treatment plan. Motivation derives from these elements.

#### Idiosyncratic Reactions

For all kinds of idiosyncratic reasons, the meaning of one person’s experience may be different from that of the majority of people in the culture or subculture. For example, a sentence completion such as “People don’t know that I secretly enjoy missing parties” is not the standard teenage response. But it explains a lot in a “hurried child,” one whom the parents have overbooked to the point of having no free time at all. Even parties do not rank as high with this student as just having a free hour to herself.

Religious or philosophical teachings can also cause a student’s responses to differ from the typical cultural responses, and this can mislead the examiner. Consider the following examples: “When I do something good in school, . . . I am embarrassed.” “It is wrong . . . to compete with others.” If this is an able or even talented child who simply refuses to perform as well as he or she can, the difficulties may be originating from teachings in the family that posit that it is wrong to compete or show pride in oneself. Oversensitivity is another personal idiosyncratic way of responding to sentence stems. This is observed in sensitivity to stems that do not appear to affect other children in the same way. This style is often

generalized to many stems. There is much evidence today that this type of oversensitivity or shyness has a physiological basis, and such children may need help but do not need to be psychologically overinterpreted. Subcultural behaviors can also be misinterpreted. One teacher became upset and referred a child for psychological evaluation because he wrote a story about his grandmother swinging a chicken round and round and drew a picture of the headless chicken jumping in the grass, blood going everywhere. Perhaps the action of the grandmother was disturbing to the child and he was working out his concern in this manner, but the action of the grandmother was nothing but common behavior in the farming subculture in which the child lived—wringing a chicken's neck as the first step in preparing it for dinner. It certainly said nothing about the level of hostility in the child (or the grandmother).

Many sources contribute to making some of our personal characteristics different from the norm of the culture in which we live. Sentence stems that appear potentially pathological can rest on nothing more than misunderstanding. Furthermore, the source of some of our idiosyncracies will never be known. Then again, who really cares how a child came to love that awful butterscotch pudding served in the cafeteria? What is important is that the child will work 2 hours on math, which he hates, in order to have a bowl of the stuff on Friday. Like individual qualities, idiosyncratic responses can guide us toward or away from certain variables that can strongly affect our treatment plans.

#### **Personal Decisions and Strategies**

Many persons have arrived at decisions and strategies as to how they will handle certain instances or classes of stimulation. Many of these decisions and strategies are hardly unique (e.g., that is how we can understand and predict some common behaviors). For example, most people have decided to stop at a red light. But often the decisions and strategies of a person are unique to the person, originate in childhood, and comprise the stuff of which therapy is made.

Many personal decisions and strategies can be apprehended from the sentence completion responses, particularly in comparing one response to another or to a cluster of other responses. For example, the stem

“When my father gives me lots of work . . .” should be examined in relation to the stem, “When my teacher corrects me . . .”; a common pattern of compliance or resistance may emerge to demands of authority. (That is why the Haak sentence completion has categories of responses in which there are several stems.) But these responses that divulge decisions and strategies arrived at by the person do not always appear with the stem designed to elicit them. For example, one young child who was quite deprived of affection produced this response: “If I had all the books in the world, I would fill up my room to the ceiling and then I would have to sleep with my mother.” Almost any situation this child viewed was related back to his abject emotional need and was converted into strategies for meeting this need.

Decisions of the client can be seen anywhere in the sentence completion, but one especially telling example is the stem. “It is no use to . . .” If the response is “act silly” or some such inane and moralistic reply, it is not of the same level of meaning as the response “think things will get better.” This latter response implies that a decision has been reached (and one that cannot wait a long time for someone to change, in this case).

When decisions and strategies have been reached in a person’s mental makeup, it implies that the person has encountered a serious problem, usually over and over again, and has made a conscious decision about how to handle it. Some of the decisions are productive and some are of no real consequence, but others may be maladaptive or even dangerous. The boy who wished to fill up his room with books in order to sleep with his mother is going to be unable to handle the crushes that come with adolescence; he is going to place so much importance on the affection he finds in a girlfriend that he will in no way be able to let such a teenage crush just break up naturally. Some of the decisions and strategies that the child divulges in the sentence completion must be made part of his treatment plan: those that work for him or her need to be reinforced as compensatory strengths; those that are pathological or productive of present and future discord need proper therapy to be eliminated. Almost nothing could be more important in a treatment plan than the “long look forward.”

## Language Translations

Most people have an individualistic and unique way of handling some of the language. When these personal expressions become too bizarre, we become concerned about thought disturbance, for instance. But because language is the codification of experiences and all persons' experiences differ, language will always be somewhat personal and will provide a window into the person's unique experience. The ability to express this experience is not, however, uniform. A great deal of verbal ability is required for a person to be able to fully articulate his or her experience and feelings. Most persons can do this to some degree, but with many the examiner must become a translator in order to understand the significance of the child's codifying of his or her experiences. The simplest translation is the individual word itself. "When I am afraid, I hiccup." "When we have a math test, I hiccup." The personal meaning of "hiccup," and probably the actual behavior, is "am anxious." The examiner understands that math tests provoke a high level of anxiety in this student, although he may not have mentioned math anxiety elsewhere.

Common reactions to diverse experiences is a translator. A teenage boy responds, "When I am mad, I drive my car too fast." He also divulges in one sentence that "at Christmas time we never get much. My dad always drives us around town to see the lights after we have had our tree, but he drives so fast he scares me." The translation in this case is that the way the father handles depression has been picked up by the son. Emotions themselves can be translated—often a sign in itself of emotional disturbance. In a normally functioning adult, emotions are differentiated and stand alone; this is not altogether achieved by a child (or

178 II. PROJECTIVE METHODS

many adults). But even with children, major emotions should not be translated to any marked degree. The most common translation for children is the crossover of sadness and anger: "When I am sad, I am mad." Unhealthy translations also occur in how the emotions are handled. Far better is the somewhat defensive "When I am sad, I am sad" than "When I am sad, I am stupid." In translations there is a common denominator, often of which the student is unaware. "I am childish and do not care

enough about my grades.” “My mother thinks I am childish.” One can suspect a conflict with the mother over this child’s school performance even though such conflict is not otherwise stated. Common denominators often tie a number of responses together and provide for the examiner a major dimension of the assessment analysis.

Consider these examples: “I cannot work when I am tired,” “When my father yells at me, I feel tired,” “When a dog comes toward me, I feel tired,” “I could be happy if I wasn’t so tired.” All of these child’s responses seem reduced to “being tired,” which raises questions about depression and anxiety, a hypothesis that can be checked out other places in the data as well.

Many children have formed certain identifications that they do not openly express.

“My father was in the army.” “When I grow up, I want to join the army”—this from a child apparently having all kinds of trouble with his distraught father. The last thing on earth the father would suspect is that the child has an identification with him. Identifications, whether conscious or unrealized, are some of the more powerful psychological variables that exist. They can be unearthed by careful attention to translations in the sentence completion—the detective work in its analysis. We don’t always “know” what a behavior indicates from our catalogue of psychological stereotypes.

Sometimes we have to figure it out.

#### **THE PSYCHOLOGICAL REPORT AND TREATMENT PLAN**

The culmination of all assessment activities should be a usable psychological report and treatment plan. Simply put, when people have referred a troubled student for assessment they expect to find out “what is wrong” with the student and what to do about it. They deserve a report that draws a comprehensive picture of major dimensions of the student’s functioning, both productive and nonproductive, and a succinct, reasonable, clear, and even creative treatment plan tied to those major dimensions of functioning.

Additionally, if an assessment is done for possible school services, the student must be considered in the context of education and school. The Haak sentence completion is an instrument specifically created to evaluate a person within a certain context (its most common use, as we learned earlier). It reveals how major dimensions of the child’s

functioning are reflected in school behavior. Because some school systems like to follow a standardized report format for the report, and because the federal rules governing special education set out certain topics in the report that must be addressed, local or federal requirements for the report may be less than appropriate in addressing the psychological organization of the child from an examiner's point of view. Nevertheless, the analysis of the psychological examination can still be worked out and reported succinctly as a major component of the report, and an organized treatment plan can be outlined, as below.

Too many psychological assessments are mere laundry lists, citing item after item and interpretation after interpretation with little effort devoted to the creation of organization or internal consistency. The sentence completion if misused can contribute to such an undesirable outcome. The examiner may use the sentence completion as a kind of "bank," picking and choosing certain sentences to quote in support of hypotheses generated elsewhere. This is an inappropriate use of the sentence completion. It is like quoting the Bible, selecting words and sentences out of context (one such report was received by an attorney of the writer's acquaintance who threw it on the desk exclaiming in frustration, "And what am I supposed to do with *this*?").

A useful assessment should usually produce from two to four or five major dimensions of concern in the child's functioning, as previously stated, both negative and positive, and explain how these concerns play

6. Sentence Completion as a Tool 179  
out in the daily life of the child. Persons working with the child should understand clearly after reading the psychological report what the dimensions of major concern are and how individual actions of the student, particularly the ones about which they are most concerned, are related to these major dimensions of behavior. This involves a lot of old-fashioned work with a pencil, clustering behaviors and drawing lines. It calls on everything the examiner has been taught and, in truth, on everything the examiner has ever seen or heard filtered through the concepts of psychology. The good (i.e., useful) psychological report is like bread: its smooth consistency is the result of much kneading and working (with

perhaps a little bit of time to let some of the issues “rise” in the mind of the examiner, if that is not taking the analogy too far).

The treatment plan should address the needs of the student that derive from his or her major dimensions of behavior and how these needs are to be met. This is often a place for creativity. The federal rules specifically prohibit dishing out only those treatments an institution has in its cupboard.

Obviously, if specialized services are needed, as, for example, physical therapy, that need will be indicated with minimal language in the psychologist’s report. The physical therapist will write the detailed, professional report for that area. But the needs that have no standard, available service are those that call on the creativity of the assessor (and the ones that are often disregarded for want of a ready fix).

Let us suppose that a child is diagnosed who has much verbal skill, a strong motivation to help others, and a desire to be a leader, but he is pushy, loud, and awkward and completely alienates his peers. What should we do? We must try to reduce the need to its barebones dimensions: The child needs to be taught how to be a social leader. Now remedies suggest themselves: bibliotherapy; a mentor who is a community leader; short-term counseling. What about getting the highly popular, civic-minded junior class president to spend some sessions with this student through the peer-tutoring program? If anybody knows how to be a peer leader, clearly he or she has figured it out. Solutions come more creatively when the bare essentials of what is needed are stated in plain English (not within some conceptual system). It is surprising how often creative solutions can be found.

The whole psychological assessment process is a venture in creativity. The artist gathers and evaluates his or her materials, forms a plan, and decides what he or she needs to do to bring the plan to fruition. So does the psychological assessor: He or she measures the relevant properties of the child, decides on a treatment plan (with input from important others), and enumerates what actions or resources are needed to bring the plan to fruition. The outcome of this plan is intended to be a happier, freer, more productive child. In such a laudable enterprise, the sentence completion technique

can play a useful part.

## REFERENCES

- Achenbach, T. M. (1991). *Child Behavior Checklist*. Burlington: University of Vermont, Department of Psychiatry.
- Agesen, N., Brun, B., & Skovgaard, B. (1964). Rotter's Sentence Completion Test. *Nordisk Psykologi* (Danish), *36*, 188–200.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- American Psychological Association. (1967–2000). PsycINFO [A computerized database of psychological abstracts]. (Available at any major library)
- Baggerly, T. N. (1999). Adjustment of kindergarten children through play sessions facilitated by fifth grade students trained in child-centered play therapy procedures and skills. *Dissertation Abstractions International*, *60*(6-A), 1918.
- Baker, D. (1988). *Establishing inter-rater reliability and criterion-related validity for the Haak Sentence Completion*. Unpublished doctoral dissertation, Texas A&M University.
- Costin, F., & Eiserer, P. E. (1949). Students' attitudes toward school life as revealed by a sentence completion test. *American Psychologist*, *4*, 289.
- Bower, E. (1969). *Early identification of emotionally handicapped children in school*. Springfield, IL: Charles C Thomas.
- Dean, R. S. (1983). *Individual Ability Profile*. Odessa, FL: Psychological Assessment Resources.
- Dean, R. S. (1984). Commentary on "Personality assessment in the schools: The special issue." *School Psychology Review*, *13*, 95–98.
- Derichs, G. (1977). Sentence completions as intake instruments. *Praxis der Kinderpsychologie und Kinderpsychiatrie* (German), *26*, 142–149.
- Erickson, E. H. (1959). Identity and the life cycle. *Psychological Issues*, *1*, 1–171.
- ### 180 II. PROJECTIVE METHODS
- Haak, R. A. (1990). Using the sentence completion to assess emotional disturbance. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context* (pp. 147–167). New York: Guilford Press.
- Haak, R. A. (1996). *Haak Sentence Completion (Elementary)* (rev.). Unpublished instrument.
- Haak, R. A. (2000). *Haak Sentence Completion (Secondary)* (rev.). Unpublished instrument.
- Hart, D. H. (1986). The sentence completion technique. In H. Knoff (Ed.), *The assessment of child and adolescent personality* (pp. 245–272). New York: Guilford Press.
- Hart, D. H., Kehle, T. J., & Davies, M. V. (1963). Effectiveness of sentence completion techniques: A review of the Hart Sentence Completion for Children. *School Psychology Review*, *12*(4), 423–434.
- Havighurst, R. J. (1953). *Human development and education*. New York: Longmans, Green.
- Ilg, F. L., & Ames, L. B. (1960). *Child behavior*. New York: Dell.
- Irving, F. S. (1967). Sentence completion responses and scholastic success or failure. *Journal of Counseling Psychology*, *14*, 269–271.
- Kaufman, A. L., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children (K-ABC)*. Odessa, FL: Psychological Assessment Resources.
- Knoff, H. M. (1963). School based personality assessment.

*School Psychology Review*, 12, 391–393.

Loevinger, J. (Ed.). (1998). *Technical foundations for measuring ego development: The Washington University Sentence Completion Test*. Mahwah, NJ: Erlbaum.

Payne, A. F. (1928). *Sentence completions*. New York: New York Guidance Clinic.

Peck, R. F., & Haak, R. A. (1973). *Coping skills of early adolescents*. Unpublished study.

Piaget, J., & Inhelder, B. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Reinis, S., & Goldman, J. M. (1980). *The development of the brain: Biological and functional perspectives*. Springfield, IL: Charles C Thomas.

Sanford, R. N., Adkins, M. M., Miller, R. B., & Cobb, E. A. (1943). Physique, personality and scholarship. *Monographs of the Society for Research in Child Development*, 8(1, Series No. 34).

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *The Stanford–Binet Intelligence Scale IV*. Itasca, IL: Riverside.

Wechsler, O. (1991). *Wechsler Intelligence Scale for Children, Third Edition*. San Antonio, TX: Psychological Corporation.

Wells, B. (2000). *The use of the Haak Sentence Completion Measure and the Child Behavior Checklist/Teacher Report Form by school psychologists in the identification of students with serious emotional disturbance*. Unpublished doctoral dissertation, University of Texas.

Wilson, L. (1949). The use of the sentence completion test in differentiating between well-adjusted and maladjusted secondary school pupils. *Journal of Consulting Psychology*, 13, 400–402.

## 6. Sentence Completion as a Tool 181

The Rorschach test has one of the most controversial histories of any instrument in the era of psychological testing. Although the controversy is still far from over, the Rorschach is continuing to gain popularity among psychologists as the instrument of choice when their purpose is to understand and describe the psychological features of an individual. This renewed popularity is evidenced in the increasing numbers of training institutions offering graduate courses in the Rorschach; in the number of internship sites requiring students to have training and skill in using the Rorschach; and in the increasing frequency with which the Rorschach is being used in clinics, hospitals, and schools. This renewed popularity is due in large measure to the continuing efforts of John E. Exner, Jr. and his associates at the Rorschach Research Foundation. The Exner Comprehensive System is the most widely used system of administering, scoring, and interpreting the Rorschach, according to data provided by Piotrowski (1996). This chapter presents a brief introduction to the Rorschach, using the procedures and norms established in Exner's Comprehensive System (Exner, 1974, 1978, 1986, 1991,

1993; Exner & Weiner, 1982, 1994) and a discussion of the use of the test with special populations of children and adolescents. A thorough knowledge of the aforementioned references is essential if one is to use the Rorschach properly and wisely, as are intensive study and training in administration, scoring, and interpretation.

#### THE EXNER COMPREHENSIVE SYSTEM

Three factors that contributed to the Rorschach controversy, according to Exner and Martin (1983), were “(1) confusion and disagreement concerning the theoretical underpinnings of the Rorschach, (2) the lack of a single, consistent administration, scoring and interpretation procedure, and (3) the oversimplified classification of the Rorschach as a projective technique” (p. 407). Exner and his associates have expended much effort in addressing the second and third critical factors.

Although projection may be present in an individual’s verbalizations about what the inkblots might be, Exner defines the Rorschach as a problem-solving (perceptual–cognitive) task, similar to the way Rorschach himself considered it to be a

182

# 7

## Using the Rorschach with Children and Adolescents: The Exner Comprehensive System

JERRY C. ALLEN

JILLAYNE HOLLIFIELD

“form interpretation test” (Exner & Martin, 1983, p. 407). If projection is present, the examiner may wish to use that material as supplemental to the formal, psychometrically based data obtained from the Structural Summary of the Comprehensive System.

Thus, the Rorschach may now be viewed as a test, administered under standard procedures and scored according to established criteria. These scores (or codes) may then be compared with the scores obtained by an appropriate reference group.

The Comprehensive System originated from Exner’s attempt to develop a single, consistent procedure of administration, scoring, and interpretation of the Rorschach test, based on empirically defensible data. The history of this endeavor can be found in

Exner and Martin (1983). The outcome resulted in the development of the system (Exner, 1974, 1978, 1986, 1991, 1993; Exner & Weiner, 1982, 1994). The development is continuing today through efforts of the Rorschach Research Foundation and other professionals' research contributions.

### **ADMINISTRATION**

Two skills are of prime importance to the Rorschach examiner who chooses to work with children and adolescents. The first is a thorough understanding of childhood developmental processes; the second is a high level of knowledge and expertise with the instrument itself and the Comprehensive System. These skills are essential to obtain complete and accurate data during the administration phase, which in turn permit appropriate scoring and interpretation. Because of the differences in verbal skill levels inherent in the various developmental levels of childhood, and because Rorschach data collection is almost entirely dependent on verbal exchange, the administration process can be quite challenging. Nevertheless, the individual child's verbal skills are always uniquely revealing.

### **Preparing the Child**

Because it is so important to be particularly careful in collecting Rorschach data from children, it is advisable for the examiner to spend adequate time with the child prior to administration to put the child at ease. Children generally have some idea as to why they are being tested, but often they do not know how the information gained in the assessment will be used. It is not uncommon for them to develop negative forms of resistance, and consequently, to perform less than optimally. Thus, it is wise for the examiner to take as much time as possible with the child, explaining the nature and purpose of the test and indicating how the results will be used, as well as answering any questions the child or the parents may have. This should be done in as honest and straightforward a manner as the situation permits, and without unnecessary elaboration. Children value honesty and will generally respond with minimal resistance to an examiner who has taken adequate preparation time before the procedure begins.

### **The Response Process**

It is vital that the examiner understand the nature of the Rorschach response process—what is involved in the formation of a response

and its ultimate delivery. Exner and his colleagues have studied in detail the processes an individual goes through in making a response from the moment the person is asked "What might this be?" and is handed the inkblot until the first response is delivered. Exner (1986) summarizes the operations as follows:

Phase I: 1. Visual input and encoding of the stimulus and its parts.

2. Classification of the stimulus and/or its parts, and a rank ordering of the many potential responses that are created.

Phase II: 3. Discarding potential answers that have low rankings.

4. Discarding other potential answers through censorship.

Phase III: 5. Selection of some of the remaining responses because of traits or styles. (p. 51)

The crucial elements in assuring the cooperation of the child and a valid Rorschach protocol are, as in the case of all assessment procedures, the time and effort the examiner expends in putting the child at ease prior to the examination proper.

7. Using the Rorschach with Children and Adolescents 183  
Exner and Weiner (1982) stress that the response process is a

complex interaction among at least three interrelated variables: (1) the set of the subject toward the test and the testing situation, (2) the evaluation of the subject regarding which of several perceived responses is most appropriate or "correct" in light of the set, and (3) the impact of the composite of response tendencies or styles plus the ongoing psychological operations of the subject. (p. 19)

Thus, the necessity of preparing the child adequately to yield the most appropriate test set is evident and will usually produce a valid record even when some amount of resistance to the task remains.

The next variable in the response process of the child involves his or her evaluation of which of all the perceived responses is most appropriate to the established test set. Children, as well as adults, see more potential responses than they are willing to verbalize. In fact, when a sample of children was encouraged to give as many responses to a card as they could in 60 seconds, they averaged slightly more than 94 total responses for the test (Exner, Armbruster, & Mittman, 1978). Nevertheless, most children, regardless of socioeconomic status or environmental

setting, restrict the number of responses they deliver. This restriction is related to the evaluation process and involves both rankordering and censorship operations (Exner, 1986). That is, on the basis of the established test set, the child selects from among numerous potential responses those that seem most appropriate or correct and censors those that do not.

Not surprisingly, Leura and Exner (1978) found that examinees do not restrict as much and, in fact, give significantly greater numbers of responses to examiners who are known to them. The largest increases are seen in Human Movement responses (*M*), Color responses (*FC*, *CF*, and/or *C*), and Blends, with significant increases in Whole (*W*) and Uncommon Detail (*Dd*) locations. Nevertheless, the proportions of both *M* and *C* responses to total number of responses remains unaffected because of the increased number of responses generated overall. Leura and Exner found the greatest difference in proportion occurring in the gray-black Shading responses. Persons whose examiners were known to them gave about 11% Shading responses, whereas the control group gave about 15%. Consequently, examiners who test children they know well should be alert to the possibility that these children's profiles may reflect some or all of the features just described. Other studies related to response productivity have yielded mixed results with regard to the potential influence of sex of the examiner (i.e., Greenberg & Gordon, 1983; Tuma & McCraw, 1975). Thus, examiners may wish to consider the potential for gender effects in Rorschach response productivity.

Another factor that may influence both the quantity and quality of children's responses is their need for reinforcement.

Children are far more concerned than adults about giving the correct responses. They are motivated by needs to please the examiner, to perform well, to do the socially acceptable thing, and to avoid potentially unpleasant outcomes. Reassuring a child that there are no right or wrong answers is not always sufficient to provide the reinforcement the child needs. Even adults have been shown to be strongly influenced by the perceived social acceptability of their responses. Exner and Leura (1976) demonstrated that adults would report that certain sexual and aggressive responses were "easy to see" when they

had received prior instructions that these responses were frequently given by successful businessmen. A second group of adults indicated that these same sexual and aggressive responses were "difficult to see" when prior instructions implied that they were produced by seriously disturbed persons. Children, too, want to do the right thing and to gain approval for their performances or, at the very least, to avoid any undesirable consequences.

The inherent lack of structure in the Rorschach is problematic for the child. Depending on the stage of development, or in some cases on developmental lag, children are to one degree or another concrete in their thinking and prefer situations in which there are rules or guidelines. Ambiguity is troublesome. As mentioned earlier, to attempt encouragement by telling the child that there are no right or wrong answers does not remove the necessity of dealing with an ambiguous situation. Exner and Weiner (1982, 1994) point out that, in fact,

#### 184 II. PROJECTIVE METHODS

the only "right" answer is that it is an inkblot. By coaxing the child to report what it might be, or by encouraging the reluctant child with "Most children see more than one thing," the examiner engages the child's preferred problem-solving operations and coping mechanisms. Therefore, although the test itself causes the surfacing of test sets, it also requires that the child use his or her unique psychological response style in the perceptual, cognitive, and affective spheres. These will be apparent in the responses generated by the child, regardless of any residual resistance.

Children's psychological response styles tend to be relatively consistent, although far less so than those of adults. Their response styles tend to become increasingly stable over the developmental years and, like those of adults, are susceptible to situational variables.

Stability of a psychological response style is determined by its utility in reducing stress for the child. Consequently, change in the cognitive, perceptual, and affective operations of the child is to be expected, and the younger the child at the time of initial testing, the more the psychological response style will change with time (Ames, Metraux, Rodell, & Walker, 1974).

The child's psychological response style is highly unique and individualistic and determines which of the many responses that

could be articulated will be selected in the child's problem-solving operations. For example, in giving the popular response to Card V, the child could report "a butterfly" (Form or *F*); "a black butterfly" (Form/Achromatic Color), a response in which form is the primary determinant but achromatic color is also used; "a butterfly flapping its wings" (Animal Movement or *FM*); or "a soft, velvety-looking butterfly" (Form/Texture or *FT*), a response in which form is the primary determinant but texture is also used. It can also be seen in whether or not the child prefers to respond to the blot as a whole (*W*) or to common (*D*) or unusual detail (*Dd*) areas. It can also be seen in whether or not the child attempts to organize details of the blot in the response. Furthermore, the child's psychological response style is apparent in the use of pairs, reflections, or special scores or contents, or in the failure to use these. Because children's psychological response styles are at best only relatively consistent, it is of utmost importance that the examiner know which of the Rorschach test variables are relatively consistent and which are highly vulnerable to change over time.

A study by Exner, Rosenthal, and Thomas (1980) underscores the lack of consistency in many of the child's psychological operations. In this study, two groups of children, initially ages 6 and 9, were examined at specified intervals over time. The only variables with high test-retest correlations over a 2-year period were *X+*% (perceptual accuracy for the total record), frequency of Active Movement, and frequency of Popular (*P*) responses. All other variables showed great disparity over the same interval. Moreover, there were substantial differences between the two groups of children. It is highly unlikely that a child's current protocol will predict accurately his or her future psychological response style.

Children change over time; as noted earlier, the younger the child at the time of initial testing, the more change is to be expected over even relatively short intervals. Older children and adolescents exhibit less change, and their records are apt to be more stable over intervals. In particular, *X+*% is consistent in the early years, perhaps even prior to school entry (Exner & Weiner, 1982), whereas Human Movement (*M*) production is not consistent until midadolescence.

Thus, the preferred problem-solving or coping style tends to fluctuate during the child's first 10–12 years, and a dominant style is not formed until somewhere between 14 and 18 years of age.

Thus, it is apparent that knowledge of both the instrument and the developmental process are essential to competent Rorschach administration with children and adolescents. The examiner must spend adequate preparation time with the child to ease tensions and minimize resistance, so that a clear picture of the psychological operations of each child may be obtained. The examiner must be aware that children evaluate and censor potential responses on the basis of how they expect their test data to be used, and also on the basis of what sorts of responses they believe are “correct” or will please the examiner. Examiners must also be alert to the potential effects of testing children who are well known to them

7. Using the Rorschach with Children and Adolescents 185  
and to possible gender effects in Rorschach response productivity. Finally, the examiner must be aware of the high degree of inconsistency in children's Rorschach records over time and the considerable potential for change in this age group.

### **Administering the Rorschach**

The actual formal administration of the Rorschach test is a relatively straightforward task; nevertheless, certain basic guidelines must be followed. The examiner must adhere to the standard administration practices in order to insure accurate scoring and interpretation. Practice is required so that the examiner can avoid inadvertently establishing unwanted test sets or reinforcing any particular type of response. The preferred seating is side by side. Seating can be altered to accommodate very young children or those with special needs (e.g., sitting on the floor), as long as standard administration is maintained.

The formal administration of the Rorschach consists of two phases: the “Administration Proper” and the “Inquiry.” In the Administration Proper phase (often termed the “Free Association” phase), after appropriate preparation of the child for the test, the examiner hands each of the inkblots in sequence to the child, who is asked, “What might this be?” Responses are recorded verbatim; therefore, the examiner needs to develop some type of shorthand method for

recording responses. After responses to all 10 cards have been obtained, the Inquiry phase begins. The purpose of the Inquiry is to allow the examiner to obtain all information needed to score the response appropriately (see Exner, 1986, pp. 70–78, for details on conducting the Inquiry).

As normally the Rorschach is just one of a battery of instruments given to the child, its description should be brief and similar to that given for other procedures. For example, the examiner may say, “And then we will be doing the inkblot test. Maybe you’ve heard of it.” If the child’s parents are present, the explanation should be delivered in a manner that insures understanding for both the parents and the child. Subsequent questions should be answered directly and honestly.

During the administration proper, there is no need to change the standard instructions, except in the case of very young children or children with language difficulties, for whom even more simplified instructions may be in order. The examiner should refer to the cards as inkblots; that is what they are. Typical instructions are as follows: “I am going to show you some inkblots and I would like you to tell me what this might be.”

Children tend to ask a lot of questions, and it is best to answer them in as direct and straightforward manner as possible.

“Where did you get these?” “I bought them.” “Do you think it looks like that?” “Right now I’m interested in what you think it looks like.” “Did I get that one right?” “If that’s the way it looks to you.” “Do other kids say things like that?” “People say all sorts of things.” “What kind of kids do you show these pictures to?” “All sorts of kids.” Some questions require lengthy explanations or are complex and should be deferred until the end of the test: “That’s a pretty complicated question. Let’s wait until we finish and then I’ll explain. I’ll write it down so I won’t forget.”

Instructions for the Inquiry should be equally straightforward and direct. “We are going to go through the cards again now, and I will read to you what you said. I want you to show me where on the blot you saw what you saw, and then tell me why it looks like that to you. I want to see it just the way you did.” The Inquiry is the most difficult part of the administration and much practice

will be necessary before one becomes proficient at it. Generally, if the examiner has given clear instructions at the outset of the Inquiry, scoring can proceed without difficulty; remember, the purpose of the Inquiry is to allow the examiner to obtain all the necessary information so that the response can be scored accurately. Some children require more encouragement and questioning even when adequate instructions have been given. "Yes, I know it looks like that to you but help me see it the way you do. Remember, I need to see it too just like you do." "What in the blot makes it look like that to you?" "Draw a line around the \_\_\_\_\_ with your finger." The examiner should not become directive, however. If the attempt at clarification is not productive, examiner and child should move to the next response. It is important to avoid establishing test sets.

#### 186 II. PROJECTIVE METHODS

**Some Typical Administration Problems**  
Occasionally, a child is extremely resistant and defensive. In this case, the examiner must decide whether to proceed with the test or to postpone it until a later date. Reviewing two areas of the administration process may be helpful in making this decision. First, have sufficient time and effort been spent in establishing rapport with the child? If not, it is possible that the problem can be remedied within the session. Second, and only in extreme cases, the examiner should consider the possibility of altering the administrative process to accommodate such difficulties as hyperactivity, unusual anxiety, and the like. Ames and colleagues (1974) and Halpern (1953), for example, advocate the use of the Inquiry immediately after each card, particularly with preschool and easily distracted children. However, Exner and Weiner (1982, 1994) caution that this procedure should be used only as a last resort, as Leura and Exner (1977) have found a resulting elevation in Common Detail (*D*) and Form (*F*) responses, especially on cards VI and X, when immediate Inquiry is used.

Another problem frequently encountered by the examiner is the brief record. Young children tend to give relatively short records, and it is not uncommon for 5- and 6-year-olds to give 15 or fewer responses. Moreover, children generally give more pure *F* responses than do adolescents or adults,

although they usually give enough responses with a sufficient variety of determinants to be interpretively useful. When this is not the case, the examiner should consider several possibilities, including intellectual or neurological difficulties. When defensiveness or resistance is present, the child is often able to control his or her affect sufficiently to yield a record that is not interpretively useful. The examiner then needs to assess whether sufficient time was spent in establishing rapport with the child or whether the use of other procedures might be more productive. When none of these possibilities is plausible, however, the presence of a brief and barren record signals the question of a psychological style in which the child copes by oversimplifying the stimulus field (Exner & Weiner, 1982, 1993). That is, the child may be ignoring important environmental cues and oversimplifying complexity to keep his or her world manageable, but at the price of almost inevitable conflict. Occasionally, the examiner will find that a child who has been somewhat resistant or defensive in the Free Association phase will respond more freely in the Inquiry phase. Klopfer (1956) and Klopfer, Ainsworth, Klopfer, and Holt (1954) have indicated the importance of noting behavioral differences in the child's performance on the two portions of the test as providing valuable information to the interpretation. Exner and Weiner (1982) suggest that by the time the Free Association has been completed, and with the additional structure of the Inquiry, the child may have become more relaxed. Thus, it is advisable to record any additional responses generated in the Inquiry, while not including these as scored responses, to supplement the interpretation. In some instances, when the scorable record is extremely brief and barren (less than 14 responses), it may be advisable to retest the child. Studies by Hulsart (1979) and Exner (1980) have shown that interpretively useful records may be obtained after brief intervals. A viable and recommended procedure is simply to say at the completion of the Free Association: "Good. Now that you have the idea of how to do these, let's go through them again. Most children see more than one thing, and you may too, now that you know how to do these." This procedure is often all that is necessary to yield an interpretively useful record.

## NORMATIVE DATA

All the major Rorschach methods currently in use in the United States include some mention of administrative procedures and interpretation guidelines for use with children.

However, only two scoring systems provide normative data obtained from stratified samples of children at various ages and exclusively designed to facilitate interpretation of children's records (Ames et al., 1974; Ames, Metraux, & Walker, 1971; Exner, 1986; Exner & Weiner, 1982). Ames and her colleagues have based their work

7. Using the Rorschach with Children and Adolescents 187  
on the scoring system of Dr. Marguerite Hertz.

The Exner Comprehensive System provides norms for children ages 5 through 16, as well as adult norms, thus giving added flexibility to the examiner who works with both children and adults. The child and adolescent norms are based on a sample of 1,580 nonpatient children who were recruited through schools and other organizations. Sample selection was partially stratified for socioeconomic and geographic status (Exner, 1986, p. 255). Normative tables provide the following information for 70 of the scoring variables at each year level: mean, standard deviation, mode, minimum and maximum range, frequency, skewness and kurtosis. These data, both for children and adolescents, are helpful to the examiner in serving as baseline data.

Although the availability of normative data specifically designed for use with children is a substantial aid to interpretation, norms, like all other standards for normal performance, must be used intelligently. Digressions from the norms may be signs of uniqueness or ideographic ways of looking at the world and may be in no way abnormal. Conversely, a child can produce a normal record and still lack or be deficient in certain adaptive skills. The normative data itself may underrepresent certain groups or be insensitive to developmental differences. Finally, because many of the Rorschach scores occur infrequently, they are not normally distributed, and rigid application of the normative data is inappropriate. In these instances, the examiner should rely more heavily on either the mode or the frequency with which the responses is likely to occur for the age of the child in question.

## SCORING

A full explanation of the Comprehensive System's scoring procedures is beyond the scope of this chapter; the interested or unfamiliar reader might consult the book describing the system (Exner, 1986). A summary of the scoring variables is given in Exner and Martin (1983, pp. 410–411).

The Comprehensive System currently includes approximately 90 possible scores (or codes) for the Rorschach responses. There are seven major categories into which the scores are placed:

1. Location (Where on the blot is the response seen?);
2. Determinants (What features in the blot contributed to the formation of the percept?);
3. Form Quality (How "good a fit" is the percept to the area of the blot to which the response was offered?);
4. Content (What is the content of the response?);
5. Popularity (Is the response one that is commonly given?);
6. Organizational Activity (To what degree does the response organize the blot elements?); and
7. Special Scores (Does the response include unusual verbal material or unusual characteristics?).

Each response a child makes to the inkblots is scored according to specified criteria.

These scores are then expressed as percentages, frequencies, or ratios on the "Structural Summary" and are compared to the normative data for purposes of interpretation.

One of the strengths of the Comprehensive System is its continuing research program investigating the psychometric qualities of the scores remaining on the Structural Summary.

Exner (1986) reported that more than 30 reliability studies of temporal consistency had been completed at the Rorschach Research Foundation by 1983. Time intervals between test and retest have varied from a few days to 3 years. Exner and Martin (1983) report that in the longest test–retest study, using 100 nonpatient subjects, reliability coefficients for 20 critical variables were uniformly high for that amount of elapsed time. For 10 variables, the coefficients were between .80 and .90; for 7, between .73 and .80; and only 3 fell at or below .70. Short-term reliability coefficients are even higher. Many studies have also been conducted investigating interscorer reliability for each of the variables. Exner (1986) reports that a standard of 90% agreement among scorers or a .25 intercorrelation among scorers was used as a standard (p.

89). The reader should be aware of ongoing criticism related to interscorer reliability in the Comprehensive System as well as other reliability and validity issues (see Exner, 1996; Wood, Nezworski, & Stejskal, 1996a, 1996b).

## 188 II. PROJECTIVE METHODS

### INTERPRETATION OF CHILDREN'S RECORDS

#### General Considerations

Skilled interpretation demands a thorough understanding of the developmental process, together with a complete comprehension of what the normative data can be expected to provide. Rorschach data have essentially the same meaning at all ages; however, the particular frequency, percentage, or weight of a response varies in interpretive significance with age. Nevertheless, the examiner must not apply the normative data concretely but, instead, must be sensitive to the unique qualities of the individual scores. Moreover, examiners must be familiar with psychopathology and with personality theory. This background is essential to making determinations regarding the prognosis and the form of treatment or remediation most likely to elicit a positive response from the child.

Children's Rorschach scores are not consistent over time; thus, the examiner must think in terms of probabilities rather than predictions, except in the case of severe pathology. Exner, Thomas, and Mason (1985) have demonstrated that only the  $X+$  is relatively stable from ages 8 through 16. In the same study, the majority of Rorschach variables were inconsistent until approximately ages 14–16, and even then the ability to modulate affect, the extent of self-focus, the effort expended in organizing, and some of the factors involved in cognitive slippage remained unstable.

Among younger children, the *EB* (Erlebnistypus or preferred response style) generally reflects an extratensive style. When this persists beyond age 12, it is highly likely to become a stable feature. On the other hand, the establishment of an introversive style appears to occur somewhat earlier and the child who demonstrates an introversive style at age 8 may well maintain it (Exner, Leura, Wylie, Armbruster, & Thomas, 1980). Similarly, the *es* (Experienced Stimulation) of young children frequently is greater than the *EA* (Experience Actual); however,

when the *EA* does become greater than the *es*, it tends to remain so in subsequent protocols. Depending on the age of the child,  $EA > es$  can be either a positive or a negative finding. This shift is expected during the developmental process and generally signals psychological maturity. A premature shift, however, could portend rigidity and inflexibility (Exner, Leura, et al., 1980). The *es* tends to be relatively stable, but because it consists of some highly unstable elements (Inanimate Movement and Diffuse Shading), it warrants the examiner's careful attention. When any of the unstable elements of the *es* are present in a child's record, particularly if the element appears in greater than the expected number, the stability of the *es* and the direction of the  $EA : es$  are both suspect. The examiner must determine whether the scores are truly representative of the child or whether they reflect temporary situational phenomena.

The interpretation of the  $EA : es$  ratio, therefore, must take several factors into account. First, the age of the child must be considered in conjunction with the appropriate normative data to establish the expected directionality of the ratio and range of values for the component scores. The examiner should note the magnitude of the difference between the *EA* and *es* values, as well as the frequencies and kinds of variables that comprise the *es*, with particular attention to those that are transitory and situational. The presence of Texture (*T*) in a child's record is vital (Leura & Exner, 1976; Pierce, 1978). Elevations in *T* can signal personal loss, whereas the absence of *T* is highly uncommon and may indicate either a tendency toward defensiveness and guardedness in interpersonal relationships or the absence of a need for or expectation of interpersonal closeness. The examiner must integrate the child's history and background information with the data to ensure meaningful and useful interpretation.

Another aspect of Rorschach interpretation where findings are unique to children is the  $FC : CF + C$  ratio. Young children do not (and are not expected to) modulate their expressions of affect, resulting in a ratio that is weighted on the right. As the developmental process progresses, the capacity for control increases, and a gradual shift from right to left generally occurs. The composition

of the scores on the right side of the ratio, however, undergoes a rather dramatic change at about age 6, when the occurrence of pure C and Color-Naming (CN) responses diminishes markedly. Thus, although a  $FC : CF + C$  ratio weighted to the right is not uncommon at age 12, the majority of responses on the right side should be  $CF$ . When the capacity for affective modulation develops in the child, it tends to persist and is reflected in a shift in the weight of the ratio, from right to left. Once again, the age of the child, the magnitude of the difference between the values for the right and left sides of the ratio, and the integration of the child's history and background are vital to accurate interpretation. An  $FC$  weighting in a child of 12 years of age or less is often premature and may signal overcontrol. Moreover for certain youngsters, an  $FC$  weighting even at age 16 may be problematic (Exner, Leura, et al., 1980).

Another common occurrence in the records of children and adolescents is the Personalized Response ( $PER$ ). This tends to be frequent in the records of very young children but disappears gradually through the course of development. The appearance of one such response in an adolescent record is not unusual; however, a high frequency of  $PER$  responses suggests defensiveness and a need to protect the self-image.

Therefore, the examiner should note the frequency of these responses relative to the age of the child, as well as the overall quality of the child's  $PER$  responses. Some responses are distinctly more defensive than others and yield substantial indications of the child's unique vulnerabilities.

The  $W : M$  ratio will also show some characteristic variability in the records of children. Commonly, the left side of the ratio will be markedly higher—particularly in the very young child, where grandiosity and limitless aspirations prevail. During the course of normal development the ratio is expected to decrease gradually until it approaches adult levels in late adolescence.

The examiner should be alert for a  $W : M$  ratio that is too constrained in relation to the age of the child.

Gordon and Tegtmeier (1982) substantiate the use of the Egocentricity Index [ $3r + (2/R)$ ] as a measure of self-focusing as opposed to self-esteem in children. However, the authors indicate that possibly these results

were an artifact produced by limiting the number of responses to two per card. These same authors (Tegmeyer & Gordon, 1983) found that the children who gave a high number of space (*S*) responses also gave significantly more *W* and total Shading responses, and more blends and a greater number of Content categories, than children who gave no more than one *S* response. They concluded that relatively high frequencies of *S* responses appear to be related to cognitive complexity and active mastery in children.

In a somewhat different approach to the question of cognitive complexity, Smith (1981) found a positive relationship between children's cognitive-developmental stage and the number and percentage of *W* responses. In addition, the level of cognitive functioning and *W* production was found to vary with the complexity of the stimulus. That is, children who had achieved the developmental stage of formal operations showed a marked increase in the number and complexity of *W* responses to the broken blots, compared to children in the concrete stage.

#### The Rorschach and Children from Disadvantaged Backgrounds

Some special considerations are in order when interpreting the records of children from low-socioeconomic-status (low-SES) backgrounds. Exner and Weiner (1982) have delineated specific findings for sex, age, and SES. Generally speaking, there appear to be some differences in the records of children from low-SES backgrounds, but these differences are not consistent across all the samples. The authors caution that although the Structural Summary may not yield differences among the various SES groups in terms of location, determinant, or content category, substantial variation may be found in the verbal material or verbal expression of these children that is not reflected in their scores. Although a child's cognitive and perceptual operations certainly influence the content of his or her verbal material, interpretations of the verbal material should be done with extreme caution (Exner & Weiner, 1982).

#### 190 II. PROJECTIVE METHODS

A study of inner-city children (Krall et al., 1983) found a lower *F* accuracy level, a lower percentage of *W* responses, and a higher frequency of *D* responses among

these children when compared to both the Ames and colleagues (1974) and the Exner and Weiner (1982) normative data. The authors also found, however, that the children in their study were similar to those in both data bases with regard to response rate and development of the *P* response.

Although caution is always warranted in content interpretation, a study of the aggressive content in the Rorschach records of inner-city children (Crain & Smoke, 1981) yielded information that may be of use to the prudent examiner. Children in the control group gave aggressive content related to interpersonal interactions and equal power, such as fighting, whereas children in the clinical group gave aggressive responses characterized by victimization and feelings of being overwhelmed, such as devouring monsters.

Other studies involving content analysis of children's Rorschach records, although not limited to children from lower-SES backgrounds, have produced findings that may be used to augment the basic interpretation. Gordon and Tegtmeier (1983) found that oral-dependent responses in the records of children do not have the same interpretive significance that they do in the records of adults. Rather, these responses in the records of children tend to be associated with various internalizing behaviors, such as withdrawal, somatic complaints, and obsessive and compulsive behaviors.

### **The Rorschach and Children with Learning Disabilities**

The Rorschach test has been shown to be sensitive to many of the variables that affect children's ability to learn. For example, Ames and Walker (1964) demonstrated that children with cognitive flexibility, as indicated on Rorschach protocols collected in their kindergarten years, were better readers in the fifth grade. These children were found to be more open to information in the environment and more able to use these cues in their learning. Conversely, Smock (1958) showed that when anxiety is artificially induced in children, it can promote rigidity and premature closure on cognitive tasks. Subsequently, Smock and Holt (1962) found significant negative correlations between perceptual rigidity and IQ, school achievement, and curiosity.

Although children with learning problems do not necessarily demonstrate diminished

intelligence or curiosity, they frequently show marked perceptual difficulties and other impediments to school achievement. In addition to documented learning disabilities, learning problems can include limited intellectual functioning and various types of neurological impairment. When a child has demonstrated limited intellectual functioning, the Rorschach may or may not be useful. Other measures may better address the education and training needs of these youngsters. On the other hand, the Rorschach can occasionally be useful, particularly with the higher functioning of these children, in detecting personal assets that can be used to the best advantage in educational interventions. Certain Rorschach features are common to this group of children; they include low number of total responses, predominance of pure *F*, few or no *M* responses, few or no chromatic *C* responses with poor integration, few or no Synthesized responses, low *F*+% and *X*+%, narrow range of Contents, low number of *P* responses, and a higher than usual occurrence of *CN* responses (Exner & Weiner, 1982).

When a child has a neurological impairment and/or attention-deficit disorder, the Rorschach can be a useful addition to the assessment battery because of the complex emotional and behavioral features attendant upon these difficulties (see Bartell & Solanto, 1995). Testing can reveal the child's preferred coping mechanisms, level and availability of controls, ability to direct cognitions, interpersonal attributes, and personality assets, all of which will be helpful in determining the most appropriate methods of interventions and education (Exner & Weiner, 1982).

Certain specific features of the Rorschach records of children with neurological impairment have been noted. One of these is a lower frequency of *M* responses. Gordon and Oshman (1981) suggest that this is related to the inability to delay in hyperactive children. Champion, Johnson, McCreary, and Dougherty (1984) propose that the lower-than-average Egocentricity Index frequently found in these children's records is related to the social stigma accompanying their learning problems. In addition, these authors attribute the lower *X*+% to problems with perceptual accuracy. Scott (1985) indicates that

the data in the “Four Square” can be extremely useful in formulating interventions.

Williams and Miles (1985) found that dyslexic children (some from the United States and some from the United Kingdom) in their sample gave Rorschach records that were very similar to one another’s but unlike those of other clinical populations or a matched control group. Specifically, the dyslexic children gave fewer responses, a high percentage of *F* responses, and a limited range of Determinants. Also, they rarely turned the cards. Although the generalizability of these findings is limited by the small number of youngsters in the sample, the authors noted that there were essentially no differences between the responses of the children from the United States and those from the United Kingdom, regardless of age or type of school attended.

A study of two groups of children with learning disabilities (Champion et al., 1984) revealed that the response patterns of children with learning disabilities were distinctly different from Exner and Weiner’s (1982) normative data for nonpatients, behavior problems, or withdrawn children. Compared to nonpatients, children with learning disabilities had lower *F*+% and *X*+%, lower Egocentricity Indices, higher Lambdas, more *Dd* and *S*, and more Shading responses. In the group of 8-year-old children with learning disabilities, the mean number of *H* percepts was at least one standard deviation below the mean of the normative sample. This group also gave more *C*\_, *V*, and *Y*, and fewer *T* than nonpatients of the same age. Children with dyslexia in the 11-year-old group gave more of all types of Shading responses, and their mean Lambda was at least one standard deviation above the mean for the 11-year-old norm group.

### The Rorschach and Children with Depression

When depression is suspected in a child, critical information can be obtained from a thorough and dependable history and from the behavioral observations of the parents and teachers. Adults, however, are not always alert to the signs of childhood depression or are preoccupied with personal concerns and fail to attend sufficiently. Parents engrossed in personal concerns may overlook the child’s cues. The teacher may misinterpret the child’s “perfect” behavior.

Consequently, from time to time, the child’s

Rorschach data will show signs of depression even when there is no report of substantiating evidence.

In adults, the two most common features found in the records of depressives are a low Egocentricity Index and elevation in the right side of the *eb* (Experience Base), or a sum of the shading responses that is greater than  $FM + m$ . Occasionally, the Egocentricity Index will be elevated rather than depressed. This is the exception rather than the rule, but in either case, there is a problem in self-focusing and evaluation of the self. When the depression is reactive rather than chronic, however, the Egocentricity Index frequently falls in the normal range (Exner, 1978, 1986). Adult depressives also frequently give more *V* (Vista) and *FD* (Form-Dimension) responses than are normally expected.

The Rorschach records of children who are depressed show features similar to those of adults, although it is unusual to find both a depressed Egocentricity Index and a higher right-sided *eb* in the record of a child. When this occurs, it is also likely that there will be an elevation in Morbid Content (*MOR*) responses (Exner & McCoy, 1981). Children who are depressed, whether chronic or situational, tend to give more *MOR* responses than do those who are not. Nevertheless, *MOR* does appear in the records of children who are not depressed at all. Thus, the features commonly associated with childhood depression are more than one *V* or *FD* response, the presence of even one Color-Shading blend, a low Egocentricity Index, *eb* elevated on the right, and an unusually high frequency of *MOR* content responses (Exner & Weiner, 1982). Exner and Weiner (1982, 1994) state that when four of these five are present, the child is experiencing either distress or, more likely, depression.

The examiner must then decide whether the condition is reactive or chronic. Both *V*

## 192 II. PROJECTIVE METHODS

responses and the Egocentricity Index are not particularly responsive to situational effects and take a long time to form. Thus, if the child's record contains both *V* responses and a low Egocentricity Index, the condition is probably chronic. Conversely, if the child's record contains *m* and *Y* responses with no elevation in *V* or depression of the Egocentricity Index, the condition is more likely to be situational or reactive.

A more complex circumstance exists when an earlier situational stressor in the life of the child has not been effectively resolved. In these instances, both  $V$  responses and a low Egocentricity Index can appear together with  $m$  and  $Y$  in the child's record, even when there is no evidence of enduring problems.

There are several other Rorschach indices that may attend depression in children. The Affective Ratio may be low because of the child's tendency to withdraw and avoid emotional situations. The frequency of  $S$  responses may be elevated. A low Lambda indicates that the child has become too involved in complexity, which actually aggravates the depression. An elevated  $W : M$  ratio indicates that the child cannot or will not economize in using resources; this is confirmed by an elevated  $Zf$ . If only the  $Zf$  is elevated, it is important for the examiner to note how much of the elevation is accounted for by  $W$  and how much consists of more complex responses. Finally, when the child's  $X+\%$  is below what is normally expected, the examiner should ascertain whether this is a function of unique and ideographic responses ( $Xu\%$ ) or actual distortion ( $X-\%$ ). The number of  $P$  responses in the child's record can be useful in determining the significance of a low  $X+\%$  (Exner & Weiner, 1982).

Obviously, when a child is depressed, the question is not only whether the experience is situational or chronic but also how acutely the distress is felt by the child. More precisely, what is the likelihood that the child will act on this experienced distress and attempt or succeed in committing suicide? Unfortunately, there are no infallible guidelines.

Exner (1978) has constructed a Suicide Constellation for Adults and attempted to construct one for children. However, the Rorschach Workshops (1987, p. 6) have issued a severe caution concerning the Children's Suicide Constellation. A crossvalidation study was conducted using a total sample of 51 cases, 36 of whom attempted and 15 of whom completed suicide. The results did not cross-validate, and Exner concluded that the "constellation is psychometrically worthless" and was not to be used. Rather, the Adult Suicide Constellation possesses a "true positive" hit rate of approximately 67% for children ages 15 and 16. The same is not true for children

below this age.

### The Rorschach and Childhood Schizophrenia

The diagnosis of schizophrenia is, of course, never made on the basis of the Rorschach record alone. A thorough and reliable history and the behavioral observations of parents and teachers are invaluable supplements to information obtained in the assessment. Part of that information will be the examiner's own observations of the child during the examination period. Thus, the examiner must be able to identify the salient behavioral features of schizophrenia. There are four major areas that pose difficulty for the individual with schizophrenia: cognition, perception, interpersonal relationships, and the ability to exert appropriate controls in each of these endeavors.

The cognitions of the child or adolescent with schizophrenia are often marked by difficulties in maintaining clear and coherent associations, strained or deviant logic, and inappropriate levels of abstraction. Initially, and particularly in the younger child, these cognitive difficulties are extremely subtle, and the examiner may attribute their occurrence to inattentiveness, anxiety, and the like. However, the logic of the child with schizophrenia tends to be highly circumstantial, with frank dissociations or associations that are tenuous at best. For example, when asked his name, a boy with schizophrenia may respond, "My name is Bill, but my dog's name is Charlie. I don't know where he is, though." Although this sort of thinking is common in young children, it usually disappears by age 7 (Exner & Weiner, 1982, 1994). Moreover, these children have marked deficiencies in their ability to abstract at appropriate levels. They are inclined to polarize, being either too concrete or too abstract. For example, in response to the comment, "Time is money," the child may say, "No, it's three o'clock." At the other extreme, these youngsters become preoccupied with abstraction, such as "This bat looks more peaceful, young, like he's got these big wings but doesn't know what to do with them yet, just trying to experience." Manifestations of these sorts of cognitive difficulties will appear in the language of the child during the Free Association and the Inquiry. Neither Ames and colleagues (1971, 1974) nor any of the other major

Rorschach systematizers deal with schizophrenia across children's age groups as succinctly as does the Comprehensive System (Exner, 1978, 1986, 1991, 1993; Exner et al., 1985; Exner & Weiner, 1982, 1994). Therefore, Rorschach findings unique to the protocols of youngsters with schizophrenia will be expressed in the scoring terminology of the Exner method.

In addition to evidence of cognitive disturbance in the child's general manner of expression, the response language of the child will contain scorable features that range on a continuum from mild to severe. At the mild extreme are slips of the tongue and redundancies (*DV*) and Incongruous Combinations (*Incom*), such as "blue chickens." Deviant Responses (*DR*), such as inappropriate or irrelevant phrases or circumstantial responses, and Fabulized Combinations (*FABCOM*), such as "two cats sitting on a rocket," are in the midrange of severity. At the severe extreme of the continuum are the Contamination (*CONTAM*) and Inappropriate Logic (*ALOG*) responses. *CONTAM* responses represent the fusion of two percepts into one, as exemplified in the response, "I see a kitty and a bird. It's a catbird." *ALOG* responses, on the other hand, contain strained or circumstantial logic, such as "They must be in love because this heart is between them." In terms of interpretive significance, the presence of one or even several of the milder responses is expected in a child's record, but the presence of only one *CONTAM* is a serious indicator of pathology.

Moreover, each of these types of scores has an inherent range of possible responses from mild to severe. For some of these special scores, Exner (1986) has developed a system for differentiating levels of severity, with Level I scores indicating mild to moderate cognitive slippage and Level II scores suggesting more severe slippage.

The child's cognitive difficulties may include an inappropriate level of abstraction.

On the Rorschach, this is often manifested in a preoccupation with abstraction and symbolism. It may take the form of an unusually high frequency of responses containing conventional symbolism, or it may emerge in responses containing highly personalized ideographic symbolism. In either case, the child's overuse of or preoccupation with symbolism increases the possibility of a

thought disturbance. It should be noted, however, that in the intelligent and otherwise healthy child, moderate use of symbolism and/or abstraction is probably a positive indicator.

Youngsters with schizophrenia also demonstrate distorted perceptions that lead to poor judgment. They tend to misperceive both themselves and others. An adolescent girl with no training or skills will insist that she is a highly talented designer who simply has no outlet for her superior gifts, for example, and will refuse to consider employment opportunities appropriate to her background. Or, a casual remark, such as "It's been nice to see you," will be misconstrued as a serious romantic innuendo.

On the Rorschach, distortions in perception are measured by the Form Quality, the  $F+\%$ , and the  $X-\%$  (Distorted Form). Scores below .70 for both the  $F+\%$  and the  $X+\%$  generally signal impaired perceptual functioning, and scores below .60 are always indicative of inaccuracy. In a brief Rorschach record, it is advisable to rely on the  $X+\%$ , because the  $F+\%$  may be representative of only a few responses. When the  $X+\%$  is below .70, the  $X-\%$  will indicate whether the low  $X+\%$  is the result of idiosyncratic responding (as represented in Form Quality scores of "unusual") or whether it is the result of frank distortion (as represented in "minus" Form Quality scores). Ideographic responding does not have the same prognostic implications as does seriously impaired perceptual accuracy. The examiner should scan the sequence of scores to determine whether clusters of "unusual" or "minus" responses occur in relation to particular determinants or con-

194 II. PROJECTIVE METHODS

tents. When patterns occur, it is highly likely that the impaired perception is linked to specified precursors as opposed to overall inaccuracy, and interventions are far more apt to be successful.

Another Rorschach indicator of perceptual accuracy is the  $P$  response, which measures the child's ability to see things as others do, or to respond conventionally.

Youngsters age 11 or older who deliver less than four  $P$  responses have an impaired ability to respond conventionally. Before age 11 this is not always true, and the examiner must use the normative data carefully to determine the significance of a low

number of *P* responses in the individual child. Moreover, *P* responses occur most frequently to Cards I, III, V, and VIII; thus if the child gives only three or four *P* responses, they should occur on these cards. When the child does this, the concern regarding perceptual distortions and poor judgment is somewhat diminished. When the child delivers three or four *P* responses to cards other than these, the concern is magnified considerably. Because of disordered thinking and misperceptions of self and others, youngsters with schizophrenia have poor social skills and therefore great difficulty in forming and maintaining interpersonal relationships. Even when participating in group activities, these children tend to be distant and emotionally withdrawn. Often they are physically withdrawn as well. The child's history and the behavioral observations of parents and teachers will provide vital information regarding patterns of interpersonal functioning, as well as onset and duration of any deviations from established patterns. Naturally, the assessment process itself provides an opportunity for the examiner to engage the child in various interactions directly. The primary Rorschach indicators of interpersonal difficulties are inferior *M* and *H* production. Although conservatism is warranted when interpreting the records of young children, the absence of *M* in a child's record generally signals a serious withdrawal from interest in people. Moreover, *M* is expected to appear with good Form Quality. A child, even as young as age 5 who delivers an *M* response with "minus" Form Quality will tend to assign inaccurate and illusory meanings to social situations and to have deficient social skills (Exner & Weiner, 1982, 1994). Other, less severe indicators of the child's difficulty in dealing with people are *M* responses with fictional or mythological human or animal content or human or animal detail content. Because it is possible for the child to give responses with human content but without human movement, the presence of good *H* in the child's record is a positive sign in the young child or in the child whose operations are not yet sufficiently complex to produce *M*. However, the absence of *H* or the presence of *H* in conjunction with *M*- is interpreted in essentially the same way as the absence of *M* or the presence of *M*- responses. An additional problem for the child with

schizophrenia is the inability to exert appropriate controls in the cognitive, perceptual, and interpersonal spheres. These children are preoccupied with anxiety-producing thoughts, largely of sexual and aggressive content. Moreover, they are unable to integrate their thoughts and feelings properly, resulting in either blunted or unsuitable expressions of affect. These difficulties are expressed in Rorschach contents of cruelty, violent aggression and flagrant sexuality.

Whereas the average child reports aggression in terms of "fighting" or some type of *MOR*, the child with schizophrenia delivers responses that reflect gore and brutality.

Similarly, although sexual responses are rare in a child, they normally tend to represent developmental concerns when they do occur.

Sexual responses occur more frequently in the records of children with schizophrenia and tend to reflect a disturbed preoccupation.

These children's inability to exert effective controls is often apparent in an *FC* :

*CF + C* ratio that is weighted to the right.

Throughout the developmental process, there is a gradual shift of the weighting from the right to the left side of the ratio.

Nevertheless, the appearance of pure *C* in the record after age 8 warrants investigation and signals the possibility of episodes of unmodulated affect.

Once the examiner has established substantial support for the presence of a thought disorder in a child, the question of an acute versus a chronic process remains. Generally speaking, when the frequency of scores on variables measuring cognitive disturbance, 7. Using the Rorschach with Children and Adolescents 195

perceptual inaccuracy, interpersonal difficulties, and problems with control are relatively

low and do not deviate excessively from the norms, the process is likely to be an acute and reactive one, probably in the early stages. In this context, the presence of increased numbers of *INCOM*, *FABCOM*,

and *ALOG* responses may be indicative of the child's attempt to manage and make sense of his or her world, in contrast to the apathy typical of the chronic process. Acute onset is usually related to a precipitating event or a situational precursor; the child is aware of his or her diminished capacities and is upset and anxious in this regard. These youngsters often spontaneously recognize the inadequacy of their responses.

The hallmarks of schizophrenia are disordered

thinking and perceptual inaccuracy. Although bizarre contents and peculiar language may augment these findings, they may not substitute for more concrete evidence of serious cognitive slippage and impaired perception. Moreover, Exner and colleagues (1985) demonstrated that many of the cognitive and perceptual operations unique to schizophrenia are highly resistant to change, even with therapeutic intervention and even over extended periods of time.

#### A LAST WORD

Now, a word of caution. We are all familiar with the need to select and use instruments with current, up-to-date norms, but we are also familiar with problems that arise when old tests are modified and new norms are established (e.g., the original vs. revised versions of the Wechsler Intelligence Scale for Children, or the Stanford–Binet, Form L-M vs. the Fourth Edition). As Exner (1986) suggests, “It would be foolhardy to suggest that the work is finished. The Rorschach continues to pose many unanswered questions” (p. 4). Although the Comprehensive System is currently the most widely used approach to the Rorschach, controversy about the system and the Rorschach continues today, particularly around such issues as interrater reliability, reliability of administration and recording of responses, and questions of validity of certain scores and indices (i.e., Exner, 1996; Wood et al., 1996a, 1996b). Research on the Exner Comprehensive System of the Rorschach is an ongoing process that results in modifications in scoring criteria, changes in norms for some scores, and refinement of definitions for some variables. The Rorschach Research Foundation (Rorschach Workshops) publishes periodic newsletters, updating for previous participants any changes that have been made in the system and keeping them abreast of ongoing research. However, not all individuals trained in the Exner system receive this information. The caution, then, is that all users of the system should keep current with the progress being made.

#### REFERENCES

- Ames, L. B., Metraux, R. W., Rodell, J. L., & Walker, R. N. (1974). *Child Rorschach responses* (rev. ed.). New York: Brunner/Mazel.
- Ames, L. B., Metraux, R. W., & Walker, R. N. (1971). *Adolescent Rorschach responses*. New York: Brunner/Mazel.
- Ames, L., & Walker, R. (1964). Prediction of later

- reading ability from kindergarten Rorschach and I.Q. scores. *Journal of Educational Psychology*, 55, 309–313.
- Bartell, S. S., & Solanto, N. V. (1995). Usefulness of the Rorschach Inkblot Test in assessment of attention deficit hyperactive disorders. *Perceptual and Motor Skills*, 80, 531–541.
- Champion, L., Johnson, P. J., McCreary, J. H., & Doughtie, E. B. (1984). Preliminary investigation into the Rorschach response patterns of children with documented learning disabilities. *Journal of Clinical Psychology*, 40, 329–333.
- Crain, W. C., & Smoke, L. (1981). Rorschach aggressive content in normal and problematic children. *Journal of Personality Assessment*, 45, 2–4.
- Exner, J. E. (1974). *The Rorschach: A comprehensive system* (Vol. 1). New York: Wiley.
- Exner, J. E. (1978). *The Rorschach: A comprehensive system: Vol. 2. Recent research and advanced interpretation*. New York: Wiley.
- Exner, J. E. (1980). But it's only an inkblot. *Journal of Personality Assessment*, 44, 562–577.
- Exner, J. E. (1986). *The Rorschach: A comprehensive system. Vol. 1. Basic foundations* (2nd ed.). New York: Wiley.
- Exner, J. E. (1991). *The Rorschach: A comprehensive system. Vol 2: Interpretation* (2nd ed.). New York: Wiley.
- Exner, J. E. (1993). *The Rorschach: A comprehensive system. Vol. 1: Basic foundations* (3rd ed.). New York: Wiley.
- Exner, J. E. (1996). A comment on “The comprehensive system for the Rorschach: A critical examination.” *Psychological Science*, 7(1), 11–13.
- 196 II. PROJECTIVE METHODS**
- Exner, J. E., Armbruster, G. L., & Mittman, B. L. (1978). The Rorschach response process. *Journal of Personality Assessment*, 42, 27–38.
- Exner, J. E., & Leura, A. V. (1976). *Variations in the ranking of Rorschach responses as a function of situational set* (Workshops Study No. 221). Unpublished manuscript, Rorschach Workshops.
- Exner, J. E., Leura, A. V., Wylie, J. R., Armbruster, G. L., & Thomas, E. A. (1980). *A longitudinal Rorschach study with children* (Workshop Study No. 207). Unpublished manuscript, Rorschach Workshops.
- Exner, J. E., & Martin, L. S. (1983). The Rorschach: A history and description of the Comprehensive System. *School Psychology Review*, 12, 407–413.
- Exner, J. E., & McCoy, R. (1981). *An experimental score for morbid content (MOR)* (Workshop Study No. 269). Unpublished manuscript, Rorschach Workshops.
- Exner, J. E., Rosenthal, N., & Thomas, E. (1980). *A Rorschach study involving retesting several times at brief intervals* (Workshop Study No. 270). Unpublished manuscript, Rorschach Workshops.
- Exner, J. E., Thomas, E. A., Mason, B. J. (1985). Children's Rorschachs: Description and prediction. *Journal of Personality Assessment*, 49, 13–20.
- Exner, J. E., & Weiner, I. B. (1982). *The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents*. New York: Wiley.
- Exner, J. E., & Weiner, I. B. (1994). *The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents* (2nd ed.). New York: Wiley.
- Gordon, M., & Oshman, H. (1981). Rorschach indices of children classified as hyperactive. *Perceptual and Motor Skills*, 52, 703–707.

- Gordon, M., & Tegtmeier, P. F. (1982). The Egocentricity Index and self-esteem in children. *Perceptual and Motor Skills*, 55, 335–337.
- Gordon, M., & Tegtmeier, P. F. (1983). Oral-dependent content in children's Rorschach protocols. *Perceptual and Motor Skills*, 57(3, Pt. 2), 1163–1168.
- Greenberg, R., & Gordon, N. (1983). Examiner's sex and children's Rorschach productivity. *Psychological Reports*, 53, 335–337.
- Halpern, F. (1953). *A clinical approach to children's Rorschachs*. New York: Grune & Stratton.
- Hulsart, B. (1979). *The effects of a second chance instructional set on the Rorschach of emotionally disturbed and culturally deprived children*. Unpublished doctoral dissertation, Long Island University.
- Klopfer, B. (1956). *Developments in the Rorschach technique* (Vol. 2). Yonkers, NY: World.
- Klopfer, B., Ainsworth, N. D., Klopfer, G., & Holt, R. (1954). *Developments in the Rorschach technique* (Vol. 1). Yonkers, NY: World.
- Krall, V., Sacks, H., Lazar, B., Rayson, R., Growe, G., Navar, L., & O'Connell, L. (1983). Rorschach norms for inner city children. *Journal of Personality Assessment*, 47, 155–157.
- Leura, A. V., & Exner, J. E. (1976). *Rorschach performance of children with a multiple foster home history* (Workshop Study No. 220). Unpublished manuscript, Rorschach Workshops.
- Leura, A. V., & Exner, J. E. (1977). *The effects of Inquiry after each card on the distribution of scores in the records of young children* (Workshops Study No. 265). Unpublished manuscript, Rorschach Workshops.
- Leura, A. V., & Exner, J. E. (1978). *Structural differences in the records of adolescents as a function of being tested by one's own teacher* (Workshops Study No. 247). Unpublished manuscript, Rorschach Workshops.
- Pierce, G. E. (1978). The absent parent and the Rorschach "T" response. In E. J. Hunter & D. S. Nice (Eds.), *Children of military families* (pp. 71–87). Washington, DC: U.S. Government Printing Office.
- Piotrowski, C. (1996). The status of Exner's Comprehensive System in contemporary research. *Perceptual and Motor Skills*, 82, 1341–1342.
- Rorschach Workshops. (1987). *1987 alumni newsletter*. Asheville, NC: Author.
- Scott, R. S. (1985). Exner's Four Square: Useful index in appraisal of LD? *Perceptual and Motor Skills*, 60, 525–526.
- Smith, N. N. (1981). The relationship between the Rorschach Whole responses and level of cognitive functioning. *Journal of Personality Assessment*, 45, 13–19.
- Smock, C. (1958). Perceptual rigidity and closure phenomenon as a function of manifest anxiety in children. *Child Development*, 29, 237–247.
- Smock, C., & Holt, R. (1962). Children's reactions to novelty: An experiential study of curiosity motivation. *Child Development*, 33, 631–642.
- Tegtmeier, P. F., & Gordon, N. (1983). Interpretation of White Space responses in children's Rorschach protocols. *Perceptual and Motor Skills*, 57, 611–616.
- Tuma, J. A., & McCraw, R. K. (1975). Influences of examiner differences on Rorschach productivity in children. *Journal of Personality Assessment*, 39, 362–368.
- Williams, A. L., & Miles, T. R. (1985). Rorschach responses

of dyslexic children. *Annals of Dyslexia*, 35, 51–66.

Wood, J. N., Nezworski, M. T., & Stejskal, W. J. (1996a). The comprehensive system for the Rorschach: A critical examination. *Psychological Science*, 7(1), 3–10.

Wood, J. N., Nezworski, N. T., & Stejskal, W. J. (1996b). Thinking critically about the comprehensive system for the Rorschach: A reply to Exner. *Psychological Science*, 7(1), 14–17.

## 7. Using the Rorschach with Children and Adolescents 197

### The Holtzman Inkblot Technique (HIT)<sup>1</sup>

was developed to overcome psychometric limitations in the Rorschach by constructing completely new sets of inkblots. Although much of the early research on the Rorschach was either irrelevant or poorly conceived, an impressive number of well-designed validity studies generally yielded negative results.

The growing realization that the Rorschach had inherent psychometric weaknesses came to a head in a symposium on failures of the Rorschach that was sponsored by the Society for Projective Techniques (Zubin, 1954).

The evidence to date seems to indicate that the HIT has answered most, if not all, of these criticisms of the Rorschach (Holtzman & Swartz, 1983). By the end of 1999, more than 800 publications using the HIT had appeared in the world literature (Swartz, Reinehr, & Holtzman, 1999). HIT research to date, therefore, generally verifies the initial promise of this newer inkblot technique (Sundberg, 1962).

### DESCRIPTION OF THE TECHNIQUE

The HIT attempts to capture the best qualities of both the projective and the psychometric approaches to the Rorschach. Unlike the Rorschach, which has only 10 inkblots in a single form, the HIT consists of two parallel forms, A and B, each of which contains 45 inkblots constituting the test series and 2 practice blots (X and Y) that are identical in both forms. The inkblots were drawn from a large pool of several thousand, many of which were created by an artist working with special papers and inks that produced brilliant colors and rich shading. Only about 1 blot in 50 survived initial screening by a group of judges who were familiar with the Rorschach.

Selection of inkblots for the final version of the HIT was aimed at maximizing the reliability of these scores, as well as maximizing the discriminatory power of the final forms in differentiating superior normals from mental hospital patients in a series of standardization studies (Holtzman, Thorpe,

Swartz, & Herron, 1961b). The two parallel forms were constructed by pairing blots on stimulus qualities as well as item characteristics from the scores, and then randomly assigning members of each pair to either Form A or Form B. The final order of presentation for the 45 inkblots in each form was arranged so that most of the “best” inkblots appear rather early in the series. In the major standardization program,  
198

# 8

## Use of the Holtzman Inkblot Technique with Children

WAYNE H. HOLTZMAN

JON D. SWARTZ

using printed versions of the original inkblots, nearly 2,000 individual protocols were collected on samples ranging from 5-year-olds to mature adults and from chronic schizophrenic patients to mentally retarded individuals. Through the cooperation of psychologists in other settings across the country, 15 different, well-defined populations were sampled to provide the standardization data. Five of these samples were normal children—122 children age 5 from nursery schools in Austin, Texas; 60 children in second through sixth grades from a middle-class, private school in Austin; 72 fourth graders from Hamden, Connecticut; 197 seventh graders from four Texas communities other than Austin; and 72 adolescents in 11th grade from Chicago high schools. This last sample was given both the Rorschach and HIT in a comparative study of the two methods (Bock, Haggard, Holtzman, Beck, & Beck, 1963).

Percentile norms subsequently were published for emotionally disturbed children and adolescents and for male juvenile delinquents (Hill, 1972). Additional normative data for four representative groups of children and adolescents seen in clinical practice have been published by Morgan (1968).

The norms for emotionally disturbed children published by Hill are based on HIT protocols collected by Conners (1965) in clinical studies of 99 emotionally disturbed children and 114 neurotic adolescents. The norms for male juvenile delinquents are based on 75 cases collected by Megargee (1965) and replicated and extended by

Mullen, Reinehr, and Swartz (1983). When taken together with the earlier sets of norms for mentally retarded and normal children, the percentile norms that are available for the HIT provide a rich source of helpful information to be used in psychodiagnosis and personality assessment.

The standardization data also were used for a number of methodological studies, including investigations of scorer agreement, internal-consistency reliability, test-retest stabilities, and intergroup differences as a preliminary basis for differential diagnosis. The results of these studies, together with percentile norms and recommendations for use of the HIT in clinical assessment or for research, were published in 1961 (Holtzman et al., 1961b). While retaining the clinical sensitivity of the Rorschach, the HIT yields 22 standardized variables that can be objectively defined, reliably scored, and efficiently handled by statistical methods. For the first time, it allows the clinician, the psychometrician, and the experimentalist to work with the same projective technique.

Greater reliability and objectivity of scoring on the HIT are obtained because of the large number of inkblots and the fact that the subject is limited to one response per blot. A scoring guide (Holtzman, Thorpe, Swartz, & Herron, 1961a) further increases objectivity by making interscorer differences negligible on most of the major variables. The existence of two parallel forms of the same technique permits accurate retesting to evaluate change over time.

Since the publication of the individual version of the HIT by the Psychological Corporation in 1961, hundreds of studies have been reported in the world literature (Swartz et al., 1999). Our own work has focused largely on the development of a group version (Herron, 1963; Holtzman, Moseley, Reinehr, & Abbott, 1963; Swartz & Holtzman, 1963; Swartz, Witzke, & Megargee, 1970); a quick group version (Swartz & Reinehr, 1983); a computer method for scoring and interpreting HIT variables (Gorham, 1967; Holtzman, 1975; Vincent, 1982); a new experimental version of the HIT (Holtzman, 1988; Swartz, 1992); and a major cross-cultural study of over 800 children in Mexico and the United States, involving repeated measures with alternate forms of the HIT 6 years in a row (Holtzman, Diaz-Guerrero, & Swartz, 1975).

Some of the HIT findings from the crosscultural study that are particularly relevant in the assessment of children are presented later in this chapter.

### TEST MATERIALS

Standard materials for the HIT consist of two parallel series, Form A and Form B; the accompanying printed Record Forms and Summary Sheets; and the *Guide for Administration and Scoring*. The inkblots are

printed on thin but tough white cardboard  $5\frac{1}{2} \times 8\frac{1}{2}$  inches in size. Cards X and Y contain practice blots that usually are not scored. These two cards appear at the beginning of both Forms A and B. Card X is a massive achromatic blot that looks like a bat or butterfly to most people. Few subjects reject this card, although some prefer to use a smaller area than the whole blot.

Card Y is suggestive of a person's torso to most subjects. Red spots of ink introduce the subject to color and often evoke responses such as "spots of blood," either given alone or interpreted with the torso.

Cards 1 and 2 in both Forms A and B are achromatic and sufficiently broken up to make a whole response difficult unless there is integration of detail, or unless the subject gives a vague concept or one in which the form of the concept fails to fit the form of the inkblot. Both cards have popular responses in smaller areas of the blots, helping to break up a response set to give only wholes. Card 3 is irregular in form and has a large red "sunburst" splotch overlaid on an amorphous black inkblot. It is difficult to give a form-definite, form-appropriate whole response to Card 3 because of the chaotic, unstructured nature of this inkblot.

Card 4 is just the opposite, containing several finely detailed popular concepts that can be interrelated, together with color and shading that produces a vista-like effect. "A battle scene" or "a cowboy watching a sunset" is a typical response to Card 4A, and "a knight carrying a spear and shield" is a typical response to Card 4B.

Cards 5A and 5B are asymmetrical, grayish-colored blots unlike any in the Rorschach. By penetrating the charcoal-like quality of these blots, one can distinguish a number of detailed objects. Together with several similar, rather wispy, amorphous, asymmetrical blots later in the series, these cards are difficult, particularly for the individual

who is searching for definite concepts having good form or who wishes to use the entire blot.

The remaining inkblots cover a wide range of stimulus variation, giving the individual ample opportunity to reveal certain aspects of his or her mental processes and personality by projecting thoughts onto otherwise meaningless inkblots. Twelve of the inkblots in Form A are black or gray; 2 are monochromatic; 11 are black with a bright color also present; and the remaining 20 are multicolored. Most of the blots have rich shading variations, which help to elicit texture responses. A similar distribution of color, shading, and form qualities is present in Form B.

### ADMINISTRATION AND SCORING

Standard procedures for administering the HIT have been developed so that published normative data may be used as aids to interpretation.

Instructions to the subject have been designed to make the task as simple as possible while eliciting sufficient information to score major variables reliably. The instructions differ from those for the Rorschach in the following ways: (1) The examiner instructs the examinee to give only one response per card; (2) a brief inquiry is given immediately after each response; (3) permissible questions by the examiner during inquiry are limited both in number and in scope, and are asked rather routinely to avoid inadvertent verbal conditioning of certain determinants or content.

Three kinds of questions are permissible as part of the brief inquiry in the standard administration.

The actual wording used can vary a great deal, so that the inquiry becomes a natural part of the conversation between examiner and subject. Typical phrasing should be as follows:

“Where in the blot do you see a \_\_\_\_\_.”

“What is there about the blot that makes it look like a \_\_\_\_\_?”

“Is there anything else you care to tell me about it?”

After establishing rapport, the examiner picks up the cards one at a time, handing each one in an upright position to the subject.

The instructions given to the subject should be informal and should stress the following points: (1) These inkblots were not made to look like anything in particular; (2) different people see different things

in each inkblot; and (3) only one response for each card is desired. The examiner uses a Record Form for recording the responses and scoring. To facilitate the recording of the location of the response, schematic diagrams for the inkblots are included on the Record Form. As each response is given, the examiner outlines a specific area used. Adjacent to the diagram is a blank space for recording the verbatim response or a shortened version of it.

Usually the subject comprehends the nature of the task quickly, and the actual inquiry can be kept to a minimum. A skilled examiner, sensitive to subtle nuances in the examiner-subject interaction, can control the flow of conversation by stimulating a reticent individual and slowing down a verbose person.

In spite of the many interesting variations in test administration that can be attempted, there is much to be said for adhering closely to the standard method of administration. This method has proved highly practical and yields objective, reliable scores on a number of important variables. Currently published normative data and statistical studies of value in the interpretation of the protocols assume close adherence to the standard method of administration.

One exception to the standard method of administration has been successfully developed for young children, whose attention span is short and who therefore may get restless halfway through the testing session. As in the case of the 122 Austin 5-year-olds, the examiner may temporarily interrupt the testing session after 20–30 cards, returning to finish the task after interpolated activity of a relaxing kind. In testing young children, the task can be structured as a playful, game-like activity, heightening the children's interest in attending to it. This technique has been used successfully for some children as young as 3 years of age, although no norms are available below the age of 5. In the course of standardization, 22 quantitative variables were developed to cover nearly all the important scoring categories and dimensions commonly employed with the Rorschach. Table 8.1 gives the names, abbreviations, brief definitions, and scoring weights for these variables. Summary scores for the individual variables are obtained by adding the weights for a given variable

across the 45 inkblots in either Form A or Form B. Three of the variables routinely are “corrected” for the number of rejections in order to provide an estimate of what the total score would have been if the subject had given a response to each of the 45 inkblots. Scoring agreement is uniformly high when trained scorers are compared. Intercorrelations between two independent scorers ranged from .89 to .995, with a median value of .98 in a sample of 40 protocols from a schizophrenic sample. When beginning scorers were compared, average scoring reliability for all variables yielded a median value of .86 in a large sample of normal adolescents. The more difficult variables to score, such as Pathognomic Verbalization and Form Appropriateness, require a greater degree of training. Qualified clinicians and research investigators, however, have had little difficulty achieving satisfactory scoring reliability.

#### RELIABILITY OF HIT SCORES FOR CHILDREN

Both internal-consistency and parallelforms reliability based on repeated testing have been reported in great detail elsewhere (Holtzman et al., 1961b, 1975). Split-half reliabilities, determined by computing the correlations between scores based on oddnumbered and even-numbered blots, are generally high. The four scores with highest internal consistency (generally about .90), regardless of populations sampled (ranging from 5-year-old children to adults with schizophrenia), are Reaction Time, Rejection, Location, and Form Definiteness.

Eight additional scores that have internal consistency reliability generally higher than .80 are Form Appropriateness, Color, Shading, Movement, Pathognomic Verbalization, Human, Animal, and Anatomy.

Odd–even reliability coefficients for the symbolic content score—Anxiety, Hostility, Barrier, and Penetration—are only slightly lower on the average and are more variable. Four scores—Space, Sex, Abstract, and Balance—occur too infrequently for accurate estimates of reliability, leaving only Popular as a score with relatively unsatisfactory internal consistency.

The most pertinent reliability measure for most clinical applications is the stability of an individual’s score across time. Unlike the Rorschach, with which spuriously high results are obtained because of retesting with

the same inkblots, the HIT has truly parallel forms that provide conservative but realistic estimates of reliability of measurement over time. The best measure of such stability is 8. Holtzman Inkblot Technique with Children 201 202

**TABLE 8.1. Names, Abbreviations, Brief Definitions, and Scoring Weights for 22 HIT Variables**  
*Reaction Time (RT)*. The time in seconds from the presentation of the inkblot to the beginning of the primary response.

*Rejection (R)*. Score 1 when subject returns inkblot to examiner without giving a scorable response; otherwise, score 0.

*Location (L)*. Tendency to break down blot into smaller fragments. Score 0 for use of the whole blot, 1 for large area, and 2 for smaller area.

*Space (S)*. Score 1 for true figure-ground reversals; otherwise, score 0.

*Form Definiteness (FD)*. The definiteness of the form of the concept reported, regardless of the goodness of fit to the inkblot. A 5-point scale with 0 for very vague and 4 for highly specific.

*Form Appropriateness (FA)*. The goodness of fit of the form of the percept to the form of the inkblot. Score 0 for poor, 1 for fair, and 2 for good.

*Color (C)*. The apparent primacy of color (including black, gray, or white) as a response determinant. Score 0 for no use of color, 1 for use secondary to form (like Rorschach *FC*), 2 when used as primary determinant but some form present (like *CF*), and 3 when used as primary determinant with no form present (like *C*).

*Shading (Sh)*. The apparent primacy of shading as response determinant (texture, depth, or vista). Score 0 for no use of shading, 1 when used in secondary manner, and 2 when used as primary determinant with little or no form present.

*Movement (M)*. The energy level of movement or potential movement ascribed to the percept, regardless of content. Score 0 for none, 1 for static potential, 2 for casual, 3 for dynamic, and 4 for violent movement.

*Pathognomic Verbalization (V)*. Degree of autistic, bizarre thinking evident in the response as rated on a 5-point scale. Score 0 where no pathology is present. The nine categories of *V* and the range of scoring weights for each is as follows: Fabulation, 1; Fabulized Combination, 2, 3, 4; Queer Response, 1, 2, 3; Incoherence, 4; Autistic Logic, 1, 2, 3, 4; Contamination, 2, 3, 4; Self-Reference, 2, 3, 4; Deterioration Color, 2, 3, 4; Absurd Response, 3.

*Anatomy (At)*. Degree of “gut-like” quality in the content. Score 0 for none; 1 for bones, X-rays, or medical drawings; and 2 for visceral and crude anatomy.

*Sex (Sx)*. Degree of sexual quality in the content. Score 0 for no sexual reference; 1 for socially accepted sexual activity or expressions (buttocks, bust, kissing); and 2 for blatant sexual content (penis, vagina).

*Abstract (Ab)*. Degree of abstract quality in the content. Score 0 for none, 1 for abstract elements along with other elements having form, and 2 for purely abstract content (“bright colors remind me of gaiety”).

*Anxiety (Ax)*. Signs of anxiety in the fantasy content as indicated by emotions and attitudes, expressive behavior, symbolism, or cultural stereotypes of fear. Score 0 for none, 1 for questionable or indirect signs, and 2 for overt or clear-cut evidence.

*Hostility (Hs)*. Signs of hostility in the fantasy content. Scored on a 4-point scale ranging from 0 for none to 3 for direct, violent, interpersonal destruction.

*Barrier (Br)*. Score 1 for reference to any protective covering, membrane, shell, or skin that might be symbolically related to the perception of body image boundaries; otherwise, score 0.

*Penetration (Pn)*. Score 1 for concept that might be symbolic of an individual’s feeling that his or her body exterior is of little protective value and can be easily penetrated; otherwise, score 0.

*Balance (B)*. Score 1 where there is overt concern for the symmetry–asymmetry feature of the inkblot; otherwise, score 0.

*Popular (P)*. Each form contains 25 inkblots in which one or more popular percepts occur. “Popular” in the standardization studies means that a percept had to occur at least 14% of the time among normal subjects. Score 1 for popular core concepts (or their precision alternatives) as listed in the scoring manual; otherwise, score 0.

*Integration (I)*. Score 1 for the organization of the two or more adequately perceived blot elements into a larger whole; otherwise, score 0.

*Human (H)*. Degree of human quality in the content of response. Score 0 for none; 1 for parts of humans, distortions, or cartoons; and 2 for whole human beings or elaborated human faces.

*Animal (A)*. Degree of animal quality in the content. Score 0 for none (including animal objects and microscopic life); 1 for animal parts, bugs, or insects; and 2 for whole animals.

the intraclass correlation obtained in a Latin square design, where half of the subjects are randomly selected to receive Form A before B, while the other half receive B before A. The most extensive studies of this kind on college students yielded stability coefficients ranging from a low of .36 for Popular to a high of .82 for Location, with an interval of 1 week between tests.

Similar studies on other normal subjects with time intervals between testing sessions ranging from 3 months to 5 years provide additional evidence of the stability of HIT scores across time. The most extensive data come from a major cross-cultural study of over 800 children in Mexico and the United States, involving repeated measures with alternate forms 6 years in a row (Holtzman et al., 1975). Beginning during the 1962–1963 school year in Austin, Texas, 133 first graders, 142 fourth graders, and 142 seventh graders were tested with the HIT as part of a large battery of perceptual, cognitive, and personality tests. Annual testing took place on the anniversary date of the initial testing until 6 years of repeated measurement had been completed. Table 8.2

presents the basic design of this study.

A complete replication of the Austin longitudinal project was begun in Mexico City in 1964 under the direction of Rogelio Diaz-Guerrero and his associates. Table 8.3 presents split-half reliability coefficients for 17 HIT scores in the first year of testing for these large samples of children in Mexico and the United States. From these statistics, it is clear that scores on the HIT generally have high reliability for schoolchildren of all ages as well as for adults.

The results for internal consistency of HIT scores among preschool children are equally high, as evidenced by the results obtained for the 122 children age 5 who were tested in Austin nursery schools as part of the standardization sample. Split-half reliability coefficients ranged from .53 for Penetration to .97 for Reaction Time. The median or average reliability for all variables among the 5-year-olds was .86.

Scores on the HIT, the Human Figure Drawing Test, and the Vocabulary and Block Design subtests of the Wechsler Intelligence Scale for Children (WISC) in the cross-cultural longitudinal project provide an unparalleled opportunity to examine the degree of test–retest stability of these measures

over time intervals varying from 1 to 5 years and for schoolchildren of all ages. On a theoretical basis, one would anticipate that the magnitude of such correlations would fall somewhere in the middle ranges—say, from .40 to .80. Correlations much higher than this would indicate rather rigid, unchanging personality characteristics; correlations much lower would reveal instability sufficiently serious to call into question the enduring nature of the measured personality traits over time. Among young children, one would expect lower stability coefficients than among adolescents or adults, because personality and cognitive development proceed more rapidly at the younger ages. And, finally, on a theoretical basis, one would also expect that stability would gradually drop as the interval of time between testing increased from 1 to 5 years. The most stable of all HIT scores is Location. Table 8.4 presents the complete set of test–retest correlations for Location, to illustrate the power of this methodology for

#### 8. Holtzman Inkblot Technique with Children 203

**TABLE 8.2. Overlapping Longitudinal Design for 6 Years of Repeated Testing**

Group	Initial age <sup>a</sup>	School grades covered
I	6.7 years	1 2 3 4 5 6
II	9.7 years	4 5 6 7 8 9
III	12.7 years	7 8 9 10 11 12

*Note.* From Holtzman, Diaz-Guerrero, and Swartz (1975). Copyright 1975 by the Hogg Foundation for Mental Health. Reprinted by permission.

<sup>a</sup>The starting ages of 6 years, 8 months; 9 years, 8 months; and 12 years, 8 months were chosen when a pilot study revealed most children in the public schools of Texas reach these exact ages at some time during the school year. Actual testing took place within 30 days of the age specified.

estimating stability across time. With the exception of the youngest children in the first year, the test–retest correlations for Location were high in both Mexico and the United States, ranging into the .80s for the older children even after several years of testing. It is interesting to note in Table 8.4 that even after an interval of 5 years, the stability of Location was still moderately high, averaging .46 for all six groups combined. It should be pointed out that the availability of parallel forms for the HIT means that an interval of 2 years takes place before the child responds again to the identical form. Two years is a sufficiently long time for memory of the initial responses to fade almost completely. The use of a staggered longitudinal design with overlapping groups, as noted in Table 8.2, also makes possible the isolation of any practice or adaptation effects regardless of the form

used. A detailed analysis of the differences that can be attributed to culture, age, sex, and trial of repeating testing has been reported elsewhere (Holtzman et al., 1975).

Only selected highlights of the findings are presented here.

In a major analysis of variance of Location scores, noticeable adaptation to repeated testing was indeed found over the 6-year period. Mexican children tended to use smaller areas of the inkblot than did American children. Similarly, children of all ages in both cultures tended to use smaller detail areas more often than whole inkblots as the test was repeated. The amount of adaptation was much greater for the Mexican children in the first 2 years of testing. Of all the variables analyzed, only Location showed this adaptation effect over a year of testing, and even then the stability of individual differences through time was unusually high.

Close behind Location in stability were Reaction Time, Form Definiteness, Movement, and Human. These variables compared favorably with scores on Vocabulary and Block Design and with the Harris–Goodenough developmental score on Human Figure Drawing, with respect to stability over a long period of time. Of the 17 HIT scores sufficiently well distributed to permit the use of product-moment correlation coefficients, 6 had generally low stability coefficients ranging from insignificant values into the .40s and .50s, with an occasional value

## 204 II. PROJECTIVE METHODS

**TABLE 8.3. Split-Half Reliability Coefficients for 17 HIT Variables**

	M_e_x_i_c_o				U_n_i_t_e_d__S_t_a_t_e_s			
	6 9	12 6	9 12		6 9	12 6	9 12	
Variable	(147) <sub>a</sub>	(141)	(149)	(133)	(142)	(142)		
Reaction Time	.94	.94	.96	.92	.97	.97		
Rejection	.95	.87	.91	.90	.93	.88		
Location	.94	.95	.95	.95	.97	.95		
Form Definiteness	.91	.88	.89	.90	.80	.86		
Form Appropriateness	.90	.82	.77	.86	.81	.67		
Color	.94	.84	.78	.93	.83	.85		
Shading	.61	.55	.58	.74	.58	.78		
Movement	.84	.83	.85	.86	.87	.88		
Pathognomic Verbalization	.86	.52	.87	.90	.76	.79		
Integration	.57	.69	.77	.58	.85	.82		
Human	.84	.81	.79	.82	.81	.83		
Animal	.92	.83	.71	.80	.73	.72		
Anatomy	.83	.86	.75	.91	.80	.69		
Anxiety	.92	.73	.70	.78	.68	.80		
Hostility	.95	.63	.66	.88	.72	.78		
Barrier	.70	.57	.46	.75	.51	.52		
Penetration	.52	.69	.63	.81	.75	.63		

*Note.* From Holtzman and Swartz (1975). Copyright 1975 by the Hogg Foundation for Mental Health. Reprinted by permission.

*a*Numbers outside parentheses indicate age of each group; numbers within parentheses indicate the number of

children in each group.  
into the .60s and .70s—Rejection, Form Appropriateness,  
Shading, Pathognomic Verbalization,  
Barrier, and Penetration. Four  
variables—Space, Sex, Abstract, and Balance—  
generally occurred too infrequently  
in samples of children to yield data amenable  
to treatment by correlation methods.  
Several generalizations concerning the  
stability of inkblot variables among the children  
and adolescents tested can be drawn  
from these findings:

1. Test–retest stability increases generally  
with an increase in the age of children.  
Older adolescents tend to have the highest  
degree of stability, while children of  
any age show higher test–retest stability  
in the later years of testing than in the  
initial years.
2. Test–retest stability dropped off in a regular  
fashion with increasing size of interval  
between tests.
3. Test–retest stability is generally slightly  
higher for American children than for  
Mexicans, regardless of age group. This  
cross-cultural difference is particularly  
marked for Integration, Hostility, and  
Barrier.
4. Most of the HIT scores show a sufficiently  
high degree of stability across  
time, ranging from .40 to .80, to justify  
their use as predictors of later behavior.  
(A successful example of such a study  
over a 9-year period is reported later in  
this chapter.) At the same time, the  
test–retest correlations are not so high as  
to suggest any kind of fixed traits that remain  
relatively invariant as children  
grow older.

#### **VALIDITY OF THE HIT FOR CHILDREN AND ADOLESCENTS**

Several hundred studies have been published  
bearing on the relationships between  
scores on HIT variables and independent  
measures of personality. Although most  
have been carried out with adults, many  
have dealt specifically with children or adolescents.  
Extensive reviews have appeared  
recently elsewhere (Holtzman, 1981, 1985;  
Holtzman & Swartz, 1983). A comprehensive  
annotated bibliography containing all  
known references to the HIT through 1999  
(Swartz et al., 1999) contains abstracts of  
these articles. Only representative highlights  
of these many findings as they pertain to use  
of the HIT with children and adolescents

are provided here. Because factor analyses of intercorrelations among the 22 HIT variables have indicated that, with few excep-

8. Holtzman Inkblot Technique with Children 205

**TABLE 8.4. HIT Location Test–Retest Correlations**

	M_e_x_i_c_o			U_n_i_t_e_d_S_t_a_t_e_s			Years
correlated	I	II	III	I	II	III	
1 & 2	.27	.57	.66	.28	.72	.76	
1 & 3	.20	.49	.60	.27	.62	.70	
1 & 4	.25	.49	.58	.26	.59	.69	
1 & 5	.24	.46	.56	.26	.51	.67	
1 & 6	.50	.26	.51	.33	.56	.59	
2 & 3	.58	.70	.71	.50	.77	.84	
2 & 4	.49	.73	.72	.54	.68	.81	
2 & 5	.42	.64	.67	.46	.58	.76	
2 & 6	.23	.57	.52	.49	.62	.75	
3 & 4	.64	.75	.76	.68	.78	.82	
3 & 5	.56	.70	.74	.70	.80	.80	
3 & 6	.44	.63	.55	.64	.77	.75	
4 & 5	.60	.77	.85	.71	.76	.86	
4 & 6	.62	.73	.68	.68	.80	.86	
5 & 6	.63	.76	.74	.79	.86	.85	

*Note.* Table covers 6 years of repeated testing; Groups I, II, and III started in Year 1 at ages 6.7, 9.7, and 12.7, respectively. From Holtzman, Diaz-Guerrero, and Swartz (1975). Copyright 1975 by the Hogg Foundation for Mental Health. Reprinted by permission.

tions, these scores tend to cluster into meaningful factors, the findings with respect to validity are arranged according to these more general dimensions.

**Factor 1: Perceptual Maturity and Integrated Ideational Activity**

Factor 1 is defined by four variables: Movement, Integration, Human, and Barrier. High scores on these four variables taken together are indicative of well-organized ideational activity, good imaginative capacity, and well-differentiated ego boundaries. All four variables increase with age among children and are significantly higher among college graduates than average adults, indicating a strong component of cognitive ability and creativity (Swartz, Reinehr, & Holtzman, 1983). Studies have shown repeatedly that these variables are indicative of creative potential.

They show significant relationships with reading comprehension among children, even after general intelligence has been held constant (Laird, Laosa, & Swartz, 1973). The energy level of movement ascribed to the percept, regardless of content, has some other interesting correlates. A high score on Movement is associated with perceived empathy in counselors, whereas low Movement is associated with the reverse (Mueller & Abeles, 1964). The degree of eye contact and smiling among psychiatric patients when interviewed is also related to high Movement scores (Lefcourt, Telegdi, Willows,

& Buckspan, 1972). Movement is correlated with the discharge or inhibition of cognitive energy, according to Covan (1976). Increased perception of movement in inkblots follows experimental inhibition of cognitive responses, while discharge of cognitive processes in a series of free-association tasks leads to a sharp decrease in reported perception of movement in inkblots. Studies of dream deprivation, whether induced by drugs (Lerner, 1966) or prevented by interrupting rapid eye movements (Feldstein, 1973), results in higher Movement scores. These findings support Rorschach's views on the fundamental similarity between movement and dreams because of the centrality of kinesthetic experience in both; the results are also consistent with Heinz Werner's sensory tonic theory of perception. Movement has a particularly strong cognitive component among young children. In the first year of testing for 6-year-olds in the longitudinal study, the complete battery of tests for the WISC was given. Factor analyses of intercorrelations among the various cognitive tests were done with Movement included as an extra variable. Among the U.S. children, Movement was clearly a major part of the first factor defined by the Verbal subtests from the WISC. Movement did not show up heavily loaded on a similar factor for the Mexicans until the 9-year-olds were analyzed. Similar results were found for the 12-year-old Mexican children. No complete analyses could be performed for the U.S. 9- and 12-year-olds, because they had not been given the complete WISC test battery. Movement deals with that component of verbal ability characterized by a lively, active imagination and the ability to project outward from one's fantasies. In this sense, it deals particularly with the expressive, imaginative aspects of verbal ability, rather than with factual information, word meanings, and analytic problem solving. Human content also has some special meaning worthy of note. As one would expect from projective theory, a high score on Human suggests high social interest, whereas lack of any Human content indicates the opposite (Fernald & Linden, 1966). One of the most interesting of the symbolic content scores is Barrier, developed by Fisher and Cleveland (1958). The score is given for references to any protective covering, membrane, shell, or skin that might be

symbolically related to the perception of body image boundaries. High Barrier is indicative of strong ego identity, whereas low Barrier suggests diffusion. High Barrier is related to being influential and independent in group processes (Cleveland & Morton, 1962), adjusting well to physical disablement (Fisher, 1963), being able to tolerate pain (Nichols & Tursky, 1967), and having a positive evaluation of one's own body (Conquest, 1963). These findings are consistent with others showing low Barrier related to juvenile delinquency (Megargee, 1965).

### **Factor 3: Psychopathology of Thought**

Factor 3 consists of Pathognomic Verbalization, Anxiety, and Hostility. Dealing with unbridled fantasies, affective expressivity,

#### **206 II. PROJECTIVE METHODS**

and loose imagination, these three variables, frequently also associated with Movement, constitute an important cluster of scores indicative of psychopathology. Among children, moderately high scores on these three variables may be a good sign rather than a bad one. In most factor analyses of inkblot variables among children, this factor often proves to be highly correlated with Factor 1, indicating creativity and imaginative capacity rather than psychopathology. Even among children, however, high scores on these variables are indicative of future emotional disturbance, as demonstrated by a long-term follow-up study of the 6-yearolds tested in Austin as part of the cross-cultural study.

Nine years after the initial testing of the first graders in Austin, 46 of them (23 girls, 23 boys) were located within the Austin schools and rated on personal adjustment by school personnel (Currie, Holtzman, & Swartz, 1974). These 46 subjects constituted more than half the group ( $n = 89$ ) that completed the 6-year period of testing and were all from families that had continued to live in Austin, Texas, over a 10-year span, thus representing a particularly stable sample in relation to current population mobility. Seven of the children were judged to have serious problems of adjustment; 7 were judged to have noticeable areas of maladjustment; 18 were judged to be generally well adjusted but with some problems in relating to others; and 14 were judged to be well adjusted. The mean scores on Pathognomic Verbalization, Anxiety, and Hostility from HIT protocols 6 years earlier

for the seven most disturbed children were two standard deviations higher than the means for children judged to be well adjusted. Low Form Appropriateness was also associated with emotional disturbance. Interestingly enough, from the many tests in the original battery, the only other one that correlated significantly with later emotional disturbance in these children was the Koppitz Scale of emotional indicators in Human Figure Drawings (see Table 8.5). Pathognomic Verbalization is the best single indicator of psychopathology. Among adults, the bizarre perception and autistic logic underlying high scores on this variable are characteristic of either schizophrenia or extreme artistic license in responding to inkblots. Highly creative artists do tend to get much higher scores than do average individuals on this variable (Holtzman,

#### 8. Holtzman Inkblot Technique with Children 207

**TABLE 8.5. Mean Test Scores from the WISC, the Human Figure Drawing, and the HIT by Four Levels of Adjustment and Correlations between Test Variables and the Adjustment Index**

Means by adjustment level	A(3)	B(2)	C(1)	D(0)	Total	Total	
Test variable ( <i>n</i> = 14) ( <i>n</i> = 18) ( <i>n</i> = 7) ( <i>n</i> = 7) ( <i>n</i> = 46) <i>SD r</i>							
WISC IQ	111.8	112.6	116.3	99.5	110.8	10.9	.21
HFD							
Koppitz indicators	0.6	0.9	1.4	2.0	1.0	1.1	-.44**
Goodenough-Harris	18.8	16.8	18.4	14.7	17.4	5.1	.21
HIT							
Form Appropriateness (FA)	35.0	35.3	35.7	25.6	33.8	9.2	.28*
Movement (M)	14.9	21.4	17.5	23.3	19.2	15.5	-.15
Pathog. Verbalization (V)	6.8	6.8	6.3	18.4	8.5	8.3	-.38**
Integration (I)	1.0	1.3	1.6	0.6	1.2	1.1	.10
Human (H)	12.8	16.9	20.7	16.7	16.2	9.7	-.19
Anatomy (At)	4.1	4.8	6.4	7.7	5.3	6.3	-.19
Anxiety (Ax)	5.6	6.8	4.9	16.0	7.5	7.4	-.37**
Hostility (Hs)	6.2	6.7	5.7	19.3	8.3	9.1	-.39**
Penetration (Pn)	2.8	3.6	3.3	3.0	3.2	2.6	-.03

*Note.* From Currie, Holtzman, and Swartz (1974). Copyright 1974 by Pergamon Press. Reprinted by permission.

\*Correlation significant beyond .05 level.

\*\*Correlation significant beyond .01 level.

Swartz, & Thorpe, 1971), but the quality of the response is noticeably different. Table 8.6 gives names, abbreviations, definitions, and scoring weights for the various categories of Pathognomic Verbalization. Normal individuals tend to give Fabulations with notable affectivity, mildly Fabulized Combinations of otherwise acceptable percepts, or even occasional Queer Responses that are often described in a playful manner. Schizophrenics, on the other hand, manifest a loss of distance between themselves and the inkblots, often giving severely Fabulized Combinations, Contaminations, Queer Responses, or special kinds

of Autistic Logic that show faulty, fantastic reasoning as a justification for the response. Embellishing a response with highly personal meaning by Self-Reference is particularly characteristic of psychotic thinking when manifested repeatedly (Swartz, 1969). A predominance of Absurd Responses is char-

208 II. PROJECTIVE METHODS

**TABLE 8.6. Names, Abbreviations, Brief Definitions, and Scoring Weights for Nine Categories of Pathognomic Verbalization (V)**

*Fabulation (FB)*. A response in which there is associative elaboration having notable affective components (16A “A kind monster . . . eyes, and he looks real sweet”). FB is scored 1 when present.

*Fabulized Combination (FC)*. An impossible, fantastic combination of otherwise fairly acceptable parts, based largely on a spatial rather than a logical relationship (14B “Some kind of get-together. Two caterpillars talking to each other in a sociable mood, don’t seem to mind having the sheep around”). FC is scored 2, 3, or 4, depending on the judged severity of pathology evident.

*Queer Response (QR)*. A response in which the subject employs peculiar or eccentric language and syntax in verbalizing the response (14A “The male part of the back ... like the muscles in the biceps formed in the V-shaped web”). QR is scored 1, 2, or 3, depending on judged severity.

*Incoherence (IC)*. A response in which there is a complete breakdown of rational control (14A “A dental of hell . . . under misunderstanding”). IC is scored 4 when present.

*Autistic Logic (AL)*. The presence of faulty, fantastic reasoning given by the subject as justification for the response (43A “A carbon copy of a person . . . because he’s lying down on carbon paper”). AL is scored 1, 2, 3, or 4, depending on judged severity.

*Contamination (CT)*. A response in which two conflicting interpretations are fused into one, or when the same area simultaneously stands for two or more interdependent but logically separate concepts (11B “That looks like a stone stain . . . looks like a heart . . . well, a stained heart”). CT is scored 2, 3, or 4, depending on judged severity.

*Self-Reference (SR)*. A response in which the subject draws himself or herself into the percept, giving the response a personal meaning (29B “A person’s face. Looks like my face . . . maybe I lost it”). SR is scored 2, 3, or 4, depending on judged severity.

*Deterioration Color (DC)*. Loose, fantastic color associations having bizarre content are given with an air of reality (4B “The yellow is a virulent disease, the yellow plague . . . the kiss of death”). DC is scored 2, 3, or 4, depending on judged severity.

*Absurd Response (AB)*. A response is categorized as absurd when a subject assigns a form-definite concept to an area of an inkblot in which by no stretch of the imagination can the form be conceived of as appropriate, and the response is not an abstract one (36A “The Empire State Building . . . whole card”). AB is scored 3 when present.

Scale Value  
 0 1 2 3 4  
 FB FC  
 No QR IC  
 pathology AL  
 present CT  
 SR  
 DC  
 AB

*Note.* The schematic diagram above shows the range of scoring weights for each of the nine V categories and the relationship of the FB and QR categories to the FC and IC categories, respectively.

acteristic of mentally retarded individuals, while a predominance of Deteriorated Color associations is indicative of severe disintegration.

Among young children, moderately high scores on Pathognomic Verbalization may simply indicate immature thought processes coupled with uncontrolled fantasies and loose imagination, rather than serious psychopathology.

Although few of the cases in either the Mexican or the U.S. sample received scores on this variable so high as to

indicate serious psychopathology, the presence of some Pathognomic Verbalization among young children may indeed be taken as a good sign, provided that the qualitative nature of the disordered thinking reveals primarily Fabulized Combinations and invalid Integrations rather than bizarre perceptions. In an analysis of longitudinal data on the three groups of U.S. children from the Mexico–United States cross-cultural investigation discussed earlier, a highly significant curvilinear developmental trend was apparent for Pathognomic Verbalization (Swartz, 1969). Across the 11 years from 6 to 17, the lowest mean score on this variable occurred in 12-year-old children, with rising means both up and down the developmental order. When groups of children ( $n = 180$ )—6, 9, and 12 years of age, respectively—were matched for sex and total Pathognomic Verbalization score, using HIT data from initial testing sessions only, it was found that with increasing age there was a significant increase in the number of children giving Fabulation responses and significant decreases in the numbers of children giving Autistic Logic, Contamination, and Absurd Responses. The numbers of children giving Fabulized Combination, Queer Responses, and Deteriorated Color responses remained quite steady across the 6-year age span (with about half in all three age groups giving Fabulized Combination responses, more than half giving Queer Responses, and only about 8% giving Deteriorated Color responses). Almost none produced verbalizations falling into the Incoherence or Self-Reference categories.

The importance of Pathognomic Verbalization in psychodiagnosis can best be illustrated by an individual case drawn from our files of schoolchildren in the Austin longitudinal study. While being tested with the HIT in the third year of repeated study, a young, apparently normal teenage boy shifted abruptly from normal responses one-third of the way through the test, thereafter giving responses heavily loaded with Pathognomic Verbalization. It is important to note that his responses to the first 12 inkblots and in the previous testing sessions over the past 2 years were generally normal, although under a high degree of self-control. Even in the remainder of this testing session, he continued to maintain an outward appearance of control, being polite, cooperative,

and attentive. Although we were quite concerned about the sudden deterioration in the quality of his responses, nothing was said to his parents or authorities because of the research nature of the data collection. Some months later, we learned to our dismay that the boy had killed his father and was hospitalized for treatment as a patient with schizophrenia. The incipient psychosis was not apparent in his general behavior or on other tests, although it was clearly revealed by his Pathognomic Verbalization score on the HIT.

More recently, Leichsenring (1991) successfully used HIT deviant verbalizations to classify acute schizophrenics, chronic schizophrenics, and borderline patients and, in another comparison, to differentiate between neurotic and normal individuals.

Signs of anxiety or hostility in the fantasy content form the basis for the Anxiety or Hostility score. Moderate-level scores on both of these symbolic content scales are normal, particularly in young children, but very high scores should be interpreted as having likely clinical significance. Zero or low positive correlations can be expected between these two variables and anxiety or hostility scales in self-report inventories. The most important evidence of their validity comes from experimental studies. Subjects who rapidly acquire the conditioned eyelid response have higher Anxiety scores than do those who do not condition easily (Herron, 1965). Individuals with high Anxiety are less tolerant of pain (Nichols & Tursky, 1967). Individuals who show a marked increase in Hostility score after a frustrating situation are those who also show a predisposition to hostility as measured by Factor I of the Buss–Durkee Inventory (von Rosenstiel, 1973). Both Anxiety and Hostility scores are directly related to observed interpersonal distance characteristic of an individual in an experimental setting; the higher the inkblot scores, the greater the distance (Greenberg, Aronow, & Rauchway, 1977). These findings all are consistent with the theoretical conception of these symbolic content scores.

In a series of studies, Kamen (1969, 1970, 1971) found several HIT variables related to State anxiety (but not Trait anxiety) as measured by the State–Trait Anxiety Inventory (STAI). A later study by Iacino

and Cook (1974), however, found HIT Anxiety to be correlated positively with the STAI Trait scale. Auerbach and Edinger (1977) found that only HIT Barrier correlated significantly (negatively) with STAI Trait anxiety; no HIT variables were found to be related to STAI State anxiety. Mittenberg and Petersen (1984), using vasomotor biofeedback, supported the validity of the HIT as a measure of anxiety but failed to characterize this anxiety as either state or trait in nature.

In a study of the aggression-reducing effects of *d*-amphetamine, Amery, Minichiello, and Brown (1984) studied 10 8- to 11-year-old boys diagnosed with attention deficit disorder with hyperactivity in a double-blind placebo-counterbalanced trial of dextroamphetamine (DAM). It was predicted that aggressive behavior and aggressive/impulsive attitudes and fantasies would be lower in the DAM phase than in the placebo phase. Findings in the predicted direction were significant on three of the five measures used, including HIT Hostility. Moreover, in a study of hostility employing 130 Caucasian, 109 African American, and 34 Hispanic forensic patients, Reinehr, Swartz, and Dudley (1984) found that ethnic group differences may exist on both objective and HIT measures of hostility.

Although the meaning of Anxiety and Hostility scores on the HIT undoubtedly is complex, it seems clear that very high scores, even among children, have sufficient validity to justify clinical interpretation for individuals. While cautioning that the HIT and other projective techniques are not direct behavioral measurements but instead measure underlying processes based on an individual's perception, Fehr (1983) sees the HIT as superior to the Rorschach in measuring both anxiety and hostility.

#### **Factor 2: Perceptual Sensitivity**

Factor 2 consists of Color, Shading, and Form Definiteness (reversed). The clustering together of these three variables is inevitable. As with scoring systems for the Rorschach, the greater the predominance of color or shading over form in a response, the higher the score. Among younger children, significant negative loadings on this factor also appear for Animal, suggesting that many children tend to use color and shading as a determinant only when they cannot find a familiar animal form. The

positive pole on this factor indicates overreactivity to the stimulus determinants; the negative pole shows primary concern for form alone as a response determinant. Among normal subjects, a high Color score has been found to be related to impulsivity (Holtzman, 1950) and to increased expression of affect (Mayfield, 1968). In her clinical use of the HIT, Hill (1972) recommends paying attention to the quality of the Color responses, particularly those given to inkblots having a high stimulus strength for Color, in making interpretations about the lability of affect.

There is little experimental evidence bearing on the validity of Shading or Form Definiteness, the other two variables that measure degree of perceptual sensitivity. Nor is there much information from correlational studies with other personality measures that would indicate the independent meaning of these variables for assessment purposes. To be sure, there is a consensus among Rorschach clinicians concerning the use of these scores for personality assessment, but the scientific evidence is too tenuous at this time to justify any confident interpretations, particularly among children.

#### **Other Factors**

The remaining three factors are less important and vary somewhat in their patterning from one population to another. Location and Form Appropriateness generally appear as defining variables for Factor 4. A high score on Location results when an individual uses smaller areas of the inkblot while ignoring the rest. This perceptual style makes it easier to find percepts that have good form. The combination of low score on Location with high score on Form Appropriateness is less common and indicates a high level of perceptual maturity and organization, particularly when accompanied by high scores on Integration.

Reaction Time and Rejection tend to be associated in a single factor, because both measure the extent of inhibition or outright perceptual inability. Both variables must be taken into account with other inkblot scores rather than interpreted alone.

Three scores—Sex, Anatomy, and Penetration—deal with bodily preoccupation and are occasionally clustered together in one factor for this reason. Blatant Sex responses are relatively rare but are significant

when they do appear, especially among children. Very high Anatomy scores are also quite unusual and significant. High Anatomy scores have been found to be closely associated with a high degree of somatic preoccupation among hospitalized patients, confirming the theoretical interpretation of Anatomy (Endicott & Jortner, 1967). Penetration frequently loads also on Factor 3, Psychopathology of Thought, suggesting that high scores on Penetration should be generally interpreted as pathological.

#### **Use of Scores for Differential Diagnosis**

Closely related to the clinical validity of individual variables within the HIT is the use of patterns of scores for differential diagnosis. The original standardization data on 15 different populations are presented in percentile norms for eight major reference groups ranging from 5-year-olds to superior adults and including psychiatric patients as well as mentally retarded individuals. Chronic schizophrenics differ from normal reference groups on almost all of the standard HIT variables.

Conners (1965) reported a number of highly significant differences between emotionally disturbed children seen in an outpatient clinic and normal controls of the same age and background. In Conners's study, disturbed children got higher scores on Rejection and Anatomy and lower scores on all other variables except Pathognomic Verbalization, Sex, Abstract, Hostility, Penetration, and Balance. Using HIT factor scores, Conners found that neurotic children appeared to be more differentiated in response and more inhibited than did hyperkinetic children. At the same time, it should be noted that neurotic children received higher scores on Form Appropriateness and Location than did children with conduct disorders.

Hill's (1972) handbook on clinical application of the HIT has provided detailed suggestions on how best to interpret HIT scores, as well as the qualitative aspects of content with respect to cognitive functioning, affective functioning, and self-identity.

In 1976, Aronow and Reznikoff compared Rorschach and HIT content scores and concluded that "the HIT is clearly the technique of choice for most research purposes" (p. 315). More recently, Pokhriyal and Ahmad (1988)—in a study of the HIT response patterns of acute schizophrenics, endogenous depressives, and normal subjects—

found that HIT scores differentiated significantly between the two clinical groups and between the normal subjects and both of the psychotic groups.

#### **CROSS-CULTURAL USE OF THE HIT**

Use of the HIT in research studies of differences in personality development among children of different cultures is particularly appropriate because of the more or less universal nature of the technique. The method has been translated into a number of languages and is relatively culture-free. The technique has been used successfully for both adults and children in cultures as widely varied as primitive groups in New Guinea, Aleutian Eskimos, peasant children in Latin America, and children from modern, industrialized societies. Our own work has concentrated on factors related to personality and cognitive development among children in Mexico and the United States (Holtzman et al., (1975). Of all the measures used in the cross-cultural longitudinal study, the HIT yielded the most striking differences between Mexican and U.S. children. One finding is of particular interest, because it sheds considerable light on the possible use of the HIT for measuring an important coping style.

#### **8. Holtzman Inkblot Technique with Children 211**

The most significant differences between Mexican and U.S. children, regardless of age, sex, or socioeconomic status, were found for seven HIT scores—Reaction Time, Pathognomic Verbalization, Location, Movement, Integration, Anxiety, and Hostility. Mexicans had a slower response time; showed less pathology, anxiety, hostility, or movement in their fantasy expressions; tended to use more small details within the blots for their responses; and showed lower ability to integrate the parts into a meaningful whole than did the American children. These differences tended to narrow with increasing age.

Most of the differences between the Mexican and U.S. children on the HIT can be understood better in terms of coping style than of any other concept. The U.S. children produced faster Reaction Time, used larger portions of the inkblots in giving responses, gave more definite form to responses, and still were able to integrate more parts of the inkblots while doing so. In addition, they incorporated other stimulus properties of the inkblots, such as Color and Shading,

into their responses more often than did the Mexican children, and they elaborated their responses by ascribing more movement to their percepts. In attempting to deal with all aspects of the inkblots in such an active manner, however, they failed more often than the Mexican children; that is, the Mexican children gave responses with better form and less often produced responses that showed deviant thinking and anxious and hostile content. In general, U.S. children tried to deal with the testing situation in much more active manner than did the Mexican children, even when unable to do so successfully.

The work of Tamm (Haroz, 1967) in the American School in Mexico City allows for a deeper insight into the meaning of these cross-cultural findings with U.S. and Mexican schoolchildren. Tamm designed a study involving bilingual Mexican and U.S. children attending the same school. Thirty children in the first, fourth, and seventh grades were tested at 6 years, 8 months; 9 years, 8 months; and 12 years, 8 months of age, respectively, to provide precise parallels to the design employed in the larger cross-cultural study between Mexico and the United States. One-half of the children were native Mexicans for whom Spanish was the primary language. These children generally came from upper-class Mexican families in which there was a strong desire on the part of the parents for their children to obtain a U.S.-style education. The remainder of the children were Americans whose fathers were businessmen or government representatives in Mexico City. The U.S. families wanted their children to develop bilingual/bicultural skills and attitudes. The curriculum in the American School was taught half in English and half in Spanish.

Tamm administered the HIT and all the subtests of the WISC to each of the 90 schoolchildren 2 years in a row. The children's test performance was analyzed in a three-way analysis-of-variance design by culture, age group, and year of testing. Of the WISC subtests, only Digit Span proved significant across cultures, the Mexican children doing slightly better than the U.S. children. The usual developmental differences were clearly apparent in both groups. On the HIT, however, marked differences were found between the Mexican and U.S. children, differences that in every respect

were essentially the same as the major differences found for HIT scores in the larger cross-cultural study. Mexican children used much more small detail and gave less Color, less Movement, less Pathognomic Verbalization, less Human content, less Anxiety, and less Hostility than did the U.S. children. The lack of any notable differences between the Mexican and U.S. children on the intelligence tests in Tamm's study, regardless of the length of time the children had spent in the American School or the children's ages, provides convincing evidence that the combination of home environment and schooling is important in the development of these mental abilities. At the same time, the dramatic differences in personality and perceptual style reflected in the HIT—differences identical to those obtained when U.S. children in Austin were compared with Mexican children in Mexico City—indicate that fundamental aspects of the U.S. and Mexican personality or "national character" remain intact, in spite of common schooling and other forces within the immediate environment of the children that would tend to produce convergence of the two cultures. The sociocultural premises underlying the U.S. and Mexican societies, and the basically different styles of coping with the challenges of life in the two cultures, provide a key to the interpretation of these results. U.S. children tend to be more actively independent and to struggle for a mastery of problems and challenges in their environment, whereas Mexican children are more passively obedient and adapt to stresses in the environment instead of trying to change them.

## CONCLUSION

The value of the HIT for clinical and research use with children depends on the reliability and validity of HIT scores for use in personality assessment and psychodiagnosis, as well as on the efficiency and ease of administration, scoring, and interpretation. Although the HIT is still a relatively young technique, the evidence to date seems to indicate that it has answered most if not all of the criticisms of the Rorschach. The availability of parallel forms and standardized variables, without sacrificing the qualitatively rich projective content of the Rorschach, provides clinicians and researchers alike with a powerful tool for the

assessment of personality in children as well as adults, and explains the growing acceptance of the HIT throughout the world. Experts in assessment have been quick to point out its advantages over the Rorschach (Anastasi, 1982; Fehr, 1983; Kleinmuntz, 1982; Peterson, 1978). The HIT is more demanding of the clinician than the Rorschach; however, the time and effort involved in administering, scoring, and interpreting the HIT need not be any greater than for the Rorschach once the technique has been mastered. Those who have learned the method well have been enthusiastic about its value for both clinical and research purposes. Some find the HIT difficult because it has 45 inkblots rather than 10, or because it has only one response per card rather than as many as the child wishes to give. Yet these are the very features that produce superior psychometric qualities, rendering the HIT more suitable for rigorous scientific validity as well as for implementation by modern computer technology. As with any major test for the assessment of personality, the final verdict on the HIT will be reached only after many years of experimental and clinical work with children and adults.

#### NOTE

1. Materials for the HIT can be obtained from the Psychological Corporation, 19500 Bulverde Road, San Antonio, TX 78259-3701.

The *Guide for Administration and Scoring* is an offprint of the sections on administration and scoring from Holtzman and colleagues (1961b). Sets of 35 mm slides are used with the group method of administration. A research guide and annotated bibliography of the HIT (Swartz et al., 1999) and other monographs in addition to Holtzman and colleagues are also available.

#### REFERENCES

- Amery, B., Minichiello, M. D., & Brown, G. L. (1984). Aggression in hyperactive boys: Response to d-amphetamine. *Journal of the American Academy of Child Psychiatry, 23*, 291–294.
- Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
- Aronow, E., & Reznikoff, M. (1976). *Rorschach content interpretation*. New York: Grune & Stratton.
- Auerbach, S. M., & Edinger, J. D. (1977). The effects of surgery-induced stress on anxiety as measured by the Holtzman Inkblot Technique. *Journal of Personality Assessment, 41*, 19–24.
- Bock, D. R., Haggard, E. A., Holtzman, W. H., Beck, A. G., & Beck, S. J. (1963). *A comprehensive psychometric study of the Rorschach and Holtzman inkblot techniques*. Chapel Hill: University of North Carolina, Psychometric Laboratory.
- Buros, O. K. (Ed.). (1949). *The third mental measurements*

- yearbook. New Brunswick, NJ: Rutgers University Press.
- Cleveland, S. E., & Morton, R. B. (1962). Group behavior and body image: A follow-up study. *Human Relations, 15*, 77–85.
- Conners, C. K. (1965). Effects of brief psychotherapy, drugs, and type of disturbance on Holtzman Inkblot scores in children. *Proceedings of the 73rd Annual Convention of the American Psychological Association, 1*, 201–202.
- Conquest, R. A. (1963). *An investigation of body image variables in patients with the diagnosis of schizophrenic reaction*. Unpublished doctoral dissertation, Case Western Reserve University.
- Covan, F. L. (1976). *The perception of movement in inkblots following cognitive inhibition*. Unpublished doctoral dissertation, Yeshiva University.
- Currie, S. F., Holtzman, W. H., & Swartz, J. D. (1974). Early indicators of personality traits viewed retrospectively. *Journal of School Psychology, 12*, 51–59.
- Endicott, N. A., & Jortner, S. (1967). Correlates of somatic concern derived from psychological tests. *Journal of Nervous and Mental Disease, 144*, 133–138.
- Fehr, L. A. (1983). *Introduction to personality*. New York: Macmillan.
- Feldstein, S. (1973). *REM deprivation: The effects of inkblot perception and fantasy processes*. Unpublished doctoral dissertation, City University of New York.
- Fernald, P. S., & Linden, J. D. (1966). The human content response in the Holtzman Inkblot Technique. *Journal of Projective Techniques and Personality Assessment, 30*, 441–446.
- Fisher, S. (1963). A further appraisal of the body boundary concept. *Journal of Consulting Psychology, 27*, 62–74.
- Fisher, S., & Cleveland, S. E. (1958). *Body image and personality*. Princeton, NJ: Van Nostrand.
- Gorham, D. R. (1967). Computer use in psychological testing. In G. Gillespie (Ed.), *Memorias del XIth Congreso, Interamericano de Psicología* (Vol. 9, pp. 1–7). Mexico City: Universidad Nacional Autónoma de México.
- Greenberg, E., Aronow, E., & Rauchway, A. (1977). Inkblot content and interpersonal distance. *Journal of Clinical Psychology, 33*, 882–887.
- Haroz, M. M. (1967). *El Holtzman Inkblot Test, el Wechsler Intelligence Scale para Children y otros tests en el estudio psicológico transcultural de niños de habla Española e Inglesa residentes en México*. Unpublished doctoral dissertation, Universidad Nacional Autónoma de México, Mexico City.
- Herron, E. W. (1963). Psychometric characteristics of a thirty-item version of the group method of the Holtzman Inkblot Technique. *Journal of Clinical Psychology, 19*, 450–453.
- Herron, E. W. (1965). Personality factors associated with the acquisition of the conditioned eyelid response. *Journal of Personality and Social Psychology, 2*, 775–777.
- Hill, E. F. (1972). *The Holtzman Inkblot Technique: A handbook for clinical application*. San Francisco: Jossey-Bass.
- Holtzman, W. H. (1950). *The Rorschach test in the assessment of the normal superior adult*. Unpublished doctoral dissertation, Stanford University.
- Holtzman, W. H. (1975). New developments in the HIT. In P. McReynolds (Ed.), *Advances in psychological*

- assessment (Vol. 3, pp. 243–274). San Francisco: Jossey-Bass.
- Holtzman, W. H. (1981). Holtzman Inkblot Technique (HIT). In A. I. Rabin (Ed.), *Assessment with projective techniques: A concise introduction* (pp. 47–83). New York: Springer.
- Holtzman, W. H. (1985). Clinical applications in personality assessment and psychodiagnosis. In J. J. Sanchez-Sosa (Ed.), *Health and clinical psychology* (pp. 3–19). Amsterdam: North Holland.
- Holtzman, W. H. (1986). The Holtzman Inkblot Technique with children and adolescents. In A. I. Rabin (Ed.), *Projective techniques for adolescents and children* (pp. 168–192). New York: Springer.
- Holtzman, W. H. (1988). Beyond the Rorschach. *Journal of Personality Assessment*, 52, 578–609.
- Holtzman, W. H., Diaz-Guerrero, R., & Swartz, J. D. (1975). *Personality development in two cultures: A cross-cultural longitudinal study of school children in Mexico and the United States*. Austin: University of Texas Press.
- Holtzman, W. H., Moseley, E. C., Reinehr, R. C., & Abbott, E. (1963). Comparison of the group method and the standard individual version of the Holtzman Inkblot Technique. *Journal of Clinical Psychology*, 19, 441–449.
- Holtzman, W. H., & Swartz, J. D. (1983). The Holtzman Inkblot Technique: A review of 25 years of research. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 4, 241–259.
- Holtzman, W. H., Swartz, J. D., & Thorpe, J. S. (1971). Artists, architects, and engineers: Three contrasting modes of visual experience and their psychological correlates. *Journal of Personality*, 39, 432–449.
- Holtzman, W. H., Thorpe, J. S., Swartz, J. D., & Herron, E. W. (1961a). *Administration and scoring guide*. New York: Psychological Corporation.
- Holtzman, W. H., Thorpe, J. S., Swartz, J. D., & Herron, E. W. (1961b). *Inkblot perception and personality: Holtzman Inkblot Technique*. Austin: University of Texas Press.
- Iacino, L. W., & Cook, P. E. (1974). Threat of shock, state anxiety, and the HIT. *Journal of Personality Assessment*, 38, 450–458.
- Kamen, G. B. (1969). Effects of a stress-producing film on the test performance of adults. *Journal of Projective Techniques and Personality Assessment*, 33, 281–285.
- Kamen, G. B. (1970). *The effects of a stress-producing film on the test performance of adults*. Unpublished doctoral dissertation, University of Missouri.
- Kamen, G. B. (1971). A second look at the effects of stress-producing film on adult test performance. *Journal of Clinical Psychology*, 27, 465–467.
- Kleinmuntz, B. (1982). *Personality and psychological assessment*. New York: St. Martin's Press.
- Laird, D. R., Laosa, L. M., & Swartz, J. D. (1973). Inkblot perception and reading achievement in children: A developmental analysis. *British Journal of Projective Psychology and Personality Study*, 18, 25–31.
- Lefcourt, H. M., Telegdi, M. S., Willows, D., & Buckspan, B. (1972). Eye contact and the human movement response. *Journal of Social Psychology*, 88, 303–304.
- Leichsenring, F. (1991). Discriminating schizophrenics from borderline patients: Study with the Holtzman Inkblot Technique. *Psychopathology*, 24, 225–231.
- Lerner, B. (1966). Rorschach movement and dreams:

- A validation study using drug-induced dream deprivation. *Journal of Abnormal Psychology*, 71, 75–86.
- Mayfield, D. G. (1968). Holtzman Inkblot Technique in acute experimental alcohol intoxication. *Journal of Projective Techniques and Personality Assessment*, 32, 491–494.
- Megargee, E. I. (1965). The relation between barrier scores and aggressive behavior. *Journal of Abnormal Psychology*, 70, 307–311.
- Mittenberg, W., & Petersen, J. D. (1984). Validation of the Holtzman anxiety scale by vasomotor biofeedback. *Journal of Personality Assessment*, 48, 360–364.
- Morgan, A. B. (1968). Some age norms obtained for the Holtzman Inkblot Technique administered in a clinical setting. *Journal of Projective Techniques and Personality Assessment*, 32, 165–172.
- Mueller, W. J., & Abeles, N. (1964). The components of empathy and their relationship to the projection of human movement responses. *Journal of Projective Techniques and Personality Assessment*, 28, 322–330.
- Mullen, J. M., Reinehr, R. C., & Swartz, J. D. (1983). Holtzman Inkblot Technique scores of delinquent adolescents: A replication and extension. *Journal of Personality Assessment*, 47, 158–160.
- Nichols, D. C., & Tursky, B. (1967). Body image, anxiety, and tolerance for experimental pain. *Psychosomatic Medicine*, 29, 103–110.
- Peterson, R. A. (1978). Holtzman Inkblot Technique. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 947–849). Highland Park, NJ: Gryphon Press.
- Pokhriyal, R., & Ahmad, H. (1988). Response patterns of acute schizophrenics and endogenous depressives on the Holtzman Inkblot Technique. *Journal of Personality and Clinical Studies*, 4, 205–207.
- Reinehr, R. C., Swartz, J. D., & Dudley, H. K. (1984). Ethnic differences in the measurement of hostility in forensic patients. *Revista Interamericana de Psicologia*, 18, 53–64.
- Sundberg, N. D. (1962). The Rorschach Americanized. *Contemporary Psychology*, 7, 250–252.
- Swartz, J. D. (1969). *Pathognomic verbalizations in normals, psychotics, and mental retardates*. Unpublished doctoral dissertation, University of Texas.
- Swartz, J. D. (1992). The HIT and the HIT 25: Comments and clarifications. *Journal of Personality Assessment*, 58, 432–433.
- Swartz, J. D., & Holtzman, W. H. (1963). Group method of administration for the Holtzman Inkblot Technique. *Journal of Clinical Psychology*, 19, 433–441.
- Swartz, J. D., & Reinehr, R. C. (1983). A quick version of administration for the group Holtzman Inkblot Technique. *Perceptual and Motor Skills*, 56, 813–814.
- Swartz, J. D., Reinehr, R. C., & Holtzman, W. H. (1983). Personality development through the lifespan: Assessment by means of the Holtzman Inkblot Technique. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 3, pp. 35–51). Hillsdale, NJ: Erlbaum.
- Swartz, J. D., Reinehr, R. C., & Holtzman, W. H. (1999). *Holtzman Inkblot Technique, Research Guide and Bibliography*. Austin, TX: Hogg Foundation for Mental Health.
- Swartz, J. D., Witzke, D. B., & Megargee, E. I. (1970).

Normative item statistics for the group form of the Holtzman Inkblot Technique. *Perceptual and Motor Skills*, 31, 319–329.

Vincent, K. R. (1982). The fully automated Holtzman interpretation. In K. Herman & R. M. Samuels (Eds.), *Computers: An extension of the clinician's mind, a reference book* (pp. 123–125). Norwood, NJ: Ablex.

von Rosenstiel, L. (1973). Increase in hostility responses in the HIT after frustration. *Journal of Personality Assessment*, 37, 22–24.

Zubin, J. (1954). Failures of the Rorschach technique. *Journal of Projective Techniques*, 18, 303–315

## Chapter 3

# ***PSYCHOMETRIC FOUNDATIONS OF ASSESSMENT***

This chapter provides an overview of the general psychometric information that clinicians must consider when interpreting any assessment test or technique. It is assumed that the reader is familiar with basic descriptive statistics such as measures of central tendency (mean, median, mode), dispersion (standard deviation, variance), and deviations from a normal distribution (kurtosis, skewness). This chapter also describes a multitude of issues that arise when making decisions about clinical patients. The emphasis is on maximizing the accuracy and effectiveness of these decisions, not on the issues per se, which have been described thoroughly by Garb (1998).

All the assessment tests and techniques reviewed in this *Handbook* follow classical test theory (CTT) rather than item response theory (IRT). The assumption in CTT is that the person's actual or obtained score on an assessment test or technique is a function of both a true score and the error associated with measuring it, and that this error is invariant across the entire range of scores. Classical test theory weights items equally regardless of their actual difficulty, so that infrequently endorsed (difficult) items are weighted equally to frequently endorsed (easy) items. To the extent that any of these assumptions cannot be met, which is more often than not, there are potential limitations in CTT.

There are many excellent sources of information on IRT (cf., Embretson & Reise, 2000), and further consideration of it is outside the scope of the present *Handbook*. However, IRT has been proposed as an alternative method in scale construction for addressing problems that arise when CTT assumptions are not met. Because IRT is not sample dependent, it allows generalization of results across samples more directly than CTT. Item response theory also allows for the effectiveness of items to be assessed at all different levels of a scale or at any specific level, if that is desired. The field of personality assessment needs to consider whether moving away from CTT toward IRT, or some other approach, would be beneficial. The following sections cover basic considerations in assessing reliability, determining validity, and making decisions based on assessment data. For additional discussion of psychometric aspects of psychological testing, readers are referred to the classic book on CTT by Gulliksen (1950) and contemporary texts by Anastasi and Urbina (1997), Kaplan and Saccuzzo (2001), and Urbina (2004).

### **RELIABILITY**

Any variable must be assessed reliably before its validity can be examined, where reliability is defined as the consistency with which a variable is measured. The reliability of an

49

#### **50 Basic Considerations**

assessment test or technique also sets the upper limits for its validity. Consequently, reliable measurement with any assessment test or technique is mandatory before its validity can even be considered.

Reliability is usually reported as a correlation coefficient, with reliability coefficients of at least .75 as a general standard for whether a variable is being assessed reliably. Three types of reliability are discussed here: (1) test-retest or temporal reliability; (2) internal consistency reliability; and (3) interrater reliability. Two other types of reliability—parallel-form or alternate-form reliability and split-half reliability—are not discussed, because they are not used with any of the assessment tests or techniques reviewed in this *Handbook*.

Three related constructs are also examined in this section on reliability: (1) standard error of measurement and cutting scores; (2) stability of test scores; and (3) redundancy of test scores.

In reviewing the reliability of a scale, clinicians need to give some thought to the appropriateness of the measure of reliability that has been used as well as to the magnitude of the correlation coefficient that has been obtained. They can compare the reliability of the scale that is being considered for use with similar scales and their reliability, if they exist. Clinicians should provide the rationale for using a scale whose reliability is significantly

lower than similar scales with comparable validity. This rationale could be included in the psychological report to inform readers of the decision that has been made.

Generally speaking, the longer a scale is, the more reliable it is likely to be, and vice versa, because the more items there are in a scale, the more adequately the scale samples the domain being measured. One of the potential shortcomings of brief scales is their limited reliability. Test developers must contend with the challenge of keeping a scale as short as possible while maintaining adequate reliability.

### **Test-Retest (Temporal) Reliability**

Test-retest reliability is assessed by administering an assessment test or technique twice to the same group of participants within some relatively short interval of time, ranging from 1 day to a few weeks. Such brief retest intervals are used when the construct being assessed is expected to vary across time, just as a person's mood is likely to change from time to time. Most of the assessment tests and techniques reviewed in this *Handbook* meet or exceed the criterion of .75 for short-term test-retest reliability when it is used as the measure of reliability. There is, however, a growing trend in the field of assessment to use internal consistency together with or instead of test-retest as the measure of reliability.

### **Internal Consistency Reliability**

A measure of internal consistency for a scale assesses the homogeneity among the items it contains and is basically the average intercorrelation among all the items on the scale. Cronbach's coefficient alpha (1951) is the measure most often used to report internal consistency. Again, the criterion of at least .75 is a general standard for an acceptable level of internal consistency for a scale, and most assessment tests and techniques will meet or exceed this criterion. Because published scales seldom fail to meet this criterion, it is easy for clinicians to overlook the importance of checking the reliability of any assessment tests and techniques that they use.

#### **Psychometric Foundations of Assessment 51**

There is ordinarily a preferred upper limit to the internal consistency of a given scale because as coefficient alpha approaches 1.00, the items become increasingly redundant and provide little, if any, additional information about the underlying construct being assessed. In addition, if the construct is relatively heterogeneous, then the scale must also embody some heterogeneity. In such an instance, a lower level of internal consistency, possibly even as low as .70, might be acceptable. A scale measuring moods that consists entirely of items assessing depression, and does not include items assessing anxiety, anger, and other affective states, might have excellent internal consistency, but still be a poor measure of the more general construct of moods, which the scale was intended to assess.

### **Interrater Reliability**

Intuitively, determining interrater reliability would seem to be simple—just count the number of times two raters agree and calculate the percentage of agreement in their judgments or decisions. This simplicity dissipates quickly when consideration is given to the chance level of agreement expected in these judgments or decisions, how frequently the measured behaviors or symptoms occur, and how many categories are being judged simultaneously. It is much easier to achieve reliable agreement when ratings are only for whether a given behavior or symptom is present than when they are quantified into three or more levels. There are three methods for measuring the consistency of interrater agreement: (1) the percentage of agreement just mentioned, (2) the interclass correlation coefficient, and (3) the kappa coefficient. These different methods have evolved because two raters can rank order their judgments similarly, and appear to be in perfect agreement, but one can be assigning consistently higher or lower scores than the other, which would mean that there is actually little agreement in their ratings. Interclass correlation coefficients provide a measure of the consistency of the scores assigned, not merely their rank order, either for a single rater or as a measure of the consistency in ratings among several raters. As for the limitation of a percentage of agreement, it is necessary to determine the chance level of agreement to evaluate whether the raters are achieving an increment in consistency beyond what would be expected by chance.

To this end, Cohen (1960) proposed the kappa coefficient, which is generally a better measure for evaluating the agreement between two raters than percentage of agreement,

because it provides a correction for the level of agreement expected by chance. If two raters are deciding the presence or absence of a behavior or symptom across a number of individuals, they should agree 50% of the time simply by chance. They would need to improve beyond this level of 50% agreement to demonstrate that they were rating the behavior or symptom reliably. In this example, if two raters agreed 85% of the time, they would demonstrate that their ratings are better than chance using the kappa coefficient. In fact, the kappa coefficient is determined by the formula:  $(\text{Observed agreement} - \text{Chance agreement}) / (1 - \text{Chance agreement})$ , so in this instance  $\text{kappa} = (.85 - .50) / (1 - .50) = .70$ .

### Standard Error of Measurement (Confidence Intervals)

In classical test theory, a person's true score on a scale or test is assumed to be composed of two components, the person's actual or obtained score and the error associated with that score. This error is expressed as the standard error of measurement (*SEM*), which is

#### 52 Basic Considerations

**Table 3.1 Changes in the standard error of measurement (*SEM*) with changes in the reliability of a scale**

$$r = .90 \text{ } SEM = 10 \times \text{SQRT}(1 - .90) = 10 \times 0.32 = 3.20$$

$$r = .80 \text{ } SEM = 10 \times \text{SQRT}(1 - .80) = 10 \times 0.45 = 4.50$$

$$r = .70 \text{ } SEM = 10 \times \text{SQRT}(1 - .70) = 10 \times 0.55 = 5.50$$

$$r = .60 \text{ } SEM = 10 \times \text{SQRT}(1 - .60) = 10 \times 0.63 = 6.30$$

$$r = .50 \text{ } SEM = 10 \times \text{SQRT}(1 - .50) = 10 \times 0.71 = 7.10$$

*Note:* The standard deviation was assumed to be 10 in these examples.

the normal variability that would be expected in an individual's score if the scale or test were administered a large number of times. The *SEM* is a direct function of the standard deviation and reliability of the scale ( $SEM = SD \times \text{SQRT}[1 - r]$ ). As the scale becomes less reliable, the *SEM* increases correspondingly. This variability, which is assumed to follow a normal distribution, is expressed in standard deviation units within which the individual's "true" score is expected to fall. If the reliability for most assessment tests or techniques is approximately .80, the *SEM* is about .5 *SD* (see Table 3.1).

If the *SEM* for a scale is 5.0 points, then two-thirds of the time, the individual's true score will be within  $\pm 5.0$  points of the obtained score. This range of  $\pm 1.0$  *SEM* is the 68% confidence interval, the range of  $\pm 2.0$  *SEM* is the 95% confidence interval, and so on. Confidence intervals sometimes are called the error bands for a score.

The *SEM* and confidence intervals become very important when cutting scores are being employed on a scale. When the obtained score on the scale is exactly at the cutting point at least 50% of the time, the individual's true score would be expected to be below the cutting score. If the cutting score for a scale is 25 to indicate the presence of some behavior or symptom, 50% of the time the individual's true score is below 25. If the *SEM* is 3.0, even when the individual has an obtained score of 28 (+1.0 *SEM*), 16% of the time the individual's true score still would be below the cutting score of 25. If the individual has an obtained score of 22 (-1.0 *SEM*), 16% of the time the individual's true score will be above the cutting score. As a general rule of thumb, an individual's obtained score should be at least .5 to 1.0 *SEM* above the cutting score so that there is reasonable probability that the individual's true score would exceed the cutting score. By the same token, it is optimal when an individual's obtained score is at least .5 to 1.0 *SEM* below the cutting score, so there is a reasonable probability that the individual's true score does not exceed the cutting score.

When a cutting score is developed on a scale, the scale is being validated as a dichotomous or categorical variable, that is, whether the person is above or below the cutting score. Consequently, higher scores on the scale above the cutting score have a higher probability that the person's true score is above the cutting score, not that the person has more symptoms or that the symptoms are more severe. Conversely, lower scores below the cutting score have a higher probability that the person's true score is below the cutting score, not that the person has fewer symptoms or the symptoms are less severe. Another way of saying the same thing is that a dichotomous or categorical variable has been validated, not a dimension of quantity or severity of the behavior or symptom.

It is important to realize that artificially dichotomizing any variable that is dimensional will result in a loss of power and is generally not recommended (Cohen, 1983). Taxometric

### Psychometric Foundations of Assessment 53

procedures have been developed by Meehl and his colleagues (cf. Waller & Meehl, 1997) as a means of determining whether a variable is dimensional, and hence should not be dichotomized, or categorical.

### Stability of Test Patterns or Profile Scores

The stability of test scores is sometimes used as an alternate way of referring to the short-term test-retest (temporal) reliability already discussed, but there is some advantage in reserving this term for descriptions of test-retest correlations over an extended period. As used here, then, stability of test patterns or profile scores refers to similarity or change in test scores across some extended interval and indicates how consistently individuals obtain the same pattern of scores on two administrations of an assessment instrument, one at the beginning and the other at the end of this period of time.

The assessment instruments presented in Chapters 6 through 14 of this *Handbook* are interpreted primarily or in part on the basis of patterns of scores that persons obtain on them. In addition to establishing the reliability of the individual scales used for this purpose, as indicated by either short-term test-retest or internal consistency data, it is important to determine whether there also is stability over time in the pattern of scores on which interpretations are based.

There has only been limited research on this topic, most of which has been restricted to the original Minnesota Multiphasic Personality Inventory (MMPI) and to the Rorschach Inkblot Method (RIM). With respect to the MMPI, Graham, Smith, and Schwartz (1986) reported 42.7%, 44.0%, and 27.7% agreement across an average interval of approximately 3 months for high-point, low point, and two-highest scales, respectively, in 405 psychiatric inpatients. Greene, Davis, and Morse (1993, March) provided similar data on 454 alcoholic inpatients who had taken the MMPI during two different hospitalizations separated by at least 6 months. Approximately 40% of the men and 32% of the women had the same high-point scale on the two successive administrations of the MMPI. However, they had the same two highest scales only 12% and 13% of the time, respectively. Finally, Livingston, Jennings, Colotla, Reynolds, and Shercliffe (2006) reported 45% and 24% agreement for high-point scale and two highest scales on the MMPI-2, respectively, in 94 injured workers evaluated after approximately 2 years (see also Chapter 6).

Two conclusions are readily apparent from these limited data. First, a pattern of test or profile scores on self-report inventories appears to be stable in fewer than 50% of the individuals who have been evaluated. Consequently, clinicians should be very cautious about making any long-term predictions from a single administration of a self-report inventory, because a different pattern of test or profile scores is more likely than not, if the person is reevaluated. Second, more research is needed on the stability of the pattern of test or profile scores for all self-report inventories.

As for the RIM, Exner (1993, chap. 11) reported retest correlations for a sample of 100 adult nonpatients examined after a 3-year interval and a separate sample of 50 adult nonpatients examined after a 1-year interval. Over the 3-year interval, 22 of 32 Rorschach variables examined had retest correlations of .75 or more (with nine at .85 or higher); and over the 1-year interval, 30 of 41 variables examined reached the .75 criterion of adequacy (with 15 at .85 or higher). Most of the examined variables with stability coefficients below

#### 54 Basic Considerations

.75 are intended to measure state dimensions of personality (e.g., moods) that are expected to change over time.

Sultan, Andronikof, and Réveillére (2006) reported retest correlations for the Rorschach in a sample of 75 French nonpatient adults who were retested after a 3-month interval. The median correlation was .53, significantly lower than the values reported by Exner (2003). There were a number of factors that appeared to moderate the stability of these coefficients: overall level of engagement in the task, level of state distress, and number of responses. In Rorschach assessment, then, there is some basis for making long-term predictions related to trait characteristics of individuals. However, such predictions are not warranted on the basis of variables measuring states of the individual, and the stability of scores derived from other performance-based measures has yet to be examined.

## Redundancy of Test Scores

Sometimes clinicians confuse the redundancy of self-report inventory scores with their validity. For example, they might administer the MCMI-III, MMPI-2, and Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996) to the same person and find elevated scores on all the scales assessing depression on these inventories (MCMI-III: Depressive [*2B*], Dysthymic Disorder [*D*], and Major Depression [*CC*]; MMPI-2: Scale 2 [*D*] and Depression [*DEP*]; BDI). The clinician might then be prone to conclude that this individual must really be depressed because of the similar pattern of elevated scores across all of these self-report inventories. In doing so, they would fail to realize that variations of the same questions are asked on all these measures. The individual has been very reliable (consistent) about reporting symptoms of depression, but this does not make the person any more depressed. This redundancy of test scores can be seen by examining the intercorrelations among all these scales, which exceed .70 in most cases.

This issue of redundant test scores should be considered carefully when selecting a battery of assessment instruments for a specific individual. At the same time, avoidance of redundancy in test selection should be distinguished from the confirmatory value of congruent findings that are obtained from different kinds of measures. As discussed in Chapter 2, patterns of convergence and divergence between self-report and performance-based test protocols can provide valuable information about an examinee's psychological condition and frame of mind.

A second type of redundancy in the interrelations among a set of variables derives from heterogeneity in the variables and may be particularly difficult to recognize. If assessors wish to predict school learning problems in adolescents, for example, they need to appreciate that a number of psychosocial variables have a low negative (–.16 to .30) relationship with academic performance, including behavior problems at school, alcohol and drug use, family instability, and poor interpersonal relationships. These variables may mistakenly be construed as being independent, when they actually reflect a complex set of specific variables that assess the same general set of problem variables. This redundancy can be demonstrated by a hierarchical regression analysis that examines incremental validity (covered in the next section) with the addition of each variable. There is almost no increment in the ability to predict academic performance as each of the just mentioned variables is added, even though they might seem to be assessing different behaviors.

### Psychometric Foundations of Assessment 55

## VALIDITY

Until an assessment test or technique produces reliable data, there is no reason to proceed to the step of attempting to validate it. Once an assessment test or technique has been determined to be reliable enough to meet the previously indicated psychometric standard, then the issue of its validity becomes paramount. The validity of an assessment test or technique consists of how well or accurately it measures the constructs it is intended to measure.

Unlike the case with reliability coefficients, there is little consensus about the magnitude validity coefficients should attain. Nevertheless, it is important to keep in mind that the reliability coefficient for a score or scale sets the upper limit for its validity coefficient. A validity coefficient for a score or scale cannot exceed the square root of its reliability coefficient. That is, a score or scale with a reliability coefficient of .80 cannot correlate higher than .89 with any other measurement; a score or scale with a reliability coefficient of .70 cannot have any validity coefficient greater than .84; and so on. Validity coefficients between two self-report scales should be larger than validity coefficients between a self-report scale and a behavioral rating, or between a self-report scale and a performance-based scale. The shared method variance between two self-report scales (i.e., the constructs are being measured in the same manner), almost inevitably produces larger correlations than the other two types of comparison.

It cannot be stated too strongly that any assessment test or technique is not valid in the abstract, but is valid for the assessment of a specific construct or set of constructs within a given assessment setting. One of the fundamental issues of validity is how well the results of an assessment test or technique in one set of circumstances generalize to other circumstances. An assessment test or technique may adequately identify a behavior

or symptom in one clinical setting, but for a variety of reasons may not work as well in another setting or with different patients. The comparability of the clinical setting and individuals used in the validation process has to be considered carefully when clinicians consider adopting an assessment test or technique they have not previously used in their particular setting.

There are multiple ways of classifying validity as well as a growing trend to consider all forms of validity as construct validity, which refers to how well an assessment test or technique measures a specific construct. Only three types of validity are considered here: convergent, discriminant, and incremental.

### **Convergent Validity**

Convergent validity is a measure of how well an assessment test or technique correlates with other assessment tests or techniques that are measuring *similar* constructs. A new scale designed to measure depression should correlate with extant measures of depression; if not, there is something amiss with one or the other of them. Whether another scale of depression is needed is addressed in discussing incremental validity.

A problem with most self-report inventories is that they assess the general negative emotional distress or misery that is common to all types of psychopathology and consequently demonstrate a substantial level of convergent validity regardless of the specific construct being assessed. This problem is reflected in discussions of whether depression and anxiety

#### **56 Basic Considerations**

are separate diagnostic entities or different manifestations of general negative emotional distress (see Barlow & Campbell, 2000). The inclusion of a diagnosis of mixed anxiety and depression in *DSM-IV-TR* (American Psychiatric Association, 2000) appears to attest to the comorbidity of these disorders.

In self-report assessment of psychopathology, a current issue is whether inventories can assess specific constructs independently of general negative emotional distress or misery. As elaborated in Chapter 6, the Restructured Clinical scales of the MMPI-2 (Tellegen et al., 2003) represent an attempt to remove general negative emotional distress (demoralization) from the clinical scales of the MMPI-2. The interested reader should consult the ongoing debate about the success of this venture (Caldwell, 2006; Nichols, 2006; Tellegen, Ben-Porath, & Sellbom, 2006). It is not clear what mood states like depression, anxiety, and anger would consist of if the general negative emotional distress were removed from them.

### **Discriminant Validity**

Discriminant validity is a measure of whether an assessment test or technique is relatively unrelated to other assessment tests or techniques that assess *dissimilar* constructs. If a scale measuring depression has a significant positive correlation with a scale measuring assertiveness and positive self-regard, there is a rather serious validity problem with one or the other scale. However, a significant and large negative correlation with a scale measuring assertiveness and positive self-regard could be an example of convergent validity because these two constructs should be inversely related.

It is difficult to know how unrelated two scales measuring dissimilar constructs actually should be, because in most cases, it would be unusual for them to be correlated negatively. Validity coefficients for two dissimilar variables should be lower than those for convergent validity for two similar constructs. Campbell and Fiske (1959) advocated multitrait, multimethods for ascertaining the convergent and discriminant validity of a set of constructs simultaneously, so that these validity coefficients can be compared directly.

### **Incremental Validity**

Incremental validity addresses two issues: (1) whether additional information from another assessment test or technique improves or enhances the clinical judgments that are made, and (2) whether any new scale or index contributes additional information beyond extant scales or indices to the prediction of relevant behaviors. There is a rather consistent literature indicating that clinicians' judgments become less accurate when they consider more than three or four variables simultaneously (see Garb, 1998). With multiple variables to evaluate, it can be difficult, sometimes impossible, for clinicians to determine which variables are the most important for a specific judgment and how to weight these variables to maximize accuracy. Whether a new scale should demonstrate incremental validity over extant scales is less

clear. This expectation can be justified to the extent that it limits the proliferation of scales, but it does not take into account the not uncommon differential validity of scales across settings and situational demands. At a minimum, the developer of a new scale or index should report its relationship with conceptually related scales or indices, along with a rationale for why this new scale or index should supplant an existing scale or index if incremental validity is not demonstrated.

#### **Psychometric Foundations of Assessment 57**

Garb (2003) has reviewed the results for the incremental validity of the addition of various assessment tests and techniques to the clinical interview in the evaluation of psychopathology in adults. There was no general pattern to the results that Garb reported, partly because of the paucity of such research. Several specific scales or indices were found to provide incremental validity, including the Borderline scale of the Personality Assessment Inventory (Morey, 1991) in the longitudinal prediction of academic performance and interpersonal problems in college students (Trull, Useda, Conforti, & Doan, 1997) and Rorschach measures of thought disorder. The importance of such demographic variables as the age of onset of the disorder, lifetime history of psychopathology, family history of psychopathology, and social and familial support have not been considered in this research.

### **CLINICAL DECISION MAKING**

The psychometric foundations of personality assessment include several special considerations that clinicians must take into account when they make decisions based on the information they obtain. The most important of these considerations are the impact of prevalence (base rate) data, the distinction between clinical and statistical judgment, the meaning of risk factors and odds ratios, the effects of heterogeneity among patients within diagnostic groups, the implications of diagnostic efficiency statistics, and setting of cutting scores. Each of these considerations is amplified in the sections that follow.

#### **Prevalence (Base Rate) Data**

No concept is more important in making decisions in clinical settings than the concept of prevalence (base rate), which is the frequency with which a given behavior or symptom occurs in a given group of individuals or patients. *DSM-IV-TR* (American Psychiatric Association, 2000) is an excellent source for basic information on the frequency of various forms of psychopathology in the general population. According to estimates, the lifetime prevalence of major depressive disorder is 10% to 25% in women and 5% to 12% in men; the lifetime prevalence of alcohol dependence in men is 15%; and the lifetime prevalence of schizophrenia is .5% to 1.5%. These prevalence estimates mean that 20% to 25% of women in the general population will experience an episode of diagnosable depression in their lifetime; nearly 15% of men will be diagnosed as alcoholic; and about 1% of people will experience a schizophrenic episode. Stated in another way, clinicians are likely to see approximately 25 depressed women patients and 15 male alcoholic patients for every schizophrenic patient.

Somewhat surprisingly, there are fewer sources of statistics on the frequency with which various forms of psychopathology are seen in clinical settings than there are concerning their prevalence, probably because these statistics are influenced by the clinical setting. A setting that specializes in the diagnosis and treatment of a specific disorder like borderline personality disorder or posttraumatic stress disorder is quite likely to have a higher prevalence for these disorders than a community clinic that is required to treat all types of patients.

The more specific the information clinicians have about the prevalence of various behaviors and symptoms in their specific setting, the more likely they are to make good decisions.

Meehl (1954) long ago recognized the need for local norms, that is, the frequency with

#### **58 Basic Considerations**

which a given behavior or symptom occurs in a specific setting. Most clinical settings are likely to attract particular types of patients. Hence the general prevalence that is presented in *DSM-IV-TR* (American Psychiatric Association, 2000) may be misleading for patients who are seen more or less frequently in any specific setting.

In settings in which the prevalence of some behavior or symptom is either very high (>80%) or very low (<20%), assessors face the formidable task of finding assessment methods that can achieve more than 80% accuracy. In many such circumstances, clinicians

can simply use the prevalence of the behavior or symptom to decide whether it is present in an individual, which will make them accurate 80% of the time without any reliance on assessment tests or techniques.

### **Clinical versus Statistical Judgment**

The discussion in Chapter 2 of the basis on which inferences can be drawn from test data distinguished between clinical and statistical judgment as alternative avenues for arriving at conclusions. Because of the importance of this distinction in the clinical decisionmaking process, and because research findings concerning the interpretive accuracy of these two methods are often ignored, some further attention to clinical versus statistical judgment is appropriate for the present chapter on psychometric foundations of psychological assessment.

To recap these approaches to decision making, statistical judgment is a quantitative approach in which empirical data are mechanically combined to provide actuarial rules for determining what test findings indicate. Clinical judgment, by contrast, is a qualitative approach in which shared beliefs of experienced clinicians are cumulated to determine what test findings signify. Readers are encouraged to consult the several sources referenced in Chapter 2 to gain an in-depth appreciation of these alternative avenues of interpretation and the controversy that has at times swirled around them.

Clinical and statistical judgment have often been pitted against each other—both in the minds of assessment psychologists and in comparative research studies—about which is the better and more accurate way to interpret test data. The empirical findings in this regard are as clear as perhaps any findings pertaining to issues in clinical psychology. As shown in the meta-analyses mentioned in Chapter 2 by Grove, Zald, Lebow, Snitz, and Nelson (2000) of 136 studies of the relative accuracy of clinical and statistical judgment, and by Ægisdóttir et al. (2006) of 67 such studies, statistical judgment is consistently as good as and typically better than clinical judgment in making accurate decisions.

Three other findings in these meta-analyses are important to note. First, the accuracy of both clinical and statistical judgments varied with the type of decision being made, the setting in which data are collected, the type of statistical formula used, and the amount of information available to the clinicians and for inclusion in the formulas. Second, the increase in accuracy of statistical over clinical judgments was modest, with statistical judgments being 13% more accurate on the average than clinical judgments in the Ægisdóttir et al. (2006) study and 10% more accurate in the Grove et al. (2000) study. Third, clinical judgment equaled the accuracy of statistical judgment in many instances and in some cases was even more accurate than a formula. As suggested in Chapter 2, clinical and statistical methods are better seen as complementary than as mutually exclusive approaches

#### **Psychometric Foundations of Assessment 59**

to decision making (see pp. 31–34). Relying on both approaches in a balanced manner based on their probable value in a given instance is thus a good way for assessors to enhance the accuracy of the conclusions they draw from their test data.

There are, however, few areas within clinical psychology where available research is ignored more routinely than in appreciating the differential accuracy of clinical and statistical judgment. Some practicing clinicians may place more confidence in their clinical judgment and give less credence to statistical judgment than is warranted by research findings. Psychologists practicing personality assessment may even be unaware of the content and implications of literature concerning the accuracy of clinical and statistical judgments. Studies reviewed by Garb (1998) suggest that clinicians are often uncertain of the variables on which a judgment should be based, or at least of the relative importance of variables pertinent to this judgment, and that they commonly rely in their decision making on variables with little demonstrated relationship to what they are trying to determine. Routine application of available statistical guidelines in making clinical decisions would contribute substantially to improving the overall quality of personality assessments and therapeutic interventions as well.

Moreover, the limitations of statistical judgments noted in Chapter 2, although important to keep in mind, do not necessarily prevent them from serving useful purposes. With respect to generalization, a statistical rule that is less accurate in a different setting or

with a demographically different group than it was in its original derivation may still be more accurate in the new circumstances than a clinical impression that is formed without attention to empirical guidelines.

Similarly, the fact that statistical judgments concern a group of individuals rather than any specific, unique individual does not preclude their providing accurate information about a person who has been examined. To the contrary, although reliable information from other sources may indicate that certain statistically based interpretations do not apply in a particular case, a set of carefully developed statistical rules is likely to include many that describe most persons accurately. In those instances in which a statistical rule clearly does not apply, or when no empirical guideline is available, assessment psychologists must rely on informed clinical judgment. They should not, however, allow their clinical impressions to dissuade them from appropriate consideration of research findings, and they should strive always to base their decisions on a well-reasoned integration of statistical and clinical judgments.

### **Risk Factors and Odds Ratios**

Risk factors are behaviors or symptoms that have been identified empirically as related to some outcome. This concept is frequently used in discussions of factors associated with dangerousness to the self (Packman, Marlitt, Bongar, & Pennuto, 2004) or to others (Monahan et al., 2005). Factors that increase the likelihood of people harming themselves include impulsivity, a sense of hopelessness, a history of previous suicidal behavior, and evidence of having a lethal plan. Because each of these factors increases the risk of dangerousness to the self, they are known as *risk factors*. With respect to dangerousness to others, Monahan et al. developed an actuarial model for violence risk assessment that exemplifies the previously noted soundness of statistical judgments when there are sufficient facts to support them.

#### **60 Basic Considerations**

Odds ratios are one way of expressing the risk of a specific behavior or symptom, provided that clinicians keep in mind relevant prevalence data. Suppose it is known that the odds of developing a schizophrenic disorder are increased by 3.5:1 if a certain condition or risk factor is present. With the previously noted lifetime prevalence of schizophrenia being 0.5% to 1.5% (American Psychiatric Association, 2000), a 3.5 increase in the odds ratio would raise the likelihood of becoming schizophrenic only to 1.75% to 5.25%, which is far from being a probable outcome. Even if the prevalence of schizophrenia were 12.5%, a 3.5 increase in the odds for a particular person would raise the likelihood just to 43.75%, and this person would still have less than a 50% probability of becoming schizophrenic. Given that the lifetime prevalence for most forms of psychopathology is well below 10%, odds ratios can sometimes be misunderstood by clinicians and by their patients as well. For infrequently occurring disorders, increased odds ratios do not substantially increase their likelihood of being present. The only exceptions, and the only disorders with a lifetime prevalence exceeding 10%, as previously noted, are major depression in women (10% to 25%) and alcohol dependence in men (15%; American Psychiatric Association, 2000).

### **Heterogeneity in Patients within the Same Diagnostic Group**

Experienced clinicians are well aware that there is considerable variability among persons within any diagnostic group. Despite this awareness, clinicians often talk about typical patients by referring to them in such terms as “alcoholics” and “borderlines,” as if persons with alcohol dependence or borderline personality disorder are all alike. One of the many exciting aspects of clinical work is the discovery of the uniqueness of each patient within a specific diagnostic group.

This erroneous idea of homogeneity in diagnostic groups may seem to be supported by the fact that group mean profiles or a set of expectations for them can be constructed for the assessment tests and techniques discussed in this *Handbook*. Moreover, these group mean profiles and expectations are different for various diagnostic groups, which could be taken as further evidence of homogeneity within those groups. Data such as these could lead clinicians to assume that all patients within a diagnostic group produce pretty much the same set of scores. To the contrary, only a small portion of patients in a diagnostic group are likely to show the same scores as the group mean profile. Because of this heterogeneity,

clinicians should be wary of any study that reports only a mean profile on an assessment test or technique. Frequency distributions of the various scores produced in each diagnostic group provide clinicians with more information on group performance on any measure than that provided by the group mean profile.

The use of polythetic criteria for diagnoses in the *Diagnostic and Statistical Manual (DSM-IV-TR)* published by the American Psychiatric Association (2000) contributes to the heterogeneity of groups of patients within specific diagnostic categories. Polythetic diagnostic criteria require only that some of the common features of a condition be present for it to be diagnosed, as in calling for any five or more of nine listed features in diagnosing Major Depressive Episode and five of nine features in diagnosing Borderline Personality Disorder. When disorders are defined by polythetic criteria, it is possible for two patients with the same disorder to share only a few, or even none, of the features of the disorder. As notable exceptions, *DSM* diagnoses of Eating Disorders and Sleep Disorders require the presence of all their listed features, which means that these diagnostic groups are likely to be relatively homogeneous.

#### **Psychometric Foundations of Assessment 61**

Polythetic diagnostic criteria reflect the heterogeneity seen in clinical practice among persons with various disorders, as well as commonly occurring overlaps and lack of firm boundaries between disorders. In this sense, they are quite appropriate to apply, provided that their heterogeneity is kept in mind. As a point of historical interest, successive editions of the *DSM* over the years have defined an increasing number of disorders by polythetic criteria, thereby increasing as well the heterogeneity of these disorders.

#### **Diagnostic Efficiency Statistics**

The term *diagnostic efficiency statistics* refers to several ways of describing the accuracy or efficiency of decisions made on the basis of assessment data. A thorough understanding of the implications of these statistics is a basic aspect of any judgment about whether a personality or behavioral characteristic is present in an individual or group of individuals. In particular, diagnostic efficiency statistics demonstrate the impact of prevalence on the accuracy of judgments that are being made and show how important it is for clinicians to be aware of available prevalence data for whatever characteristics are being judged.

As background for defining and illustrating diagnostic efficiency statistics, clinicians should recognize that, in whatever kind of setting they are conducting assessments, they are constantly making judgments, such as deciding in mental health settings whether a person is anxious, depressed, schizophrenic, suicidal, passive-dependent, obsessive-compulsive, or interpersonally aversive. Most such states and dispositions can exist and be measured in degrees (e.g., highly anxious, mildly depressed, somewhat dependent, severely obsessive), but they can also be characterized in a dichotomous fashion as being present or absent, or at least as being more likely to be present than absent, and vice versa.

In similar fashion, the development and validation of an assessment instrument also involves dichotomous distinctions. If an assessment specialist wants to develop a new test for assessing problems with alcohol, an essential requirement is identifying some gold standard against which the new test can be validated. A frequently used standard in developing a test for alcohol problems is a group of people participating in an alcohol treatment program. Scores on the new test in this alcohol treatment group can then be compared with the scores of other persons in the same treatment facility who are not in the alcohol program or of normal individuals in the community. Which of these two possible comparison groups is used can substantially affect the diagnostic efficiency statistics, as illustrated next.

The use of persons being treated in an alcohol program as the gold standard is somewhat problematic because the individuals the test developer really wants to identify are people whose current and ongoing alcohol use is having a negative impact on their work, social, and interpersonal functioning. Such people are difficult to identify in sufficient numbers to validate an assessment test or technique, whereas persons in alcohol treatment programs, who may or may not currently be experiencing these negative impacts, are plentiful and relatively easily accessed. Because convenience samples of treatment program participants are often used as the gold standard in validating alcohol problems assessment measures,

practitioners using one of these measures should read the description of its validation samples carefully and decide how comparable it is to people in their particular setting. Table 3.2 shows a model for reporting the diagnostic efficiency statistics for a hypothetical new assessment test for alcohol problems. The various terms in Table 3.2 are defined next, following which actual examples are used to demonstrate the impact of different

## 62 Basic Considerations

**Table 3.2 Defining terms for diagnostic efficiency statistics**

Gold Standard Criterion

Present Absent

Present

*True positives*

**a**

*False positives*

**b**

New Test

Absent

*False negatives*

**c**

*True negatives*

**d**

True positives (sensitivity) =  $a/(a + c)$

False negatives =  $c/(a + c)$

True negatives (specificity) =  $d/(b + d)$

False positives =  $b/(b + d)$

Overall correct classification =  $(a + d)/(a + b + c + d)$

*Note:* **a**=Alcoholic patients identified as alcoholics by the newestest; **b**=Normal individuals misidentified by the new test as alcoholics; **c** = Alcoholics misidentified as normal individuals by the new test; **d** = Normal individuals identified as normal individuals by the new test.

prevalence rates on the diagnostic efficiency statistics. Deferring for the moment the method of determining an optimal cutting score on the new assessment test for identifying individuals with alcohol problems, scores above the cutting score are hypothesized as indicating that alcohol problems are present, and scores below the cutting score as indicating that alcohol problems are not present. In this first example, it is assumed that the comparison group for the alcohol treatment group is normal individuals from the community.

Once the new test has been given to both groups of individuals, there are four possible outcomes:

1. *True positives*, which consist of alcohol patients who are identified correctly by the new test as having alcohol problems.
2. *False positives*, which consist of normal individuals who are identified incorrectly by the new test as having alcohol problems.
3. *False negatives*, which consist of alcohol patients who are identified incorrectly by the new test as not having alcohol problems.
4. *True negatives*, which consist of normal individuals who are identified correctly by the test as not having alcohol problems.

The true positive rate also is known as the *sensitivity* of the test, and the true negative rate is known as the *specificity* of the test. The *hit rate* or *overall correct classification* rate is the total of true positives plus true negatives divided by the total number of participants in both groups, which gives the percentage of correct classifications. An assessment test or technique with a high rate of true positives or sensitivity would be very good at identifying that alcohol problems are present, other things being equal; whereas an assessment test or technique with a high rate of true negatives or specificity would be very good at identifying that alcohol problems are absent.

## Psychometric Foundations of Assessment 63

**Table 3.3 Diagnostic efficiency statistics for prevalence = 50%**

True positives (sensitivity) =  $85/100 = 85.0\%$

False negatives =  $15/100 = 15.0\%$

True negatives (specificity) =  $80/100 = 80.0\%$

False positives =  $20/100 = 20.0\%$   
Overall correct classification =  $165/200 = 82.5\%$   
Gold Standard Criterion

Alcoholic Patients Normal Individuals

Present

*True positives*

**85**

*False positives*

**20**

New Test

Total Absent

*False negatives*

**15**

*True negatives*

**80**

**100 100**

Table 3.3 illustrates how percentages for the diagnostic efficiency statistics are calculated. In this example, it is assumed that there are 100 participants in each group and that the new test for alcohol problems correctly identifies 85 of the alcoholic patients as having alcohol problems and 80 of the normal individuals as not having alcohol problems. The true positive rate (sensitivity) is accordingly 85.0%, and the true negative rate (specificity) is 80.0%.

Much can be learned by observing the changes that occur in diagnostic efficiency statistics as the prevalence changes. In the present hypothetical example, a prevalence of 50% is created by assigning an equal number of persons to the alcohol patient and normal groups. Designing experiments to have equal sample sizes, and thereby artificially creating a 50% prevalence rate, has the advantage of increasing the power of the statistical techniques. In their focus on the usefulness of a new measure, test developers and practitioners may assume this 50% prevalence and overlook that the prevalence of alcoholism, and all other forms of psychopathology as well, varies widely with the type of setting in which it occurs. Accordingly, several different prevalence rates much lower than 50% need to be considered, because they are likely to affect some aspects of diagnostic efficiency.

To illustrate the potentially limiting effect of prevalence rates on diagnostic efficiency statistics, Table 3.4 shows these statistics for the hypothetical new test for alcohol problems when the prevalence is reduced to 15.0%, which is the previously noted lifetime prevalence for alcohol dependence (American Psychiatric Association, 2000). In this example, it is assumed that the sensitivity and specificity of the new test remain at 85% and 80%, respectively. The hit rate or overall correct classification rate has decreased slightly, from 82.5% to 81.0%.

Figure 3.1 presents the same information as Table 3.4, but uses a different conceptual layout, which many individuals find easier to understand the bases for the various calculations

#### **64 Basic Considerations**

##### **Table 3.4 Diagnostic efficiency statistics for prevalence = 15%**

True positives (sensitivity) =  $13/15 = 86.7\%$

False negatives =  $2/15 = 13.3\%$

True negatives (specificity) =  $68/85 = 80.0\%$

False positives =  $17/85 = 20.0\%$

Overall correct classification =  $81/100 = 81.0\%$

Gold Standard Criterion

Alcoholic Patients Normal Individuals

(cf. Gigerenzer, 2002). The results are identical in both Table 3.4 and Figure 3.1, so it does not make any difference which approach the clinician uses for laying out the data.

Assuming in this example that the sensitivity and specificity are unchanged is a typical assumption when an assessment test or technique is used with similar individuals in similar settings. A test validated to assess public speaking problems among college students in

a large state university counseling center would be expected to show similar sensitivity and specificity among students in another large state university, and in most other colleges and universities for that matter. However, a test of psychopathy validated in a maximum security penitentiary would be expected to show very different sensitivity and specificity in a university counseling center.

Although sensitivity and specificity remain the same in Table 3.4, the relative number of true positives to false positives in the two groups has changed drastically. In Table 3.3, where the prevalence is 50%, there are 85 true positives and 20 false positives, whereas in Table 3.4, where the prevalence is 15%, there are 13 true positives and 17 false positives.

100 People  
 Prevalence 15 Alcoholics 85 Normal  
 Test Result 13 Positive 2 Negative 17 Positive 68 Negative  
 True positives (sensitivity) =  $13/15 = 86.7\%$   
 False negatives =  $2/15 = 13.3\%$   
 True negatives (specificity) =  $68/85 = 80.0\%$   
 False positives =  $17/85 = 20.0\%$   
 Overall correct classification =  $81/100 = 81.0\%$

**Figure 3.1 Alternative conceptual layout for diagnostic efficiency statistics.**

**Psychometric Foundations of Assessment 65**

**Table 3.5 Calculating positive and negative predictive power**

Positive predictive power =  $\text{True positives} / (\text{True positives} + \text{False positives})$   
 Negative predictive power =  $\text{True negatives} / (\text{True negatives} + \text{False negatives})$   
 Using data from Table 3.3 where prevalence = **50%**  
 Positive predictive power =  $85 / (85 + 20) = 81.0\%$   
 Negative predictive power =  $80 / (80 + 15) = 84.2\%$   
 Using data from Table 3.4 where prevalence = **15%**  
 Positive predictive power =  $13 / (13 + 17) = 43.3\%$   
 Negative predictive power =  $68 / (68 + 2) = 97.1\%$

With the prevalence decreased from 50% to 15%, an individual hypothesized to have alcohol problems on the basis of the new alcohol test is actually more likely to be a normal individual. That is, the new test is incorrect in more instances than it is correct at a prevalence of 15%, even though the sensitivity and specificity of the test are unchanged and remain quite respectable.

Once an assessment test or technique has been validated, however, assessors should be less interested in sensitivity and specificity than in the ratio of true positives to false positives and true negatives to false negatives. The percentages derived from these ratios are reported respectively as *positive predictive power* and *negative predictive power*. As shown in Table 3.5, positive predictive power is the number of true positive scores divided by the total number of true positive and false positive scores. Using the data from Tables 3.3 and 3.4, it can be seen in Table 3.5 that the positive predictive power for the hypothetical test is 81.0% when the prevalence rate is 50% and 43.3% when the prevalence rate is 15%. Negative predictive power is the number of true negative scores divided by the total number of true negative and false negative scores. The negative predictive power is 84.2% when the prevalence rate is 50% and 97.1% when the prevalence rate is 15%.

The effects of varying the probability of false positive (10%, 20%, 25%) and false negative (10%, 20%, 25%) outcomes in conjunction with the change in the prevalence (2%, 5%, 10%, 20%, 25%, 50%) are summarized in Table 3.6. When the prevalence is 50%, there are significant improvements in the hit rate or overall classification rate over the prevalence even when the false positive and false negative percentages are 25%. However, as the prevalence decreases to 10%, 5%, and 2%, there is limited improvement in the hit rate or overall classification rate over the prevalence and in some cases there appears to be an actual decrease in the accuracy of classification.

These data demonstrate the difficulty of assessing characteristics or predicting behaviors that have a very low or very high prevalence in the setting in which people are being evaluated, particularly in finding an assessment method that is more accurate than simply identifying every person as not having the characteristic (when the prevalence is very low) or identifying every person as having the characteristic (when the prevalence is very high). On this basis, beginning with Meehl (1954), assessment specialists have consistently pointed out that tests work best when the prevalence rates for what is being measured hover

around 50%, become increasingly less efficient as prevalence diverges from 50%, and must be used cautiously in situations involving very low or very high prevalence.

As also shown in Table 3.6, clinicians should attend not only to the hit or overall classification rates, which can sometimes be misleading, but also to the positive (*PPP*) and negative (*NPP*) predictive power of whatever measures they use. The *NPP* is excellent at

**66 Basic Considerations**

**Table 3.6 Changes in PPP and NPP as a function of prevalence and percentage of false positives and false negatives**

Prevalence									
True		Positive		False		Negative		True	
Positive		False		True		Negative		False	
Positive	N	Hit Rate	PPP	NPP	Positive	N	Hit Rate	PPP	NPP
2	90	10	4,410	490	5,000	90.0	15.5	99.8	
2	90	10	3,920	980	5,000	80.2	8.4	99.7	
2	90	10	3,675	1,225	5,000	75.3	6.8	99.7	
2	80	20	4,410	490	5,000	89.8	14.0	99.5	
2	80	20	3,920	980	5,000	80.0	7.5	99.5	
2	80	20	3,675	1,225	5,000	75.1	6.1	99.5	
2	75	25	4,410	490	5,000	89.7	13.3	99.4	
2	75	25	3,920	980	5,000	79.9	7.1	99.4	
2	75	25	3,675	1,225	5,000	75.0	5.8	99.3	
5	90	10	1,710	190	2,000	90.0	32.1	99.4	
5	90	10	1,520	380	2,000	80.5	19.1	99.3	
5	90	10	1,425	475	2,000	75.8	15.9	99.3	
5	80	20	1,710	190	2,000	89.5	29.6	98.8	
5	80	20	1,520	380	2,000	80.0	17.4	98.7	
5	80	20	1,425	475	2,000	75.3	14.4	98.6	
5	75	25	1,710	190	2,000	89.3	28.3	98.6	
5	75	25	1,520	380	2,000	79.8	16.5	98.4	
5	75	25	1,425	475	2,000	75.0	13.6	98.3	
10	90	10	810	90	1,000	90.0	50.0	98.8	
10	90	10	720	180	1,000	81.0	33.3	98.6	
10	90	10	675	225	1,000	76.5	28.6	98.5	
10	80	20	810	90	1,000	89.0	47.1	97.6	
10	80	20	720	180	1,000	80.0	30.8	97.3	
10	80	20	675	225	1,000	75.5	26.2	97.1	
10	75	25	810	90	1,000	88.5	45.5	97.0	
10	75	25	720	180	1,000	79.5	29.4	96.6	
10	75	25	675	225	1,000	75.0	25.0	96.4	
20	90	10	360	40	500	90.0	69.2	97.3	
20	90	10	320	80	500	82.0	52.9	97.0	
20	90	10	300	100	500	78.0	47.4	96.8	
20	80	20	360	40	500	88.0	66.7	94.7	
20	80	20	320	80	500	80.0	50.0	94.1	
20	80	20	300	100	500	76.0	44.4	93.8	
20	75	25	360	40	500	87.0	65.2	93.5	
20	75	25	320	80	500	79.0	48.4	92.8	
20	75	25	300	100	500	75.0	42.9	92.3	
25	90	10	270	30	400	90.0	75.0	96.4	
25	90	10	240	60	400	82.5	60.0	96.0	
25	90	10	225	75	400	78.8	54.5	95.7	
25	80	20	270	30	400	87.5	72.7	93.1	
25	80	20	240	60	400	80.0	57.1	92.3	
25	80	20	225	75	400	76.3	51.6	91.8	
25	75	25	270	30	400	86.3	71.4	91.5	
25	75	25	240	60	400	78.8	55.6	90.6	

**Table 3.6 (Continued)**

Prevalence
True
Positive
False
Negative
True
Negative
False
Positive <i>N</i> Hit Rate <i>PPP</i> <i>NPP</i>
25 75 25 225 75 400 75.0 50.0 90.0
50 90 10 90 10 200 90.0 90.0 90.0
50 90 10 80 20 200 85.0 81.8 88.9
50 90 10 75 25 200 82.5 78.3 88.2
50 80 20 90 10 200 85.0 88.9 81.8
50 80 20 80 20 200 80.0 80.0 80.0
50 80 20 75 25 200 77.5 76.2 78.9
50 75 25 90 10 200 82.5 88.2 78.3
50 75 25 80 20 200 77.5 78.9 76.2
50 75 25 75 25 200 75.0 75.0 75.0

Note: *PPP* = Positive predictive power; *NPP* = Negative predictive power.

all prevalences, and at its worst is still accurate 75% of the time with a 50% prevalence that yields 25% false negatives and 25% false positives. The *PPP*, by contrast, is less than 50% accurate, that is, inaccurate more often than it is accurate, for all combinations of false positive and false negative values with all prevalence rates less than 20%, with one exception. In other words, any judgment that the behavior or symptom is present (*PPP*) when the prevalence is less than 20% will be inaccurate more often than it is accurate. The relatively small improvements in the hit rate or overall classification rate over the prevalence as the prevalence decreases to 10%, 5%, and 2%, and the generally very limited positive predictive power (*PPP*), raise the issue of the cost involved in making these judgments in low prevalence situations. It is possible that such small improvements in classification accuracy are not worth the cost in professional time and expense necessary to collect and evaluate assessment data. This is a complex question that will not be pursued in the present discussion. Clinicians must realize, however, that it is imperative to have some reasonable estimate of the prevalence of characteristics and behaviors that are being classified and the percentage of false positive and false negative outcomes that can be tolerated when any scale and its cutting score are used in a new setting. Even a small change in these psychometric features can have an appreciable effect on the accuracy of classification, as is evident in Table 3.6. Assessors should also investigate whether adjustments in cutting scores when scales are used in different settings might enhance the accuracy of classification; this is discussed in the next section.

Finally, with respect to prevalence, Wiggins (1973, p. 252) has pointed out that any test or predictor with a validity coefficient greater than .00 cannot result in a decreased proportion of correct decisions compared with judgments based on prevalence, the reason for this being that prediction from the prevalence rate sets the false positive rate at 0%. Hence, the negative result when the hit rates in Table 3.6 are compared with the prevalence assumes that both have the same percentage of false positive and false negatives, which is not correct. For example, if the prevalence for alcoholic dependence is 15% and false positives plus false negatives equal 25%, the hit rate is 75%, which is a 10% decrease from the prevalence (see last row of first section of Table 3.6). However, any classification made from the prevalence would set the false positive rate at 0% and the false negative rate at 15%, thereby accounting for the 10% decrease from the prevalence to the 75% hit rate.

**68 Basic Considerations**

The relative disadvantages of false positive and false negative outcomes are often more important to consider than the overall classification rate of a measure or scale, particularly in clinical evaluations. Prevalence rates predict all negative outcomes, so that the only errors are false negatives, whereas cutting scores on a measure of symptoms or problem behavior classify accurately some percentage of persons who have the symptom or problem

(true positives) and classify inaccurately some percentage of persons who do not (false positives). In most clinical settings, false negatives tend to be more disadvantageous than false positives. Failure to identify a disorder and provide treatment for a distressed person (false negative) is a more serious error than recommending treatment for a well-functioning person who does not need it (false positive), and failing to detect suicidal risk in people who are in fact poised to take their own lives is a more serious error than instituting suicide precautions for people who are in fact unlikely to harm themselves. Although such clinical realities must always be kept in clinicians' minds, they do not alter the primary point of the present discussion, which concerns the effects of prevalence and the percentage of false positives and negatives on the accuracy of classification. Articles by Meehl and Rosen (1955) and Streiner (2003) are recommended for additional reading about these and other matters related to diagnostic efficiency statistics.

### **Cutting Scores**

The final consideration to be discussed about clinical decision making is the derivation of appropriate cutting scores for particular scales in specific settings. The optimal cutting score can be determined fairly easily by plotting the frequency distribution of hits and misses on the same axes, as illustrated in Figure 3.2, because the optimal cutting score is the point

**Figure 3.2 Effects of different cutting scores on false positives and false negatives.**

*Psychometric Foundations of Assessment* 69

at which the two frequency distributions overlap. These examples assume that both groups are of equal size and have equal variances. Rorer and Dawes (1982) provide information on how to select the optimal cutting score when these assumptions are not met. These authors also show how an optimal cutting score can be estimated even when the prevalence in the local setting is not known. As can be seen in Figure 3.2a, the alcoholic persons who fall above this cutting score are true positives, and the normal individuals are false positives. The alcoholic persons who fall below this cutting score are false negatives, and the normal individuals are true negatives. Raising (Figure 3.2b) or lowering (Figure 3.2c) this cutting score changes the percentage of individuals classified in each category. It is thus possible not only to identify an optimal cutting score, but also to investigate the relative effects on false positives and false negatives of altering this cutting score.

Because the hit rate may not be as important as the relationship between the number of false positive and false negative outcomes in most clinical settings, clinicians should consider all these variables in selecting the most appropriate cutting score for a particular scale in a particular setting. A statistical test of the difference between the means of the two distributions provides much less information than the data concerning the prevalence of the behavior and the relative number of false positives and false negatives associated with a cutting score. As few as 100 patients in each group is sufficient to provide reasonable estimates of the prevalence of the behavior and the frequency of false positives and false negatives (Meehl, 1973), and a smaller number of patients can provide approximations of these variables while more data are being collected. This type of information is generally available in most applied settings, although most clinicians do not realize its importance. Consideration of base rates and optimal cutting scores also applies directly to the *DSM-IV-TR* (American Psychiatric Association, 2000). Finn (1982) showed that the polythetic criteria used to define many diagnoses such as Borderline Personality Disorder, as mentioned previously, can result in a changed prevalence rate and different rates of false positives and false negatives, depending on how many of the criteria are required to make the diagnosis. If the number of required criteria for diagnosing Borderline Personality Disorder were increased above five, the percentage of false positives would decrease; conversely, decreasing the number of required criteria below five would increase the frequency of false positives. This type of information on the changing rate of false positives and false negatives as the number of required criteria is altered does not appear in the *DSM-IV-TR*, and clinicians must look elsewhere for assistance in estimating how such alterations influence the prevalence of specific disorders in a particular setting.

### **IMPEDIMENTS TO ACCURATE DECISIONS**

In any setting in which assessment data are collected, there are likely to be several impediments to making accurate decisions, most of which reflect implicit theories that everyone

has about the causes and explanations of behavior. Prime examples of such implicit theories are social roles and stereotypes that people construct in their minds to ease and simplify their everyday interactions. Instead of learning what behaviors to expect from each person with whom they come in contact, people use social roles and stereotypes to make decisions about other individuals more rapidly than would otherwise be possible, and with less information. Being aware of social roles, as in expecting that store clerks will help you find a product you are looking for and will sell it to you, allows you to carry out the daily

#### **70 Basic Considerations**

requirements of life more quickly than having to deduce the expected conduct of each person you encounter. In similar fashion, stereotypes provide generalizations about certain types of people that simplify social interactions with them.

People typically collect information from their everyday experience in ways that support their implicit theories and produce numerous inherent biases. In this process, the apparent accuracy of social roles and stereotypes is commonly buttressed by propensities to note only information that confirms them and to ignore contradictory information. In addition, people tend to interpret ambiguous data that support whatever concepts of social roles or stereotypes they have formed.

It is easy to see how the processes that produce and buttress social roles and stereotypes can cloud the judgment of assessment psychologists and influence their conclusions in untoward ways. The paragraphs that follow describe some specific impediments to accurate decision making and discuss ways of minimizing their impact. Several general texts that explore this issue in depth are recommended for further reading (e.g., Dawes, 1996; Gilovich, Griffin, & Kahneman, 2002).

#### **Vividness (Saliency)**

All clinicians encounter persons whose symptoms or clinical presentation are so dramatic or vivid that they are easily recalled. These cases are frequently mentioned to colleagues and students to illustrate various aspects of the assessment and treatment process. The ease of recalling these individuals with vivid or salient characteristics often leads clinicians to overestimate how frequently these characteristics occur among the typical persons they evaluate in their office, agency, or institutional setting.

Accordingly, presentation of vivid case examples to illustrate psychological assessment issues should include information about their prevalence. Including this information helps ensure that those who read or hear about these dramatic cases have an accurate appreciation of the frequency with which they are likely to occur.

#### **Confirmatory Bias**

Once clinicians formulate a hypothesis about a person, they should strive to keep an open mind as they examine other data to determine whether they confirm or disconfirm the hypothesis. Confirmatory bias results when the data search attends only to data that support a preliminary judgment and ignore data that contradict it. Looking specifically for redundancy among test scores can exacerbate this problem, as can any preexisting biases a clinician may have concerning the implications of certain assessment findings. Along with confirmatory bias, clinicians may find themselves engaging in hindsight bias, which consists of recalling the data that supported their conclusions and forgetting data that pointed in other directions.

Confirmatory bias can also support illusory correlations, which involve seeing relationships between variables that are actually unrelated. Illusory correlations emerge when clinicians intuit some relationship between two variables and then look for information to confirm this relationship, instead of proceeding in the more appropriate opposite way by inferring relationships only on the basis of data that warrant them.

After clinicians have formulated data-based hypotheses about an individual they have examined, they should then be alert to any information that would disconfirm their

#### **Psychometric Foundations of Assessment 71**

hypotheses. This emphasis on looking for disconfirming information is necessary because of the ease with which clinicians can inadvertently interpret ambiguous data as confirming their hypotheses. Clinicians should also write down their conclusions, usually in the form of their reports, prior to receiving any explicit feedback on the accuracy of their judgments.

Proceeding in this way prevents assessors who have been mistaken (which all clinicians are at least once in a while) from reshaping what they meant to conclude and thereby continuing to believe that their judgment was accurate, without learning anything from their mistake.

### **Lack of Feedback on the Accuracy of Interpretations**

Feedback on the accuracy of the statements that assessment psychologists make in their reports may not always be available. Ideally, an assessment and the preparation of an assessment report are followed by some discussion with a referring person or agency concerning the extent to which the report provided clear and useful conclusions and was consistent with other known facts about the person who was examined. In addition, as discussed in Chapter 2, examiners may have opportunities for feedback sessions that verify or disprove interpretations they have made.

Often, however, feedback concerning the accuracy of their interpretations may not be available to assessors. They conduct their examination, construct their interpretations, prepare and submit their reports, and move on to the next case. In many agencies, institutions, and organizations—particularly when evaluations have been conducted for administrative rather than clinical purposes—it may be unusual for clinicians to discuss their reports with anyone who can provide useful feedback.

In the absence of feedback, and without being confronted with any contradictory information, clinicians can maintain belief in the accuracy of their judgments. As a more desirable process, discussing assessment results with referring parties and with people they have examined helps clinicians keep abreast of when they are being clear and accurate and when they are not, and by so doing improve the quality of their decision making and report writing.

### **Lack of Awareness of Relevant Variables**

Clinicians are often insufficiently aware of the variables or factors on which they are basing their decisions. When asked to make these factors explicit, they frequently emphasize variables that are unrelated to the behavior or symptom about which they have made a judgment. Even when they are familiar with the factors relevant to a particular judgment, they often do not know very much about the optimal weighting of these factors. Insufficient awareness among clinicians of the factors and their optimal weighting for determining whether a characteristic is present impedes the accuracy of their decision making and accounts for the previously mentioned superiority of statistical to clinical prediction when the necessary empirical data are available.

With this consideration in mind, it cannot be emphasized too strongly that assessment psychologists consulting in clinical settings should be familiar with the empirical literature concerning the specific types of psychopathology that are frequently encountered in their setting. This familiarity with the literature should include knowing which variables are most relevant for identifying each of these types of psychopathology and what if any

#### **72 Basic Considerations**

statistical procedures are available for combining these variables in some weighted fashion. Frequency counts of the presence and absence of variables thought to be relevant for each type of psychopathology can be developed with little difficulty in the clinician's particular setting when such data are not available in the empirical literature.

### **Task Complexity**

Clinicians' cognitive capacity and memory functions can easily be overwhelmed by the sheer amount of data that must be processed in conducting a psychological evaluation. This complexity and the challenge it poses to accurate decision making are likely to increase if examiners do not take extensive notes on an examinee's behavior during the assessment process. Audiotaping or videotaping an assessment procedure is an efficient way to capture all the information provided by people's behavior during testing as well as by their test responses. These tapes can then be consulted as necessary to assist in formulating interpretations and preparing a written report.

Using computerized interpretations of assessment measures that are available is another means of decreasing examiners' reliance on their memory. Computer-based test interpretation is discussed in Chapter 2 of this *Handbook*.

### **Incomplete Information**

The accuracy of decision making can also be impeded when clinicians lack important bits of information. The nature of most settings in which assessment psychologists consult is such that some types of information are routinely not available. Incomplete or missing information in particular can result in clinicians failing to appreciate fully the importance of certain variables and their prevalence. To continue with the example of an alcohol treatment program, clinicians working in such a setting can readily come to the conclusion that no one can recover from alcoholism without participating in a formal treatment program. This belief is instilled by working with the people they encounter in their treatment program, who in fact did not recover without treatment.

What clinicians working in alcohol treatment programs do not acquire, however, and which would disconfirm their belief, is information about individuals with alcohol problems who find ways of recovering on their own without being treated in a formal program. Generally speaking, statements about alcoholism that are generated within an alcohol treatment program cannot be considered valid for all persons who are alcohol dependent, because the subset of persons being sampled comprises only alcoholics who are receiving formal treatment. Incomplete information about any factor that has a substantial bearing on drawing a particular conclusion can contribute to clinicians forming mistaken beliefs and judgments and making inaccurate decisions.

### **CONCLUDING COMMENTS**

Clinicians who conduct psychological evaluations should not feel overwhelmed by the considerable information presented in this chapter or dejected by the many troublesome issues that are raised. The information provides a basic primer in psychometrics and **Psychometric Foundations of Assessment 73**

should, for the most part, be familiar to readers who received instruction in statistics and measurement courses prior to beginning their study of personality assessment methods. As for the troublesome issues, which are all too often ignored in many applied settings, they will not go away and must constantly be kept in clinicians' minds as they conduct psychological evaluations. At least one potential solution is provided for each of these issues in the course of the chapter, and there are also four general considerations for examiners to keep in mind that will serve them well in carrying out their clinical decision-making responsibilities.

First, whatever the type of setting in which clinicians are providing diagnostic consultation, they need to be well informed about the prevalence of the psychological characteristics and disorders that are seen in this setting and about the most salient behaviors or symptoms that differentiate among them. Being well informed requires knowing the extant empirical literature and keeping up to date with changes that occur in the field.

Second, attempting to understand and describe the nature of people, their personality assets and limitations, their adjustment problems and disorders, if any, is a complex task that cannot be carried out quickly. Despite the popularity of Dr. Phil, Ann Landers, and Oprah, no clinical assessment can be made in a few minutes of discussion or based on a few sentences describing the problem.

Third, personality assessment is a probabilistic rather than a causal endeavor. No behavior or symptom has just a single cause, and the frequency of its occurrence depends on multiple variables. Once it is truly understood that any statement about an individual is a probabilistic statement, not a fact, both clinicians and the individuals they assess are less likely to be misled.

Finally, clinicians cannot rely on their memory to assimilate and store all the complex information available about the prevalence of and relationships among all of the behaviors or symptoms in all forms of psychopathology. Regular recourse to appropriate books and journals and increased reliance on computers to store and process relevant information are keys to maintaining and enhancing personality assessments.

### **REFERENCES**

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical judgment. *Counseling Psychologist, 34*, 341–382.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Barlow, D. H., & Campbell, L. A. (2000). Mixed anxiety-depression and its implications for models of mood and anxiety disorders. *Comprehensive Psychiatry*, *41*, 55–60.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.
- Caldwell, A. B. (2006). Maximal measurement or meaningful measurement: The interpretive challenges of the MMPI-2 Restructured Clinical (RC) Scales. *Journal of Personality Assessment*, *87*, 193–201.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.
- 74 Basic Considerations**
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- Dawes, R. (1996). *House of cards: Psychology and psychotherapy built on myth*. New York: Free Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Exner, J. E., Jr. (2003). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations and principles of interpretation* (4th ed.). Hoboken, NJ: Wiley.
- Finn, S. E. (1982). Base rates, utilities, and DSM-III: Shortcomings of fixed-rule systems of psychodiagnosis. *Journal of Abnormal Psychology*, *91*, 294–302.
- Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.
- Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment*, *15*, 508–520.
- Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.
- Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.
- Graham, J. R., Smith, R. L., & Schwartz, G. F. (1986). Stability of MMPI configurations for psychiatric inpatients. *Journal of Consulting and Clinical Psychology*, *54*, 375–380.
- Greene, R. L., Davis, L. J., Jr., & Morse, R. P. (1993, March). *Stability of MMPI code types in alcoholic inpatients*. Paper presented at the midwinter meeting of the Society for Personality Assessment, San Francisco.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical predication: A meta-analysis. *Psychological Assessment*, *12*, 19–30.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Belmont, CA: Wadsworth/Thompson Learning.
- Livingston, R. B., Jennings, E., Colotla, V. A., Reynolds, C. A., & Shercliffe, R. J. (2006). MMPI-2 code-type congruence of injured workers. *Psychological Assessment*, *18*, 126–130.
- Meehl, P. E. (1954). *Clinical versus statistical judgment*. Minneapolis: University of Minnesota Press.
- Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225–302). Minneapolis: University of Minnesota Press.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–216.
- Monahan, J., Steadman, H. J., Robbins, P. C., Appelbaum, P., Banks, S., Grisso, T., et al. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services*, *56*, 810–815.
- Morey, L. C. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Nichols, D. S. (2006). The trials of separating bath water from baby: A review and critique of the MMPI-2 Restructured Clinical Scales. *Journal of Personality Assessment*, *87*, 121–138.
- Packman, W. L., Marlitt, R. E., Bongar, B., & Pennuto, T. O. (2004). A comprehensive and concise assessment of suicide risk. *Behavioral Sciences and the Law*, *22*, 667–680.
- Rorer, L. G., & Dawes, R. M. (1982). A base-rate bootstrap. *Journal of Consulting and Clinical Psychology*, *50*, 419–425.

# Assessment of Childhood Anxiety

KELLY ROBINSON

Historically, anxiety has been conceptualized as a circumstance of adulthood. Children were not thought to experience anxiety in any unique way, and scientific studies of anxiety disorders have, until recently, focused on adults. However, while the majority of theories about anxiety have focused on symptoms in adults, they have been based on limited empirical evidence in general and have been born of many diverse psychological perspectives. Yet, most theorists agree that there is not a single cause but multiple factors interacting with each other which lead to the development of anxiety disorders. While the quest for genetic links, family characteristics, environmental influences, and proximal antecedents that cause anxiety continues, it is likely that all of these elements act on and shape each other to form each person's individual anxiety experience. Therefore, the process of assessing anxiety must examine many aspects of an individual's functioning, such as physiological arousal, cognitions, behaviors, and subjective interpretations. The anxiety itself must also be assessed by measuring its severity, duration, and pervasiveness in the individual's life (Herbert, 1994).

## THEORIES OF ANXIETY

Of the many models developed to explain anxiety, three seem to predominate. They include behavioral, cognitive, and biological models. Behavioral models are primarily based on classical and operant conditioning. According to classical conditioning, anxiety-provoking stimuli elicit physical sensations that alert the individual to the presence of danger (i.e., the feared stimuli). Pairing a neutral stimulus with the anxietyprovoking stimulus causes the individual to associate the physical symptoms of anxiety with the neutral stimulus. Once the individual has made this association, exposure to the neutral stimulus alone results in anxiety. This classical conditioning model coupled with reinforcement, such as what occurs when anxiety is reduced through avoidance, may further increase the likelihood that the anxiety response will persist. While the behavioral model has been used to explain the creation of phobias, clearly the model does not account for the vast majority

of cases of anxiety. Even many phobic individuals cannot identify instances in which neutral stimuli were paired with  
508

# 23

## Assessment of Childhood Anxiety

KELLY ROBINSON

feared objects or situations (Kearney & Wadiak, 1999).

Cognitive theories of anxiety explain symptoms by identifying the normal processes by which children assimilate information in the world. Newly acquired information is placed within the cognitive structures of previously acquired information.

People compare the new information with the old in order to make predictions about their environment. When individuals learn to expect negative consequences from situations they perceive as aversive or dangerous, they avoid those situations. Thus, when those aversive situations are unavoidable, and individuals must endure encounters that they perceive are threatening, they experience feelings of anxiety (Kearney & Wadiak, 1999).

Biological theories of anxiety focus on genetic studies. Though a single causal gene has not been found, results of several studies suggest that there are genetic links to anxiety disorders, particularly panic disorder, obsessive-compulsive disorder, and certain types of phobias (Kearney & Wadiak, 1999).

### CHILDREN'S ANXIETY

Children who experience pathological levels of anxiety represent a category of individuals for whom assessment and diagnosis becomes particularly challenging. While it is still important to assess all the major areas of a child's life that were discussed previously (proximal antecedents, family characteristics, physiological functioning, etc.), one must also bear in mind the developmental appropriateness of anxiety during stages of a child's life. In other words, a certain amount of anxiety is healthy, so it is important to determine if anxiety is present and if it is problematic for the child.

Anxiety is a normal and developmentally appropriate response that allows humans to adapt to their environment. From an evolutionary

perspective it allows for survival in the “fight or flight” action a person takes when faced with danger, and in a performance capacity a limited amount of anxiety helps a person function at his or her best. Children progress through stages of normal anxiety throughout their lifespan. Therefore, it is important to be knowledgeable about these stages so as not to confuse normal developmental anxiety from abnormal anxiety that leads to dysfunction. Children tend to progress through the developmental anxiety trajectory beginning in infancy with fear of sudden noises and of being dropped or startled. At age 1 to 2, children begin feeling anxious when not in close proximity to caregivers. Then fear of strangers emerges, and when children reach the preschool years, fear of monsters, animals, and the dark are common. Finally, in early to middle childhood, children may develop specific fears of objects or situations, such as school, physical danger, and supernatural phenomena (Beidel & Stanley, 1993). Although these anxieties represent normal developmental stages, abnormal or pathological anxiety also follow these stages, with pathological anxiety differing in degree. Evidence that anxiety is a common part of normal development is suggested by studies of subclinical anxiety symptoms. These studies show that individuals who have identifiable anxiety symptoms but who do not meet diagnostic criteria for an anxiety disorder and individuals who are asymptomatic have similar functional outcomes. Qualitatively, clinical levels of anxiety may appear similar to normal developmental anxiety. However, pathological anxiety differs in that it is generally inappropriate to the situation, involuntary, irrational, and functionally limiting (Anderson, 1994). Investigation into several areas may be useful in determining whether anxiety is at a level that may be considered problematic or pathological. First, context is important. Is the anxiety out of proportion to the feared stimulus, situation, or event? Given that children experience a progression of developmentally typical fears, the context may be wider than for adults (Herbert, 1994). However, it is still possible to assess the degree of the anxiety and its age appropriateness. It is also prudent to examine the frequency and intensity of the child’s anxiety. For example,

is the child's level of anxiety within a range that is present in most other children within the same age range? Also, does the anxiety occur with more frequency and in regard to a wider range of stimuli than in most other children of the same age? Whereas constant low levels of anxiety can be just as maladaptive as high levels of anxiety, examining the frequency and intensity should provide a clearer picture of how a child's anxiety impacts his or her functioning.

The intensity of children's anxiety often varies relative to situations, times, places, or people with whom the child engages. Therefore, scrutinizing the child's environment and also the child's perspective of his or her own anxiety is important. How the child makes sense of his or her anxiety may provide insight into any purpose that the anxiety may serve. Whereas conducting a functional analysis of antecedents and consequences regarding the types of problems the child experiences due to anxiety, it will not provide information about how the child views the anxiety as a problem. Attempting to see anxiety from the perspective of the child may also help in developing treatment goals and strategies (Herbert, 1994).

### **ANXIETY DISORDERS**

Many anxiety problems go undetected due to the difficulty that school personnel and parents have in identifying the disorder, as anxiety frequently results in only relatively mild behavioral problems. Furthermore, anxiety disorders in children have a high comorbidity rate with other disorders, such as depression and attention-deficit/hyperactivity disorder, which further obscures the presentation of symptoms and makes identification more difficult. This difficulty in identifying anxiety disorders in children is an important issue in regard to assessment, as most anxiety problems begin in childhood and progress to a chronic degree into adulthood, where they may have a profound and detrimental impact on the lives of the individuals affected (Albano, Chorpita, & Barlow, 1996). According to the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DMS-IV; American Psychiatric Association, 1994), there are nine anxiety disorders with which a child may be diagnosed, including separation anxiety disorder (SAD), specific phobia, social phobia,

generalized anxiety disorder (GAD) (previously referred to in the literature as overanxious disorder), panic disorder with or without agoraphobia, obsessive–compulsive disorder (OCD), posttraumatic stress disorder (PTSD), and acute stress disorder. SAD is the only disorder that is unique to children. Each disorder shares common features but may be differentially diagnosed based on the primary focus of the child's anxiety.

The prevalence of anxiety disorders in children and adolescents ranges from approximately 9 to 27% (Albano et al., 1996).

However, there is wide variability in the reporting of prevalence rates, which is influenced by a number of factors such as diagnostic and referral processes, likelihood of participation in studies, use of general populations versus clinical samples, relationship of reporter of symptoms, and age of population.

Differences in the reports of prevalence rates aside, anxiety disorders are the most common psychiatric disorder diagnosable in children and adolescents (Bell-Dolan & Brazeal, 1993). Anxiety disorders seem to increase in adolescence with phobic symptoms more prevalent in girls than in boys, with the exception of public-speaking phobias. The symptoms of anxiety peak earlier in girls than in boys, but boys lose specific fears more quickly than do girls, usually around age 10. Less gender variability is noted in younger children. The most common disorders comorbid with anxiety are depression and dysthymia, ADHD, conduct and oppositional disorders, and other specific anxiety disorders. There is little evidence that GAD, phobias, and SAD continue into adult life. Recovery rates for these anxiety disorders range from 60 to 75%, and those that do progress into adulthood may not continue with the same pattern of symptoms. OCD, on the other hand, has shown considerable stability into adulthood. Boys tend to demonstrate a younger age of onset, and the disorder also seems to be more prevalent among boys. While most individuals with OCD report no previous history of psychopathology, the disorder is sometimes comorbid with depression, other anxiety disorders, and Tourette syndrome (Anderson, 1994).

### **Separation Anxiety Disorder**

A child may be diagnosed with SAD if he or she exhibits developmentally inappropriate

and excessive anxiety in regard to separation from home or an attached figure. SAD is diagnosed when the child exhibits three or more symptoms, such as worry about harm coming to attached figures, school refusal, or reluctance to go to sleep without an attached figure present. The symptoms must persist for at least 4 weeks, and onset must be before age 18 (American Psychiatric Association, 1994). The prevalence rate for children with SAD ranges from 3 to 5% in children, and may be a little less in adolescents. One-third of children with SAD have comorbid GAD secondary to SAD, and another third are diagnosed with a comorbid depressive disorder (Last, Strauss, & Francis, 1987).

The focus of SAD is fear associated with separation from home or an attached figure. It is important to remember when assessing for SAD that anxiety about being away from home and family members is a normal part of development from the age of approximately 7 months to 6 years (Bernstein & Borchardt, 1991). However, children with SAD differ from typical children in that they often experience fear of losing loved ones through catastrophic events. Children with SAD tend to demonstrate extreme avoidance behaviors and complain of somatic symptoms, including panic attacks. A key feature of the disorder that distinguishes children with SAD from children experiencing normal developmental anxiety is that SAD interferes with a child's daily functioning, such as friendships, participation in groups, and academics.

### **Specific Phobias**

Specific phobias are characterized by marked and persistent fear or anxiety when exposed to, or when anticipating exposure to, a particular object or situation. Exposure to the feared stimulus results in an immediate anxiety response, which may include a panic attack. In children the anxiety response may take the form of crying, tantrums, freezing, or clinging. Adults must recognize the fear as excessive or unreasonable in order for a diagnosis to be made, but this recognition is not required for children (American Psychiatric Association, 1994). The most important factor in differentially diagnosing specific phobias in children is identifying true phobias versus normal developmental fears. Common childhood fears include heights, darkness, thunder, injections,

insects, and dogs (Albano et al., 1996). One feature that distinguishes phobias from typical fears is the degree to which the fear is maladaptive. School phobia is a unique specific phobia because school phobia may look similar to SAD, social phobia, or truancy. While truancy may be an indication of oppositional behavior, the school phobic child's avoidance of school is involuntary. Anxiety about going to school may not even be in the child's conscious awareness, as symptoms frequently manifest through somatic complaints before or during the school day. The fear is typically associated with specific occurrences at school rather than with separation from home or attached figures (Blagg & Yule, 1994). To diagnose a specific phobia of school, social phobia, which centers around a fear of embarrassment or humiliation, must also be ruled out as a possible diagnosis. School phobia tends to surface shortly after beginning school for the first time (ages 5–6), following a transition from elementary to middle/high school (ages 11–13), and after age 14. However, school phobias experienced by older children and adolescents may be suggestive of a more severe developing pathology (Blagg & Yule, 1994).

### **Social Phobia**

Social phobia is characterized by marked and persistent fear in social and performance situations. In children the anxiety must be present in peer situations, not just during interactions with adults, and children must demonstrate the capacity to form age-appropriate relationships with others. Exposure to feared situations provokes an anxiety response, which may include panic attacks, and in children, may also include crying, tantrums, and shying away from social situations. Adults must recognize that the fear is unreasonable or excessive in order to be diagnosed, whereas in children this criterion does not have to be met. Symptoms must be present for at least 6 months in order for social phobia to be diagnosed in individuals under the age of 18 (American Psychiatric Association, 1994).

23. Assessment of Childhood Anxiety 511

Social phobia is rarely diagnosable under the age of 10, and has been diagnosed equally in males and females (Albano et al., 1996). Unlike SAD, social phobia is not a disorder specific to children. Therefore,

much of the research examining the disorder has been taken from the adult literature. The primary focus of fear in social phobia is fear of embarrassment. Children and adults with the disorder become excessively concerned about rejection, humiliation, and negative evaluation. Not surprisingly, children with social phobia tend to be shy and have few friends. Cognitions are overwhelmingly negative and self-deprecating (Albano et al., 1996). Symptoms of social phobia may include somatic complaints and obsessing about visible signs of embarrassment, such as blushing, shaking, and sweating. Although anyone in embarrassing situations may have these same worries, the difference between children with social phobia and typical children is the degree of worry. However, an apparent ability to perform during these social activities does not necessarily mean that the disorder is absent. Although social phobic individuals attempt to avoid situations in which they may become embarrassed, they will often endure these situations rather than draw attention to themselves. Despite an apparent ability to perform, they may experience an extreme degree of internal distress during these times (Albano et al., 1996). School refusal is not an uncommon feature of the disorder, and one must be careful to differentially diagnose social phobia, school phobia, and SAD by keeping in mind that the anxiety focus of social phobia involves fear of embarrassment in social situations rather than fear of more specific occurrences at school, or a separation from home.

### **Generalized Anxiety Disorder**

GAD is diagnosed when an individual experiences excessive worry or anxiety more days than not, about a number of situations, and the worry is difficult to control.

Anxiety symptoms may include restlessness, fatigue, difficulty concentrating, irritability, muscle tension, and disturbances in sleep.

Children must demonstrate only one of these symptoms, whereas adults must experience three or more symptoms in order for a diagnosis to be made (American Psychiatric Association, 1994). Although GAD may begin any time during the school-age years, the mean age of onset appears to be between age 10 and 13 (Last et al., 1987). GAD is highly comorbid with depressive disorder and other anxiety disorders.

The presentation of GAD in children appears similar to that of adults except that children are more likely than adults to have somatic complaints. According to DSM-IV, a diagnosis of GAD is warranted in a child only if at least one physiological symptom is present. The focus of GAD is excessive worry about a number of life events, which may include past behavior, competence in sports, academics, peer relationships, or future events. Frequently children with GAD demonstrate perfectionist approaches in performing tasks and set exceedingly high standards for themselves. They are also quite hard on themselves when they do not achieve their goals or when they perceive that they have failed in some way (Albano et al., 1996).

### **Panic Disorder**

Panic disorder is marked by recurrent and unexpected panic attacks followed by 1 month or more of worry about the future occurrence of attacks or the implications of attacks. Panic disorder may also include agoraphobia, which is anxiety about being in situations in which escape would be difficult or help unavailable if a panic attack should occur, such as in public places. Individuals with agoraphobia avoid these situations, likely resorting to extreme behavior in order to do so (e.g, refusing to leave the house) (American Psychiatric Association, 1994).

Panic disorder has gone largely undiagnosed in children until fairly recently, and as such there is a paucity of literature in regard to specific epidemiological information such as age of onset and gender predominance. However, the disorder appears to be more common in females than in males, and there may be a link between pubertal onset and panic attacks (Kearney & Allan, 1995). Cognitive and language limitations may make it difficult for children to adequately describe the physical sensations

### 512 VII. ADVANCED TOPICS

that accompany panic attacks, making the disorder difficult to diagnose. Furthermore, children often lack the specific detail that adults use in describing these sensations. For example, children frequently use vague terms to describe panic attacks. They may express a fear of getting sick, whereas adults tend to describe fears of dying, going crazy, or losing control. As children reach adolescence, their fears become more specific,

as they can describe feeling a sense of breathlessness, rapid heartbeat, and dizziness and feeling as if they are not in their own body, and some describe a fear of dying. Differential diagnostic concerns are related to the similarity in appearance between panic disorder and separation anxiety disorder, as many children and adolescents cope with their panic attacks by staying close to a family member or friend who makes them feel safe. The focus of the anxiety should be used to differentiate between the diagnoses (Albano, Chorpita, & Barlow, 1996).

### **Obsessive–Compulsive Disorder**

For a diagnosis of OCD to be made, a person must experience either obsessions or compulsions. Adults must realize that the obsessions or compulsions are excessive or unreasonable in order to be diagnosed, but children do not have to meet this criteria. The obsessions or compulsions must consume 1 hour or more per day, cause the individual extreme distress, and interfere with normal functioning (American Psychiatric Association, 1994). OCD is commonly comorbid with depression, Tourette syndrome, and other anxiety disorders (Albano et al., 1996).

The expression of OCD in children is similar to that of adults. However, the content of children's obsessions and compulsions most often changes over time (Albano et al., 1996). Common themes include contamination, sexual themes, religiosity, aggressive or violent images, recurring songs or rhymes, and fear of illness. Children's compulsive behavior tends to involve handwashing, checking, and arranging items in particular orders, with washing being the most common ritual (Last & Strauss, 1989). However, children frequently engage in developmentally appropriate rituals that are not compulsions, such as arranging toys or dolls and saying goodnight to family members. These rituals do not seem excessive and are typically different in content than compulsions. Typical childhood rituals tend to dissipate at about 10 years of age and are not paired with obsessive thoughts (Swedo, Rapoport, Leonard, Lenane, & Cheslow, 1989). Again, the degree of distress a child experiences, particularly when unable to perform the ritual, is a useful indicator of pathology (Albano et al., 1996).

### **Posttraumatic Stress Disorder**

For an individual to be diagnosed with PTSD, he or she must have been exposed to a traumatic event, during which a serious threat to the safety of self or others was present. The individual's response to the experience must have been one of fear, helplessness, or horror or in children, disorganized or agitated behavior. Reexperiencing of the event must also take place and may take the form of repetitive play, dreams, flashbacks, distress when exposed to reminders of the event, and physiological symptoms upon exposure to internal or external cues of the event. A number of avoidance or numbing behaviors and increased arousal must also be identified, and the duration of the symptoms must be at least 1 month (American Psychiatric Association, 1994). The symptoms of PTSD usually, but not always, arise within 3 months of the traumatic event. Bad dreams are particularly common in young children and may take the form of general nightmares about monsters, etc., rather than about the traumatic event. Children do not tend to reexperience the event in the same way that adults do, instead reliving the event through play. Children may not be able to report many of their symptoms, so observations by parents and teachers are crucial to evaluate hopelessness and loss of interest in previously enjoyed activities. Children may demonstrate a sense of foreshortened future by believing that they will not be around to perform adult activities. Physical symptoms, such as aches and pains, irritability, and angry outbursts, are also common features. PTSD may also be linked to the development of other anxiety disorders (American Psychiatric Association, 1994).

### 23. Assessment of Childhood Anxiety 513

#### **Acute Stress Disorder**

Acute stress disorder is diagnosed when a person has been exposed to a traumatic event in which physical harm was threatened and the individual's response included fear, hopelessness, or horror and is followed by several dissociative symptoms. The symptoms last at least 2 days but do not linger beyond 4 weeks. A primary feature of the disorder is a lack of emotional responsiveness and a loss of ability to experience pleasure. The person with acute stress disorder may experience amnesia related to the traumatic event (American Psychiatric Association, 1994).

## DIAGNOSTIC PROCEDURES

Each of the anxiety disorders just described is based on a categorical diagnostic system, DSM-IV. DSM provides discrete categories based on a symptom checklist. If a person has the requisite symptoms listed in the manual, a disorder can be diagnosed. No consideration is given to individuals who fall short of the criteria by only one symptom. Although the criteria are general and allow for some variation in symptoms, individual differences that cause the presentation of the disorder to look different from one person to the next are not included in DSM descriptions. Another drawback of DSM and other categorical systems of diagnosis is that neither the singular presentation of childhood symptoms nor developmental issues are given special consideration. The criteria for adults and children are virtually identical. Due to these factors, overlapping symptoms across diagnoses and the heterogeneity of symptom manifestation, the usefulness of categorical systems for diagnosing children has been called into question.

Dimensional systems, developed by clustering symptoms that commonly occur together, are used with increasing frequency. Rating scales are the primary means of measurement in dimensional systems, as they produce profiles of an individual's symptoms across a broad spectrum. For example, a child may score high on one scale (e.g., anxiety) and low on another scale (e.g., aggression), providing a more complete picture of the child's functioning. In addition to allowing for individual differences in symptom expression, developmental variables may be accounted for in dimensional systems (Albano et al., 1996).

Despite dimensional systems' growing popularity, DSM is the most widely used and recognized classification system among mental health professionals and allows for ease of discussion about general psychiatric disorders. Therefore, future diagnostic systems appear to be moving toward a combined approach that ties together assessment and treatment goals (Albano et al., 1996).

### Rating Scales

Rating scales are used for a variety of purposes, such as to aid in diagnosis, screen for possible pathology, gather information about behavior across settings, uncover personality traits, and measure treatment progress. Rating scales are also used by

many types of professionals in schools, hospitals, and clinical settings. To determine which scale to use in a particular situation, it is important to consider the goal of obtaining information and the setting in which the information will be used. Self-report rating scales are the most common method used to assess anxiety and depression in children in clinical and research settings (Silverman & Rabian, 1999). Whereas children and adolescents are thought to be the most accurate reporters of their own feelings and internal states, parents and teachers are better recorders of observable behaviors (Silverman & Rabian, 1999).

Rating scales are popular because they are inexpensive and easy to use. One of their many purposes is to screen and/or diagnose groups of children, in which case the instrument must differentiate children with clinical levels of anxiety from those who are asymptomatic. One factor that may affect the validity of rating scales is the tendency for individuals to answer items in such a way as to make themselves appear in the most positive light. To limit this effect, administration of directions should be carefully worded to encourage honest responding, such as telling children that there are no right or wrong answers. When screening children using rating scales, the clinician must be mindful that a scale may reveal more false-positive responses than true-

#### 514 VII. ADVANCED TOPICS

positive responses, meaning that children may appear more anxious than they really are (Costello & Angold, 1988). Therefore, once a rating scale has identified a child as having clinical levels of anxiety, further investigation into the child's cognitive and socioemotional functioning, by way of interviews or other diagnostic procedures, must be conducted to support or disconfirm the results of the rating scale.

Diagnosing anxiety disorders should be accomplished through diagnostic interviews, with rating scales used as a means of obtaining severity of symptoms. Rating scales may be used to identify and quantify symptoms and behaviors once the presence of an anxiety disorder has been established (Silverman & Rabian, 1999).

A potential problem with using rating scales to measure anxiety is that rating scales do a relatively poor job of differentiating anxiety and depression, which is further

complicated by the disorders' high comorbidity rates. However, it may be beneficial to look for patterns of responses that point more to worry than to depressive symptoms, such as loss of interest in pleasurable activities, lack of motivation, and low self-esteem.

Behaviors that mediate anxiety may be identified through rating scales. For example, a child's responses may indicate that he or she avoids situations that lead to anxiety. It may also be possible to use rating scales to assess other variables that are secondarily affected by anxiety, such as peer or familial relationships (Silverman & Rabian, 1999). However, the validity of using rating scales to assess treatment progress is questionable, as improved scores may not represent a reduction in anxiety but rather reduced reporting of symptoms, or normal fluctuations in anxious thoughts and feelings (Silverman & Rabian, 1999).

### **Interviews**

The most common method used to assess child psychopathology is the interview. There are structured, semistructured, and unstructured interviews with structured interviews requiring the least training and diagnostic experience to administer. In most cases, interviews are designed to detect symptoms consistent with DSM criteria. Many interviews include both child and parent versions. Interviews are crucial in diagnosing anxiety disorders because the more information a clinician has to make a diagnostic decision, the more reliable the decision should be. However, inherent problems with interviewing children can lower the reliability of diagnosis. Perhaps the biggest problem with using interviews to assess children is the reliance of the interviews on DSM structure. As stated previously, DSM criteria do not adequately detect developmental differences in the expression of symptoms, and most DSM criteria are based on adult symptomatology. Consequently, wide variation in children's answers to interview questions may occur, which clinicians may interpret differently (Kearney & Wadiak, 1999). Another DSM-related issue affecting the reliability of interviews is that child anxiety disorders and their criteria change with each new publication of DSM (Kearney & Wadiak, 1999). Furthermore, diagnosing specific types of anxiety disorders is less reliable than diagnosing the presence

of an anxiety disorder, generally.

### **Behavioral Observations**

Behavioral observations are critical in the assessment of anxiety for several reasons.

First, assessment through the use of rating scales and interviews is limited by the desire of respondents to appear in a positive light or to please the administrator. In addition, anxiety symptoms are often expressed differently across settings, with some symptoms only appearing in certain environments or situations. A critical component of diagnosing anxiety disorders is determining the pervasiveness and severity of symptoms. Therefore, functional behavioral analyses should be conducted and antecedent and consequent events to the anxiety explored. Behavioral observations may also be helpful in developing appropriate and practical treatment goals for the anxious child.

In addition to functional behavior analyses, behavioral observations may take the form of behavioral approach tests (BATs), observational ratings, role-play tests, and self-monitoring. BATs are more frequently employed when specific phobia, social phobia, or generalized anxiety is suspected. A typical BAT includes brief stages, during which the child is exposed to a feared stimulus, and response measurements are taken.

An example of a five-stage BAT includes an initial adaptation stage, during which the child is placed in an environment and permitted time to adapt to the setting. Next, during the baseline stage, the clinician records the child's level of fear (e.g., by measuring heart rate), then the same procedure is conducted while the child is walking and moving during the walking baseline stage. The child is then exposed to a feared stimulus for a brief period, or until the fear subsides. Finally, the child is placed in the original setting with the feared stimulus absent, and the length of time to reach baseline levels is measured. The stages of a BAT may be conducted in simulated or natural settings. Natural settings are preferred, although external validity in either setting has not been well established.

Observational rating systems primarily measure motor behaviors that are consistent with anxious responding, such as verbalizations, trembling, avoiding, poor eye contact, and body rigidity. The Behavior Assessment System for Children (BASC)—Student Observation

System is one example of an observational rating system. However, further research regarding reliability and validity of these measures is required.

Role-play tests are typically used with children diagnosed with social phobia and are administered by asking the child to respond during role plays as if they were in anxiety arousing situations. The test may include videotaping so that the child's responses may be analyzed. However, reliability and validity of role-playing tests have not been demonstrated.

Self-monitoring requires the child to record his or her own anxious responses by keeping diaries of antecedents, anxiety-producing events, and consequences or by recording thoughts that are associated with the anxiety. However, self-monitoring approaches may be limited by the child's language and cognitive capabilities and his or her willingness to comply with the procedure (Kearney & Wadiak, 1999).

#### **Physiological Assessment**

Due to the increase of physiological arousal in individuals with anxiety disorders, physiological assessment may be a particularly informative and useful tool. This relatively new area of study involves the measurement of bodily responses and vital functions in the assessment of anxiety, such as measuring heart rate, blood pressure, respiration, blood volume, skin temperature, and electrical activity in tense muscles (Kearney & Wadiak, 1999). However, to date, methodology regarding the use of physiological assessment of anxiety has not been well established.

Normative, reliability, and validity data on physiological assessment have not been established.

#### **INSTRUMENTS**

The following section discusses a number of instruments commonly used to measure symptoms of anxiety. Although not a comprehensive list, this section provides examples of general behavior rating scales, anxiety scales, and interviews. The instruments listed vary with respect to format and respondent, and psychometric properties of each are reviewed.

#### **Behavior Assessment System for Children**

The Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) is a comprehensive diagnostic system for use with 4- to 18-year-olds, which includes a Structured Developmental History, Student

Observation System, Teacher Rating Scales (TRS), Parent Rating Scales (PRS), and Self-Reports of Personality (SRP). The SDH and SOS do not have norms and cannot be readily integrated with information gathered from the rating scales but are an important inclusion in making differential diagnosis of emotional and behavioral disorders based on multiple sources of information. The TRS, PRS, and SRP provide scores based on national norms, gender within the norm group, or in comparison to a group of seriously emotionally disturbed children. In addition to clinical scales and composite scores for Internalizing and Externalizing behaviors, an Adaptive Scale is provided. The Adaptive Scale was constructed using items in concordance with current literature and is beneficial in assessing the mentally retarded. Use of the scale in its entirety is helpful in profiling strengths as well as weaknesses.

516 VII. ADVANCED TOPICS

The BASC was developed using exceptional standardization procedures, as item analyses were carefully conducted to ensure adequate discrimination and assessed for discrepancies in performance by gender or ethnicity. Where the sample did not match the population on ethnicity, maternal education, geographical region, and special education placement, weighting was used to create unbiased norms. While the BASC also includes a clinical sample, the largest component of which included behavior disorders and attention-deficit/hyperactivity disorder, this sample was not adjusted, and white males were overrepresented. Internal consistency estimates were good for all scales with alphas ranging in the .80s and .90s, and the clinical sample has similar internal consistencies. However, alphas for individual scales were lower than for composites. For example, the PRS revealed average coefficients in the mid to upper .70s (Jones & Witt, 1994). Test-retest reliabilities, with a 1-month interval between administrations, were also in the mid-.80s to mid-.90s with the exception of some in the .70s for the PRS for adolescents. Interrater reliabilities for the TRS were somewhat low for the preschool form of the BASC, but composite correlations ranged from .69 to .89. Moderate correlations were found between the PRSs for all age groups. Latent trait analyses and factor analyses were conducted

to establish the scales and composites ensuring good construct validity, and the three-factor structure of Internalizing, Externalizing, and Adaptive Skills was supported. Correlations between the TRS and several other teacher rating scales were high. The PRS was also correlated highly with the Child Behavior Checklist and moderately with the Personality Inventory for Children—Revised and the Conners Parent Rating Scales. The SRP correlations with other self-report inventories varied and indicate that the BASC-SRP may measure something different than some of the other scales. Validity of the BASC was also supported by the performance of many groups of children with previously acquired diagnoses of conduct disorder, behavior disorder, depression, emotional disturbance, attention-deficit/hyperactivity disorder, learning disability, mild mental retardation, and autism. Convergent and discriminant validity was demonstrated. While the correlations were relatively high between the TRS and PRS, with the exception of those at the preschool level, they were only low to moderate between the SRP and PRS and between the SRP and TRS. In summary, the BASC's easy-to-use format, well-constructed manual, and sound psychometric properties make it an excellent choice in assessing a child's behavioral and emotional functioning. While the reliability and validity of the TRS and PRS are highest, the SRP is adequate, but the reading level required of 8- to 11-year-olds may limit its use with this age group (Sandoval, 1998).

#### **Personality Inventory for Children—Revised**

This rating scale, to be completed by the child's guardian, preferably the biological mother, is to be used for children ages 3–16 (PIC; Wirt, Lachar, Klinedinst, & Seat, 1984). The test was designed in the late 1970s to elicit 33 scaled scores, 16 profile scores, and 17 experimental scores describing a child's behavior, emotional and cognitive processes, and family characteristics. While the entire Personality Inventory for Children (PIC) contains 600 true/false items, it is possible to administer only certain sections of the test (e.g., eliminating the supplemental experimental scales). The format of the test is easy to understand and follow, and only a sixth- to seventh-grade reading level is required of the respondent.

As the PIC was normed using biological mothers, validity may be affected if someone other than the biological mother completes the scale. Also, the form should be completed in one sitting, if possible, so that answers are not contaminated if the respondent talks to someone else about the items between test sessions. Despite the length of the PIC, clinicians should not be discouraged from administering additional instruments to other family members or to those individuals familiar with the child in order to obtain a comprehensive picture of the child's functioning.

The area of primary concern in administering the PIC relates to the test's outdated norms. The original test was normed between 1958 and 1962 in a relatively small geographical area. Although new scale items were developed in the 1970s, the old norms were applied. Furthermore, the manual lacks specific age and gender norm differences, which do not coincide with the effects of development on the characteristics that the scales purport to measure. The revised manual lists studies from the 1977 manual to support reliabilities for each scale, which ranged from .46 to .94 for a psychiatric outpatient sample, from .50 to .89 for a sample of normal children, and from .68 to .97 with another sample of normal children. One internal consistency study conducted with a heterogeneous clinic sample estimated alpha between .57 and .86 for the scales. A vast array of validity studies conducted on the PIC indicated excellent results. However, a national restandardization of the PIC is needed to determine its current diagnostic utility (Knoff, 1989).

#### **Piers–Harris Children's Self-Concept Scale**

The Piers–Harris Children's Self-Concept Scale (CSCS; Piers & Harris, 1969) was originally designed to measure children and adolescents' self-concepts. The scale, which can be used with children in grades 4–12, and requires a third-grade reading level, includes 80 yes-or-no questions about how the child feels about him or herself in a self-report format. The Piers–Harris produces six cluster scores (Behavior, Intellectual and School Status, Physical Appearance and Attributes, Anxiety, Popularity, and Happiness and Satisfaction), an overall self-concept scale, a Response Bias Index, and an Inconsistency Index. The comprehensive manual,

revised in 1984, is easy to follow and does a good job of cautioning the user against interpreting the scale without support from other instruments and in listing its weaknesses. Although the Piers–Harris has not been renormed since the original standardization sample, which included 1,183 students in Pennsylvania, several recent reliability and validity studies have supported its continued use. Internal consistency estimates for the Total Score ranged from .73 to .81, and test–retest reliabilities ranged from .42 to .96. Validity studies comparing the Piers–Harris with other measures of selfconcept, personality, and behavior revealed moderate relationships. Factorial validity studies have resulted in conflicting results. Some studies supported the six-factor structure, but others found additional factors or indicated factor instability. Therefore, interpretation of cluster scores may not be valid. The ease of use, brevity, and sound reliability and validity make the Piers–Harris an attractive option for assessing self-concept and provide a base for the further investigation of anxiety related issues. However, the outdated norms and the limited geographical region in which the test was originally normed are areas that require future investigation (Epstein, 1985).

### **State–Trait Anxiety Inventory**

The State–Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1969), for ages 9–16 and adults, assess anxiety on two levels, state and trait anxiety. State anxiety is represented by the authors as an immediate sense of how one is feeling, whereas trait anxiety is a more permanent condition. Each level is measured by 20 items in counterbalanced order. The theory behind the state and trait anxiety distinction is welloutlined in the manual, which is also easy to read and interpret.

The STAI was normed on 377 high school juniors, 982 college freshman, 484 introductory psychology course students, 461 male neuropsychiatric patients, 161 general medical and surgical patients, and 212 state prisoners. Separate norms are provided for males and females. Normative means and standard deviations are provided for particular psychiatric diagnoses, medical patients, and prisoners.

Internal consistency coefficients range from .83 to .92 for high school and college students' state scores and from .86 to .92

for trait scores. Test–retest reliabilities were reported as follows: State Anxiety with a 1-hour interval between test administrations was .33 for males and .16 for females, with a 20-day interval was .54 for males and .27 for females, and with a 104-day interval was .33 for males and .31 for females. Trait anxiety was .84 for males and .76 for females after 1 hour, .86 for males and .76 for females after 20 days, and .73 for males and .77 for females after 104 days. Criterion-

#### 518 VII. ADVANCED TOPICS

related validity studies indicated that the STAI college women were highly correlated with other measures. Factor analyses generally supported the distinction between state and trait anxiety (Dreger, 1978).

#### Revised Children's Manifest Anxiety Scale

The Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Richmond, 1985) was designed to assess the level of children's anxiety across five scales; Physiological Anxiety, Worry/Oversensitivity, Social Concerns/Concentration, Total Anxiety, and Lie. The test can be used with children and adolescents from the ages of 6–19, requires a third-grade reading level, and is formatted into a 37-item self-report. The extremely well-organized manual contains information regarding the theory of anxiety on which the test is based and several case studies that demonstrate how the test was used as part of a larger psychoeducational battery. The RCMAS manual is easy to use and contains the necessary information for interpretation of each of the five scores; however, the authors caution against using the three individual anxiety scaled scores for more than hypothesis generation due to their low reliability. They do recommend using the separate norms provided according to age, sex, and ethnicity (Gresham, 1989). The large standardization sample of 4,972 included 44% white males, 44% white females, 5.8% African American males, and 6% African American females and covered a variety of geographic regions throughout the United States.

Internal consistency studies revealed alpha levels for Caucasian and African American males and females for each age level for the Total Anxiety Score in addition to standard errors of measurement for the Total Anxiety Score. Coefficient alphas ranged from .42 to .84, so interpretation for groups with lower reliability estimates should be

made cautiously. Internal consistency for each of the anxiety subscales was lower than optimal for direct interpretation. Test-retest reliability estimates were reported for the Total Anxiety Score at .68 and for the Lie Scale Score at .58 with an interval of 9 months, but when the interval was 3 weeks, test-retest reliability was reported at .98. No test-retest coefficients were provided for groups older than the seventh grade. The authors' claim that the RCMAS is a measure of trait, as opposed to state, anxiety was supported by a validity study that indicates that the Total Anxiety Score is more highly correlated with a measure of trait anxiety than state anxiety on the STAIC (Revised Children's Manifest Anxiety Scale; Stewart, 1989). The factor analysis produced a five factor solution, but theoretically invalid rotation procedures reported in the manual may render the five-factor structure uninterpretable (Grisham, 1989).

#### **Child Anxiety Scale**

The Child Anxiety Scale (CAS; Gillis, 1980) is a self-report questionnaire that was derived from the Early School Personality Questionnaire (ESPQ) developed by Coan and Cattell in 1959. The 20-item scale can be administered to children from ages 5–12 and is easy for young children to understand due to its dichotomous response method of choosing either a red or blue circle in response to pictorial and audiotaped items. Young children may still have difficulty understanding the nature of the task, and neither a lie nor social desirability scale is provided. However, the bright and colorful nature of the scale, in addition to its brevity, make it an appealing choice for use with children.

The author gives separate age and grade norms but does not separate norms by gender, claiming that the tendency for females to score higher than males was not strong enough to warrant separate norms. The 2,105 children sample is described in the manual according to geographic region and ethnicity. The one area of concern regarding the norming sample relates to the small number of children from urban areas that were included in the norming sample. Seventy-five percent of the children were sampled from areas with populations less than 50,000, indicating that use of the instrument with children from urban settings should be interpreted cautiously. Test-retest

and internal consistency coefficients were good. Test–retest coefficients for firstthrough third-graders ranged from .82 to .92, and a Kuder–Richardson 20 coefficient for internal consistency was .81. However, validity studies are severely lacking, as the 23. Assessment of Childhood Anxiety 519 manual reports no correlations between the CAS and other measures of anxiety. However, factorial validity is supported by factor analyses conducted on the ESPQ (Maxwell, 1985).

### **Multidimensional Anxiety Scale for Children**

The Multidimensional Anxiety Scale for Children (MASC; March, 1997) is a self-report measure that can be used to assess anxiety in children and adolescents ages 8–19.

The MASC is a brief measure consisting of items rated on a 0- (never true about me) to 4-point (always true about me) Likert scale, with 4 indicating the most severe emotional problems. The MASC produces a total of four factor scales, a total scale, and a validity scale, which amount to 13 subscales, including Physical Symptoms (Tense Symptoms Subscale, Somatic Symptoms Subscale, Total), Harm Avoidance (Perfectionism Subscale, Anxious Coping Subscale, Total), Social Anxiety (Humiliation Fears Subscale, Performance Fears Subscale, Total), Separation/Panic, Total Anxiety, Anxiety Disorders Index, and an Inconsistency Index. Total testing time is 15 minutes (Caruso, 2001).

Although reliability and validity of the scale were generally acceptable, the reliability of some specific subscales were questionable.

For example, the Perfectionism and Anxious Coping Subscales of the Harm Avoidance Scale, the Performance Fears Subscale, and the Anxiety Disorders Index all had internal consistency reliability coefficients of less than .65 in the normative sample.

However, the Anxiety Disorders Index consists of items tapping several types of anxiety disorders, which may account for the heterogeneity of test items. Validity was demonstrated through confirmatory factor analysis, which supported the four factor structure. The four general scores and the Anxiety Disorders Index effectively discriminated between a random subset from the normative sample and a separate group of children with DSM-IV diagnosed anxiety disorders. Caution should be used in diagnosing OCD, however, because children

with OCD were not included in the group used to develop the Anxiety Disorder Index Scale. Further validity studies are warranted to determine if the scale can differentiate between different types of anxiety disorders (Christopher, 2001).

The norming sample consisted of 2,698 children and adolescents but lacked an adequate representation of minorities, particularly Hispanic children, in relation to census proportions. The racial distribution of the sample was as follows: 53.3 % Caucasians, 39.2 % African Americans, 7% Hispanic/Latin Americans, 1.4 % Asian Americans, 2.4% Native Americans, and 3 % other. Separate norms are provided for males and females in 4-year intervals but not for racial groups. The test has a fourth-grade reading level, and items may be read to young test takers, but no validity data regarding the ability of nonnative English speakers to respond to test items is provided (Caruso, 2001).

The MASC-10 is a short version of the original test providing a single score to be used for screening purposes. The validity of this measure was less than optimal, and internal consistency coefficients were also questionable. Overall, the MASC appears to be a useful screening tool. However, further validity studies are required to support its diagnostic value (Caruso, 2001).

#### **Anxiety Scales for Children and Adults**

The Anxiety Scales for Children and Adults (ASCA; Battle, 1993) was developed to measure the presence and level of anxiety in individuals via a self-report form. There are two forms, form Q for children and form M for adults. Both forms list items that, according to the authors, correspond to common symptoms of anxiety. The child form contains 25 dichotomously scored items, while the adult form contains 40 items from 1 (*never*) to 5 (*always*).

No mention of item analysis, internal consistency analysis, or factor analysis is made in the manual. Due to a lack of information that one would expect in the manual in regard to development of the test items, the basis on which the items were chosen (e.g., according to DSM criteria) is not apparent (Merenda, 1995). According to a review by Oehler-Stinnett (1995), no distinction is made by the authors between real or imagined anxiety-producing stimuli, state or trait anxiety symptoms, and fear, worry, or

panic. Test-retest reliability is reported at  
520 VII. ADVANCED TOPICS

.82 to .96, but time between testing was 1 week for the child's form and 2 weeks for the adult form. Furthermore, the manual is lacking in empirical evidence for construct validity so it is unclear how closely the test measures anxiety and if there are any specific types of anxiety that the test measures better than others. Predictive validity is not discussed in detail. Studies of concurrent validity revealed that the ASCA Form Q was correlated with the State-Trait Anxiety Inventory for Children at  $r = .64$  and the Nervous Symptoms subtest of the California Test of Personality at  $r = .66$ . Form M was correlated with the Taylor Anxiety Scale for Adults at  $r = .75$  and the Nervous Symptoms subtest of the California Test of Personality at  $r = .63$  (Oehler-Stinnett, 1995).

However, the ASCA was also highly correlated with several depression inventories, calling into question interpretability of the scores as used to measure anxiety. The manual also lacks studies regarding performance on the ASCA by age, socioeconomic status, race, or clinical populations. In summary, one should use caution in using this form for diagnostic or treatment decisions due to the lack of empirical support for the test's validity (Merenda, 1995).

#### **Social Phobia and Anxiety Inventory**

The Social Phobia and Anxiety Inventory (SPAI; Turner, Beidel, & Dancu, 1996) is a self-report rating scale that assesses the presence of social phobia characteristics in individuals 14 and over. The test produces three scores, a Social Phobia Score (SP), Agoraphobia Score (Ag), and a Difference score (SP minus Ag), from a 6-point Likert scale assessing how frequently the test taker experiences anxiety in response to given circumstances.

A sixth-grade reading level is required to take the test without assistance.

Although the SPAI is easy to score, the rationale for scoring methods is not fully explained in the manual. Furthermore, the authors endorse the use of clinical judgment in interpreting the pattern of responses given by examinees. For example, the test does not delineate a separate score for anxiety in particular situations, such as with strangers, but a trained clinician may glean this information by reviewing the pattern of item scores.

Of the 308 college students used in test

development, 182 were used to calculate test-retest reliability, which resulted in the following scores:  $r(173) = .85$  for SP,  $r(173) = .74$  for Ag, and  $r(173) = .86$  for Difference scores, suggesting adequate stability across test administrations. Internal consistency scores were also high, as Cronbach's alpha was estimated at .96 for SP and .85 for Ag scores. Validity studies supported the use of the SPAI for identifying social phobia in individuals. Although factor analysis was performed, factor structures varied from group to group. However, confirmatory factor analyses performed by the authors supported the two-factor (SP and Ag) solution (Walcott, 2001). Furthermore, the authors' claim that the Difference score produces a purer measure of social phobia has been debated. The apparent redundancy of test items may cause concern over the validity of interpreting scores based on these items. The authors do provide information regarding interpretation of the SPAI for different genders, ages, and races, although sample sizes on which this information was based were relatively small (Engelhard, 2001). Although the SPAI has primarily been used with adults, a newer version for children, the SPAI-C, has been developed (Walcott, 2001).

### Test Anxiety Profile

The Test Anxiety Profile (TAP; Oetting & Deffenbacher, 1980) provides a measure of test anxiety in a variety of different testing situations for individuals in grades nine through college. Two components of anxiety are assessed across six different testing situations, (multiple choice test, time-limit test, "pop" quiz, essay test, giving talk, and math test) providing 12 scores. In addition, two anxiety scores may be used to assess a student's anxiety feelings and thoughts in unique testing situations. The Feeling Anxiety (FA) score purports to measure self-perceived emotions and physiological responses to particular testing situations, and the Thought Interference (TI) score was constructed to measure a person's cognitions that may interfere with thought processes required to perform efficiently on an exam. Although the authors have studied test anxiety extensively, and the TAP appears to have adequate psychometric properties, the manual does not describe in detail the norming sample and the procedures used to

construct individual items. It appears that the sample consisted of 600 Colorado State University students with reliability and validity studies based on this sample and a sample of 61 high school students. However, the internal consistency estimates reported for the college student sample ranged from .88 to .96. Furthermore, a test-retest reliability study with a 7- to 10-week interval between tests revealed coefficients ranging from .66 to .81. Criterion validity studies resulted in correlations from .28 to .59. Discriminant validity studies appeared to conform to patterns of perceived thoughts and emotions before and after test taking in the anxiety literature; however, no evidence was provided to support that the FA and TI scales measure what they are intended to measure (Brown, 1985). In summary, the TAP appears to have adequate psychometric properties to be a useful diagnostic tool, but other factors, such as study skills and test-taking strategies should be examined as possibly influencing scores on this measure.

#### **Schedule for Affective Disorders and Schizophrenia, Third Edition**

The SADS (Endicott & Spitzer, 1978) is a semistructured interview developed in the mid-1970s primarily as a research tool to identify groups of patients with the same psychiatric symptoms. The SADS was designed to decrease the unreliability of diagnostic practices by providing questions in a sequential format based on the Research Diagnostic Criteria, a forerunner to DSM-III (RDC; Spitzer, Endicott, & Robins, 1975). The SADS takes approximately 1½ to 2 hours to administer and is divided into two parts, focusing on the current episode and past history, respectively. However, not all DSM diagnoses can be obtained by using the SADS, most notably PTSD. The interview results in a total of 24 scales including a Global Assessment Scale. The SADS has been reliable in diagnosing many disorders, but the reliability depends on the validity of measurement with each specific disorder. Concurrent validity studies comparing the diagnoses of the SADS with those of the Diagnostic Interview Schedule were weak, indicating that using the SADS for diagnostic purposes rather than as it was intended, as a research instrument, may not be useful. Intensive training and practice is recommended to use the SADS, as the format is somewhat confusing and assumes a certain level

of knowledge of the subject area. Furthermore, the terminology may be outdated and translation into current symptom terminology may be necessary.

The Kiddie-Schedule for Affective Disorders (K-SADS; Kaufman, Birmaher, Brent, Ryan, & Rao, 2000) has been modified from the SADS to be used with school-age children (Carmer, 1995). The K-SADS addresses most anxiety disorders through parent and child interviews, and like the SADS, continues to focus on past and present episodes of psychopathology. Due to the unreliability of diagnosing the specific anxiety disorders, several different versions of the K-SADS have also been developed in attempts to address structural inadequacies of the interviews (Silverman, 1994).

#### **Anxiety Disorders Interview Schedule for DSM-IV: Child Version**

The Anxiety Disorders Interview Schedule for DSM-IV: Child Version (ADIS for DSM-IV: C) was revised from the ADIS-C in 1996, which was revised from the original 1983 version of the ADIS (Silverman & Albano, 1996). The interview can be used with children ages 7–17, who are suspected of having any of the anxiety disorders listed in DSM-IV. The semistructured interviews were developed to be consistent with DSM-IV criteria, and include a child version as well as a detailed parent version. In addition to sections devoted to the anxiety disorders, separate sections that assess mood and externalizing disorders, school refusal behavior, and screening sections for substance abuse, psychosis, selective mutism, eating disorders, somatoform disorders, and learning disorders are included. According to the authors, the ADIS for DSM-IV:C is an effective instrument for use with research and clinical populations. It is recommended that both the child and parent interviews be administered to obtain a comprehensive diagnosis.

All diagnostic questions can be answered dichotomously (e.g., yes/no), and interference scores that rate the degree to

#### **522 VII. ADVANCED TOPICS**

which symptoms interfere with the child's life may be given on a 0–8 scale. The child interview begins with an explanation of the "Feelings Thermometer," which allows the child to visually rank his or her feelings of anxiety by choosing the appropriately filled thermometer.

The manual reports internal reliability estimates

of between .64 to 1.00 for each of the specific anxiety disorders with an overall kappa of .75. Test-retest reliability coefficients obtained from a sample of 50 outpatients with a 10-14-day interval between test administrations ranged from .64 to .84 for individual anxiety disorders and .75 overall. The authors report that criterion validity for symptom scale scores and for symptom summary scores is good for both the parent and child interviews (Albano & Silverman, 1996.)

## CONCLUSION

To more fully understand the way in which children and adolescents experience anxiety, further research must investigate the nature of worry in typically developing children versus children who develop symptoms at clinically significant levels. Thus far, most of the literature regarding the assessment of anxiety has been developed through investigations of adult symptoms. Even the prevalence of specific anxiety disorders and the frequency of occurrence of anxiety symptoms in children are largely unknown, and to date, extensive investigation into the progression of anxiety symptoms from childhood into adolescence and adulthood is lacking (Albano et al., 1996). Consistent with these future studies are continued explorations into optimal diagnostic methods for children and adolescents, which may include alternatives to categorical systems. In addition to further inquiry into anxiety assessment, advances in treatment methods, including pharmacological developments, will follow new research. Currently, adrenergic, serotonergic, and GABA(4-aminobutyrate) ergic neurotransmitter systems have been explored in regard to anxiety (Bernstein, 1994). Yet once again, a paucity of literature exists that focuses on the use of anxiety medications with children and adolescents. The effectiveness of medications used to treat different childhood anxiety disorders has also not been well studied. These issues are particularly important given the increasingly high rate of anxiety disorders detected in school children. While attention and behavior problems were once thought to be the most serious childhood disturbances in school, recent research has uncovered that internalizing problems, such as anxiety disorders, may be even more prevalent. Thus, school-based approaches will be new areas of practice and research in

psychology, and new advances in assessment, as well as treatment, will surely follow.

## REFERENCES

- Albano, A. M., Chorpita, B. F., & Barlow, D. H. (1996). Childhood anxiety disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 196–241). New York: Guilford Press.
- Albano, A. M., & Silverman, W. K. (1996). *Anxiety Disorders Interview Schedule for DSM-IV: Clinician manual*. San Antonio, TX: Psychological Corporation.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Anderson, J. C. (1994). Epidemiological issues. In T. H. Ollendick, N. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 43–65). New York: Plenum Press.
- Battle, J. (1993). *Anxiety scales for children and adults*. Austin, TX: Pro-Ed.
- Beidel, D. C., & Stanley, M. A. (1993). In C. G. Last (Ed.), *Anxiety across the lifespan: A developmental perspective* (pp. 167–203). New York: Springer.
- Bell-Dolan, D., & Brazeal, T. J. (1993). Separation anxiety disorder, overanxious disorder, and school refusal. *Child and Adolescent Psychiatric Clinics of North America*, 2, 583–580.
- Bernstein, G. A. (1994). Psychopharmacological interventions. In T. H. Ollendick, J. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 169–186). New York: Plenum Press.
- Bernstein, G. A., & Borchardt, C. M. (1991). Anxiety disorders of childhood and adolescence: A critical review. *Journal of the American Academy of Child and Adolescent Psychiatry*, 30, 519–532.
- Blagg, N., & Yule, W. (1994). School refusal. In T. H. Ollendick, J. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 169–186). New York: Plenum Press.
- Brown, S. D. (1985). Test Anxiety Profile. In *The ninth mental measurements yearbook* (Vol. 2, pp. 1543–1545). Lincoln: University of Nebraska Press.
- 23. Assessment of Childhood Anxiety 523**
- Carmer, J. C. (1995). Schedule for Affective Disorders and Schizophrenia, Third Edition. In *The twelfth mental measurements yearbook* (pp. 918–919). Lincoln: University of Nebraska Press.
- Caruso, J. C. (2001). Multidimensional Anxiety Scale for Children. In *The fourteenth mental measurements yearbook* (pp. 799–800). Lincoln: University of Nebraska Press.
- Christopher, R. (2001). Multidimensional Anxiety Scale for Children. In *The fourteenth mental measurements yearbook* (pp. 801–803). Lincoln: University of Nebraska Press.
- Coan, R. W., & Cattell, R. B. (1959). The development of the Early School Personality Questionnaire. *Journal of Experimental Education*, 28, 143–152.
- Costello, E. J., & Angold, A. (1988). Scales to assess child and adolescent depression: Checklists, screens, and nets. *Journal of the American Academy of Child and Adolescent Psychiatry*, 27, 726–737.
- Dreger, R. M. (1978). State-Trait Anxiety Inventory. In *The eighth mental measurements yearbook* (Vol. 1, pp. 1094–1096). Lincoln: University of Nebraska Press.
- Endicott, J., & Spitzer, R. L. (1978). *A diagnostic interview:*

*The Schedule for Affective Disorders and Schizophrenia*. New York: Department of Research Assessment and Training, New York State Psychiatric Institute.

Engelhard, G. (2001). Social Phobia and Anxiety Inventory. In *The fourteenth mental measurements yearbook* (pp. 1161–1163). Lincoln: University of Nebraska Press.

Epstein, H. (1985). Piers–Harris Children’s Self-Concept Scale. In *The ninth mental measurements yearbook* (Vol. 2, pp. 1168–1169). Lincoln: University of Nebraska Press.

Gillis, J. S. (1980). *Child Anxiety Scale*. Champaign, IL: Institute for Personality and Ability Testing Inc.

Gresham, F. (1989). Revised Children’s Manifest Anxiety Scale. In *The tenth mental measurements yearbook* (pp. 695–697). Lincoln: University of Nebraska Press.

Herbert, M. (1994). Etiological considerations. In T. H. Ollendick, N. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 3–20). New York: Plenum Press.

Jones, K. M. & Witt, J. C. (1994). Rating the ratings of raters: A critique of the Behavior Assessment System for Children. *Child Assessment News*, 4, 10–11.

Kaufman, J., Birmaher, B., Brent, D. A., Ryan, N. D., & Rao, U. (2000). K-SADS-PL. *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 1208.

Kearney, C. A., & Allan, W. D. (1995). Panic disorder with or without agoraphobia. In A. R. Eisen, C. A. Kearney, & Schaefer (Eds.), *Clinical handbook of anxiety disorders in children and adolescents* (pp. 251–281). Northvale, NJ: Jason Aronson.

Kearney, C. A., & Wadiak, C. (1999). In S. D. Netherton, D. Holmes, & C. E. Walker (Eds.), *Child and adolescent psychological disorders: A comprehensive textbook* (pp. 282–303). New York: Oxford University Press.

Knoff, H. M. (1989). Personality Inventory for Children, Revised Format. In *The tenth mental measurements yearbook* (pp. 625–630). Lincoln: University of Nebraska Press.

Last, C. G., & Strauss, C. C. (1989). Obsessive–compulsive disorder in childhood. *Journal of Anxiety Disorders*, 3, 295–302.

Last, C. G., Strauss, C. C., & Francis, G. (1987). Comorbidity among childhood anxiety disorders. *Journal of Nervous and Mental Disease*, 175, 726–730.

March, J. (1997). *Multidimensional Anxiety Scale for Children*. Tonawanda, NY: Multi-Health Systems.

Maxwell, S. (1985). Child Anxiety Scale. In *The ninth mental measurements yearbook* (Vol. 1, pp. 297–298). Lincoln: University of Nebraska Press.

Merenda, P. F. (1995). Anxiety Scales for Children and Adults. In *The twelfth mental measurements yearbook* (pp. 78–79). Lincoln: University of Nebraska Press.

Oehler-Stinett, J. (1995). Anxiety Scales for Children and Adults. In *The twelfth mental measurements yearbook* (pp. 79–81). Lincoln: University of Nebraska Press.

Oetting, E. R., & Deffenbacher, J. L. (1980). *Test Anxiety Profile*. Fort Collins, CO: Rocky Mountain Behavioral Science Institute.

Piers, E. V., & Harris, D. B. (1969). *Piers–Harris Children’s Self-Concept Scale (The Way I Feel About Myself)*. Los Angeles, CA: Western Psychological

Services.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children*. Circle Pines, MN: AGS.

Reynolds, C. R., & Richmond, B. O. (1985). *Revised Children's Manifest Anxiety Scale*. Los Angeles, CA: Western Psychological Services.

Sandoval, J. (1998). Behavior Assessment System for Children. In *The thirteenth mental measurements yearbook* (pp. 128–131). Lincoln: University of Nebraska Press.

Silverman, W. K. (1994). Structured diagnostic interviews. In T. H. Ollendick, N. J., King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 293–315). New York: Plenum Press.

Silverman, W. K., & Albano, A. M. (1996). *The Anxiety Disorders Interview Schedule for DSM-IV—Child and Parent Versions*. San Antonio, TX: Psychological Corporation.

Silverman, W. K., & Rabian, B. (1999). Rating Scales for Anxiety and Mood Disorders. In D. Shaffer & C. P. Lucas (Eds.), *Diagnostic assessment in child and adolescent psychopathology* (pp. 127–166). New York: Guilford Press.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. (1970). *State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Spitzer, R. L., Endicott, J., & Robins, E. (1975). *Research Diagnostic Criteria (RDC)*. New York: Biometrics Research, New York State Psychiatric Institute.

Stewart, K. J. (1989). Revised Children's Manifest Anxiety Scale. In *tenth mental measurements year-*

#### 524 VII. ADVANCED TOPICS

*book* (pp. 697–699). Lincoln: University of Nebraska Press.

Swedo, S. E., Rapoport, J. L., Leonard, H., Lenane, M., & Cheslow, D. (1989). Obsessive-compulsive disorder in children and adolescents: Clinical phenomenology of 70 consecutive cases. *Archives of General Psychiatry*, 46, 335–341.

Turner, S. M., Beidel, D. C., & Dancu, C. V. (1996). *Social Phobia and Anxiety Inventory*. Tonawanda, NY: Multi-Health Systems.

Walcott, D. D. (2001). Social Phobia and Anxiety Inventory. In *The fourteenth mental measurements yearbook* (pp. 1163–1164). Lincoln: University of Nebraska Press.

Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1984). *Multidimensional description of child personality: A manual for the Personality Inventory for Children*. Los Angeles, CA: Western Psychological Services.

Historically, anxiety has been conceptualized as a circumstance of adulthood. Children were not thought to experience anxiety in any unique way, and scientific studies of anxiety disorders have, until recently, focused on adults. However, while the majority of theories about anxiety have focused on symptoms in adults, they have been based on limited empirical evidence in general and have been born of many diverse psychological perspectives. Yet, most theorists agree that there is not a single cause but multiple factors interacting with each other which lead to the development of anxiety disorders.

While the quest for genetic links, family characteristics, environmental influences, and proximal antecedents that cause anxiety continues, it is likely that all of these elements act on and shape each other to form each person's individual anxiety experience. Therefore, the process of assessing anxiety must examine many aspects of an individual's functioning, such as physiological arousal, cognitions, behaviors, and subjective interpretations. The anxiety itself must also be assessed by measuring its severity, duration, and pervasiveness in the individual's life (Herbert, 1994).

### THEORIES OF ANXIETY

Of the many models developed to explain anxiety, three seem to predominate. They include behavioral, cognitive, and biological models. Behavioral models are primarily based on classical and operant conditioning. According to classical conditioning, anxiety-provoking stimuli elicit physical sensations that alert the individual to the presence of danger (i.e., the feared stimuli). Pairing a neutral stimulus with the anxietyprovoking stimulus causes the individual to associate the physical symptoms of anxiety with the neutral stimulus. Once the individual has made this association, exposure to the neutral stimulus alone results in anxiety. This classical conditioning model coupled with reinforcement, such as what occurs when anxiety is reduced through avoidance, may further increase the likelihood that the anxiety response will persist. While the behavioral model has been used to explain the creation of phobias, clearly the model does not account for the vast majority of cases of anxiety. Even many phobic individuals cannot identify instances in which neutral stimuli were paired with

508

# 23

## Assessment of Childhood Anxiety

KELLY ROBINSON

feared objects or situations (Kearney & Wadiak, 1999).

Cognitive theories of anxiety explain symptoms by identifying the normal processes by which children assimilate information in the world. Newly acquired information is placed within the cognitive

structures of previously acquired information. People compare the new information with the old in order to make predictions about their environment. When individuals learn to expect negative consequences from situations they perceive as aversive or dangerous, they avoid those situations. Thus, when those aversive situations are unavoidable, and individuals must endure encounters that they perceive are threatening, they experience feelings of anxiety (Kearney & Wadiak, 1999).

Biological theories of anxiety focus on genetic studies. Though a single causal gene has not been found, results of several studies suggest that there are genetic links to anxiety disorders, particularly panic disorder, obsessive-compulsive disorder, and certain types of phobias (Kearney & Wadiak, 1999).

### **CHILDREN'S ANXIETY**

Children who experience pathological levels of anxiety represent a category of individuals for whom assessment and diagnosis becomes particularly challenging. While it is still important to assess all the major areas of a child's life that were discussed previously (proximal antecedents, family characteristics, physiological functioning, etc.), one must also bear in mind the developmental appropriateness of anxiety during stages of a child's life. In other words, a certain amount of anxiety is healthy, so it is important to determine if anxiety is present and if it is problematic for the child.

Anxiety is a normal and developmentally appropriate response that allows humans to adapt to their environment. From an evolutionary perspective it allows for survival in the "fight or flight" action a person takes when faced with danger, and in a performance capacity a limited amount of anxiety helps a person function at his or her best.

Children progress through stages of normal anxiety throughout their lifespan. Therefore, it is important to be knowledgeable about these stages so as not to confuse normal developmental anxiety from abnormal anxiety that leads to dysfunction. Children tend to progress through the developmental anxiety trajectory beginning in infancy with fear of sudden noises and of being dropped or startled. At age 1 to 2, children begin feeling anxious when not in close proximity to caregivers. Then fear of strangers emerges, and when children reach the

preschool years, fear of monsters, animals, and the dark are common. Finally, in early to middle childhood, children may develop specific fears of objects or situations, such as school, physical danger, and supernatural phenomena (Beidel & Stanley, 1993).

Although these anxieties represent normal developmental stages, abnormal or pathological anxiety also follow these stages, with pathological anxiety differing in degree.

Evidence that anxiety is a common part of normal development is suggested by studies of subclinical anxiety symptoms.

These studies show that individuals who have identifiable anxiety symptoms but who do not meet diagnostic criteria for an anxiety disorder and individuals who are asymptomatic have similar functional outcomes.

Qualitatively, clinical levels of anxiety may appear similar to normal developmental anxiety. However, pathological anxiety differs in that it is generally inappropriate to the situation, involuntary, irrational, and functionally limiting (Anderson, 1994).

Investigation into several areas may be useful in determining whether anxiety is at a level that may be considered problematic or pathological. First, context is important. Is the anxiety out of proportion to the feared stimulus, situation, or event? Given that children experience a progression of developmentally typical fears, the context may be wider than for adults (Herbert, 1994).

However, it is still possible to assess the degree of the anxiety and its age appropriateness.

It is also prudent to examine the frequency and intensity of the child's anxiety. For example, is the child's level of anxiety within a range that is present in most other children within the same age range? Also, does the anxiety occur with more frequency and in regard to a wider range of stimuli than in most

23. Assessment of Childhood Anxiety 509  
other children of the same age? Whereas constant low levels of anxiety can be just as maladaptive as high levels of anxiety, examining the frequency and intensity should provide a clearer picture of how a child's anxiety impacts his or her functioning.

The intensity of children's anxiety often varies relative to situations, times, places, or people with whom the child engages. Therefore, scrutinizing the child's environment and also the child's perspective of his or her own anxiety is important. How the child

makes sense of his or her anxiety may provide insight into any purpose that the anxiety may serve. Whereas conducting a functional analysis of antecedents and consequences may provide valuable information regarding the types of problems the child experiences due to anxiety, it will not provide information about how the child views the anxiety as a problem. Attempting to see anxiety from the perspective of the child may also help in developing treatment goals and strategies (Herbert, 1994).

## ANXIETY DISORDERS

Many anxiety problems go undetected due to the difficulty that school personnel and parents have in identifying the disorder, as anxiety frequently results in only relatively mild behavioral problems. Furthermore, anxiety disorders in children have a high comorbidity rate with other disorders, such as depression and attention-deficit/hyperactivity disorder, which further obscures the presentation of symptoms and makes identification more difficult. This difficulty in identifying anxiety disorders in children is an important issue in regard to assessment, as most anxiety problems begin in childhood and progress to a chronic degree into adulthood, where they may have a profound and detrimental impact on the lives of the individuals affected (Albano, Chorpita, & Barlow, 1996). According to the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DMS-IV; American Psychiatric Association, 1994), there are nine anxiety disorders with which a child may be diagnosed, including separation anxiety disorder (SAD), specific phobia, social phobia, generalized anxiety disorder (GAD) (previously referred to in the literature as overanxious disorder), panic disorder with or without agoraphobia, obsessive-compulsive disorder (OCD), posttraumatic stress disorder (PTSD), and acute stress disorder. SAD is the only disorder that is unique to children. Each disorder shares common features but may be differentially diagnosed based on the primary focus of the child's anxiety.

The prevalence of anxiety disorders in children and adolescents ranges from approximately 9 to 27% (Albano et al., 1996).

However, there is wide variability in the reporting of prevalence rates, which is influenced by a number of factors such as diagnostic and referral processes, likelihood of participation in studies, use of general populations

versus clinical samples, relationship of reporter of symptoms, and age of population. Differences in the reports of prevalence rates aside, anxiety disorders are the most common psychiatric disorder diagnosable in children and adolescents (Bell-Dolan & Brazeal, 1993). Anxiety disorders seem to increase in adolescence with phobic symptoms more prevalent in girls than in boys, with the exception of public-speaking phobias. The symptoms of anxiety peak earlier in girls than in boys, but boys lose specific fears more quickly than do girls, usually around age 10. Less gender variability is noted in younger children. The most common disorders comorbid with anxiety are depression and dysthymia, ADHD, conduct and oppositional disorders, and other specific anxiety disorders. There is little evidence that GAD, phobias, and SAD continue into adult life. Recovery rates for these anxiety disorders range from 60 to 75%, and those that do progress into adulthood may not continue with the same pattern of symptoms. OCD, on the other hand, has shown considerable stability into adulthood. Boys tend to demonstrate a younger age of onset, and the disorder also seems to be more prevalent among boys. While most individuals with OCD report no previous history of psychopathology, the disorder is sometimes comorbid with depression, other anxiety disorders, and Tourette syndrome (Anderson, 1994).

### **Separation Anxiety Disorder**

A child may be diagnosed with SAD if he or she exhibits developmentally inappropriate and excessive anxiety in regard to separation from home or an attached figure. SAD is diagnosed when the child exhibits three or more symptoms, such as worry about harm coming to attached figures, school refusal, or reluctance to go to sleep without an attached figure present. The symptoms must persist for at least 4 weeks, and onset must be before age 18 (American Psychiatric Association, 1994). The prevalence rate for children with SAD ranges from 3 to 5% in children, and may be a little less in adolescents. One-third of children with SAD have comorbid GAD secondary to SAD, and another third are diagnosed with a comorbid depressive disorder (Last, Strauss, & Francis, 1987). The focus of SAD is fear associated with

separation from home or an attached figure. It is important to remember when assessing for SAD that anxiety about being away from home and family members is a normal part of development from the age of approximately 7 months to 6 years (Bernstein & Borchardt, 1991). However, children with SAD differ from typical children in that they often experience fear of losing loved ones through catastrophic events. Children with SAD tend to demonstrate extreme avoidance behaviors and complain of somatic symptoms, including panic attacks. A key feature of the disorder that distinguishes children with SAD from children experiencing normal developmental anxiety is that SAD interferes with a child's daily functioning, such as friendships, participation in groups, and academics.

### **Specific Phobias**

Specific phobias are characterized by marked and persistent fear or anxiety when exposed to, or when anticipating exposure to, a particular object or situation. Exposure to the feared stimulus results in an immediate anxiety response, which may include a panic attack. In children the anxiety response may take the form of crying, tantrums, freezing, or clinging. Adults must recognize the fear as excessive or unreasonable in order for a diagnosis to be made, but this recognition is not required for children (American Psychiatric Association, 1994). The most important factor in differentially diagnosing specific phobias in children is identifying true phobias versus normal developmental fears. Common childhood fears include heights, darkness, thunder, injections, insects, and dogs (Albano et al., 1996). One feature that distinguishes phobias from typical fears is the degree to which the fear is maladaptive. School phobia is a unique specific phobia because school phobia may look similar to SAD, social phobia, or truancy. While truancy may be an indication of oppositional behavior, the school phobic child's avoidance of school is involuntary. Anxiety about going to school may not even be in the child's conscious awareness, as symptoms frequently manifest through somatic complaints before or during the school day. The fear is typically associated with specific occurrences at school rather than with separation from home or attached figures (Blagg & Yule, 1994). To diagnose a specific phobia

of school, social phobia, which centers around a fear of embarrassment or humiliation, must also be ruled out as a possible diagnosis. School phobia tends to surface shortly after beginning school for the first time (ages 5–6), following a transition from elementary to middle/high school (ages 11–13), and after age 14. However, school phobias experienced by older children and adolescents may be suggestive of a more severe developing pathology (Blagg & Yule, 1994).

### **Social Phobia**

Social phobia is characterized by marked and persistent fear in social and performance situations. In children the anxiety must be present in peer situations, not just during interactions with adults, and children must demonstrate the capacity to form age-appropriate relationships with others. Exposure to feared situations provokes an anxiety response, which may include panic attacks, and in children, may also include crying, tantrums, and shying away from social situations. Adults must recognize that the fear is unreasonable or excessive in order to be diagnosed, whereas in children this criterion does not have to be met. Symptoms must be present for at least 6 months in order for social phobia to be diagnosed in individuals under the age of 18 (American Psychiatric Association, 1994).

#### **23. Assessment of Childhood Anxiety 511**

Social phobia is rarely diagnosable under the age of 10, and has been diagnosed equally in males and females (Albano et al., 1996). Unlike SAD, social phobia is not a disorder specific to children. Therefore, much of the research examining the disorder has been taken from the adult literature. The primary focus of fear in social phobia is fear of embarrassment. Children and adults with the disorder become excessively concerned about rejection, humiliation, and negative evaluation. Not surprisingly, children with social phobia tend to be shy and have few friends. Cognitions are overwhelmingly negative and self-deprecating (Albano et al., 1996). Symptoms of social phobia may include somatic complaints and obsessing about visible signs of embarrassment, such as blushing, shaking, and sweating. Although anyone in embarrassing situations may have these same worries, the difference between children with social phobia and typical children is the degree of

worry. However, an apparent ability to perform during these social activities does not necessarily mean that the disorder is absent. Although social phobic individuals attempt to avoid situations in which they may become embarrassed, they will often endure these situations rather than draw attention to themselves. Despite an apparent ability to perform, they may experience an extreme degree of internal distress during these times (Albano et al., 1996). School refusal is not an uncommon feature of the disorder, and one must be careful to differentially diagnose social phobia, school phobia, and SAD by keeping in mind that the anxiety focus of social phobia involves fear of embarrassment in social situations rather than fear of more specific occurrences at school, or a separation from home.

### **Generalized Anxiety Disorder**

GAD is diagnosed when an individual experiences excessive worry or anxiety more days than not, about a number of situations, and the worry is difficult to control.

Anxiety symptoms may include restlessness, fatigue, difficulty concentrating, irritability, muscle tension, and disturbances in sleep.

Children must demonstrate only one of these symptoms, whereas adults must experience three or more symptoms in order for a diagnosis to be made (American Psychiatric Association, 1994). Although GAD may begin any time during the school-age years, the mean age of onset appears to be between age 10 and 13 (Last et al., 1987). GAD is highly comorbid with depressive disorder and other anxiety disorders.

The presentation of GAD in children appears similar to that of adults except that children are more likely than adults to have somatic complaints. According to DSM-IV, a diagnosis of GAD is warranted in a child only if at least one physiological symptom is present. The focus of GAD is excessive worry about a number of life events, which may include past behavior, competence in sports, academics, peer relationships, or future events. Frequently children with GAD demonstrate perfectionist approaches in performing tasks and set exceedingly high standards for themselves. They are also quite hard on themselves when they do not achieve their goals or when they perceive that they have failed in some way (Albano et al., 1996).

## Panic Disorder

Panic disorder is marked by recurrent and unexpected panic attacks followed by 1 month or more of worry about the future occurrence of attacks or the implications of attacks. Panic disorder may also include agoraphobia, which is anxiety about being in situations in which escape would be difficult or help unavailable if a panic attack should occur, such as in public places. Individuals with agoraphobia avoid these situations, likely resorting to extreme behavior in order to do so (e.g., refusing to leave the house) (American Psychiatric Association, 1994).

Panic disorder has gone largely undiagnosed in children until fairly recently, and as such there is a paucity of literature in regard to specific epidemiological information such as age of onset and gender predominance. However, the disorder appears to be more common in females than in males, and there may be a link between pubertal onset and panic attacks (Kearney & Allan, 1995). Cognitive and language limitations may make it difficult for children to adequately describe the physical sensations

### 512 VII. ADVANCED TOPICS

that accompany panic attacks, making the disorder difficult to diagnose. Furthermore, children often lack the specific detail that adults use in describing these sensations. For example, children frequently use vague terms to describe panic attacks. They may express a fear of getting sick, whereas adults tend to describe fears of dying, going crazy, or losing control. As children reach adolescence, their fears become more specific, as they can describe feeling a sense of breathlessness, rapid heartbeat, and dizziness and feeling as if they are not in their own body, and some describe a fear of dying. Differential diagnostic concerns are related to the similarity in appearance between panic disorder and separation anxiety disorder, as many children and adolescents cope with their panic attacks by staying close to a family member or friend who makes them feel safe. The focus of the anxiety should be used to differentiate between the diagnoses (Albano, Chorpita, & Barlow, 1996).

## Obsessive–Compulsive Disorder

For a diagnosis of OCD to be made, a person must experience either obsessions or compulsions. Adults must realize that the

obsessions or compulsions are excessive or unreasonable in order to be diagnosed, but children do not have to meet this criteria. The obsessions or compulsions must consume 1 hour or more per day, cause the individual extreme distress, and interfere with normal functioning (American Psychiatric Association, 1994). OCD is commonly comorbid with depression, Tourette syndrome, and other anxiety disorders (Albano et al., 1996).

The expression of OCD in children is similar to that of adults. However, the content of children's obsessions and compulsions most often changes over time (Albano et al., 1996). Common themes include contamination, sexual themes, religiosity, aggressive or violent images, recurring songs or rhymes, and fear of illness. Children's compulsive behavior tends to involve handwashing, checking, and arranging items in particular orders, with washing being the most common ritual (Last & Strauss, 1989). However, children frequently engage in developmentally appropriate rituals that are not compulsions, such as arranging toys or dolls and saying goodnight to family members. These rituals do not seem excessive and are typically different in content than compulsions. Typical childhood rituals tend to dissipate at about 10 years of age and are not paired with obsessive thoughts (Swedo, Rapoport, Leonard, Lenane, & Cheslow, 1989). Again, the degree of distress a child experiences, particularly when unable to perform the ritual, is a useful indicator of pathology (Albano et al., 1996).

### **Posttraumatic Stress Disorder**

For an individual to be diagnosed with PTSD, he or she must have been exposed to a traumatic event, during which a serious threat to the safety of self or others was present. The individual's response to the experience must have been one of fear, helplessness, or horror or in children, disorganized or agitated behavior. Reexperiencing of the event must also take place and may take the form of repetitive play, dreams, flashbacks, distress when exposed to reminders of the event, and physiological symptoms upon exposure to internal or external cues of the event. A number of avoidance or numbing behaviors and increased arousal must also be identified, and the duration of the symptoms must be at least 1 month (American Psychiatric Association, 1994). The symptoms

of PTSD usually, but not always, arise within 3 months of the traumatic event. Bad dreams are particularly common in young children and may take the form of general nightmares about monsters, etc., rather than about the traumatic event. Children do not tend to reexperience the event in the same way that adults do, instead reliving the event through play. Children may not be able to report many of their symptoms, so observations by parents and teachers are crucial to evaluate hopelessness and loss of interest in previously enjoyed activities. Children may demonstrate a sense of foreshortened future by believing that they will not be around to perform adult activities. Physical symptoms, such as aches and pains, irritability, and angry outbursts, are also common features. PTSD may also be linked to the development of other anxiety disorders (American Psychiatric Association, 1994).

### 23. Assessment of Childhood Anxiety 513

#### **Acute Stress Disorder**

Acute stress disorder is diagnosed when a person has been exposed to a traumatic event in which physical harm was threatened and the individual's response included fear, hopelessness, or horror and is followed by several dissociative symptoms. The symptoms last at least 2 days but do not linger beyond 4 weeks. A primary feature of the disorder is a lack of emotional responsiveness and a loss of ability to experience pleasure. The person with acute stress disorder may experience amnesia related to the traumatic event (American Psychiatric Association, 1994).

#### **DIAGNOSTIC PROCEDURES**

Each of the anxiety disorders just described is based on a categorical diagnostic system, DSM-IV. DSM provides discrete categories based on a symptom checklist. If a person has the requisite symptoms listed in the manual, a disorder can be diagnosed. No consideration is given to individuals who fall short of the criteria by only one symptom. Although the criteria are general and allow for some variation in symptoms, individual differences that cause the presentation of the disorder to look different from one person to the next are not included in DSM descriptions. Another drawback of DSM and other categorical systems of diagnosis is that neither the singular presentation of childhood symptoms nor developmental issues are given

special consideration. The criteria for adults and children are virtually identical. Due to these factors, overlapping symptoms across diagnoses and the heterogeneity of symptom manifestation, the usefulness of categorical systems for diagnosing children has been called into question.

Dimensional systems, developed by clustering symptoms that commonly occur together, are used with increasing frequency.

Rating scales are the primary means of measurement in dimensional systems, as they produce profiles of an individual's symptoms across a broad spectrum. For example, a child may score high on one scale (e.g., anxiety) and low on another scale (e.g., aggression), providing a more complete picture of the child's functioning. In addition to allowing for individual differences in symptom expression, developmental variables may be accounted for in dimensional systems (Albano et al., 1996).

Despite dimensional systems' growing popularity, DSM is the most widely used and recognized classification system among mental health professionals and allows for ease of discussion about general psychiatric disorders. Therefore, future diagnostic systems appear to be moving toward a combined approach that ties together assessment and treatment goals (Albano et al., 1996).

### **Rating Scales**

Rating scales are used for a variety of purposes, such as to aid in diagnosis, screen for possible pathology, gather information about behavior across settings, uncover personality traits, and measure treatment progress. Rating scales are also used by many types of professionals in schools, hospitals, and clinical settings. To determine which scale to use in a particular situation, it is important to consider the goal of obtaining information and the setting in which the information will be used. Self-report rating scales are the most common method used to assess anxiety and depression in children in clinical and research settings (Silverman & Rabian, 1999). Whereas children and adolescents are thought to be the most accurate reporters of their own feelings and internal states, parents and teachers are better recorders of observable behaviors (Silverman & Rabian, 1999).

Rating scales are popular because they are inexpensive and easy to use. One of their many purposes is to screen and/or diagnose

groups of children, in which case the instrument must differentiate children with clinical levels of anxiety from those who are asymptomatic. One factor that may affect the validity of rating scales is the tendency for individuals to answer items in such a way as to make themselves appear in the most positive light. To limit this effect, administration of directions should be carefully worded to encourage honest responding, such as telling children that there are no right or wrong answers. When screening children using rating scales, the clinician must be mindful that a scale may reveal more false-positive responses than true-

#### 514 VII. ADVANCED TOPICS

positive responses, meaning that children may appear more anxious than they really are (Costello & Angold, 1988). Therefore, once a rating scale has identified a child as having clinical levels of anxiety, further investigation into the child's cognitive and socioemotional functioning, by way of interviews or other diagnostic procedures, must be conducted to support or disconfirm the results of the rating scale.

Diagnosing anxiety disorders should be accomplished through diagnostic interviews, with rating scales used as a means of obtaining severity of symptoms. Rating scales may be used to identify and quantify symptoms and behaviors once the presence of an anxiety disorder has been established (Silverman & Rabian, 1999).

A potential problem with using rating scales to measure anxiety is that rating scales do a relatively poor job of differentiating anxiety and depression, which is further complicated by the disorders' high comorbidity rates. However, it may be beneficial to look for patterns of responses that point more to worry than to depressive symptoms, such as loss of interest in pleasurable activities, lack of motivation, and low self-esteem.

Behaviors that mediate anxiety may be identified through rating scales. For example, a child's responses may indicate that he or she avoids situations that lead to anxiety. It may also be possible to use rating scales to assess other variables that are secondarily affected by anxiety, such as peer or familial relationships (Silverman & Rabian, 1999). However, the validity of using rating scales to assess treatment progress is questionable, as improved scores may not represent a reduction

in anxiety but rather reduced reporting of symptoms, or normal fluctuations in anxious thoughts and feelings (Silverman & Rabian, 1999).

### **Interviews**

The most common method used to assess child psychopathology is the interview. There are structured, semistructured, and unstructured interviews with structured interviews requiring the least training and diagnostic experience to administer. In most cases, interviews are designed to detect symptoms consistent with DSM criteria. Many interviews include both child and parent versions. Interviews are crucial in diagnosing anxiety disorders because the more information a clinician has to make a diagnostic decision, the more reliable the decision should be. However, inherent problems with interviewing children can lower the reliability of diagnosis. Perhaps the biggest problem with using interviews to assess children is the reliance of the interviews on DSM structure. As stated previously, DSM criteria do not adequately detect developmental differences in the expression of symptoms, and most DSM criteria are based on adult symptomatology. Consequently, wide variation in children's answers to interview questions may occur, which clinicians may interpret differently (Kearney & Wadiak, 1999). Another DSM-related issue affecting the reliability of interviews is that child anxiety disorders and their criteria change with each new publication of DSM (Kearney & Wadiak, 1999). Furthermore, diagnosing specific types of anxiety disorders is less reliable than diagnosing the presence of an anxiety disorder, generally.

### **Behavioral Observations**

Behavioral observations are critical in the assessment of anxiety for several reasons. First, assessment through the use of rating scales and interviews is limited by the desire of respondents to appear in a positive light or to please the administrator. In addition, anxiety symptoms are often expressed differently across settings, with some symptoms only appearing in certain environments or situations. A critical component of diagnosing anxiety disorders is determining the pervasiveness and severity of symptoms. Therefore, functional behavioral analyses should be conducted and antecedent and consequent events to the anxiety explored. Behavioral observations may also be helpful

in developing appropriate and practical treatment goals for the anxious child. In addition to functional behavior analyses, behavioral observations may take the form of behavioral approach tests (BATs), observational ratings, role-play tests, and self-monitoring. BATs are more frequently employed when specific phobia, social phobia, or generalized anxiety is suspected. A typical BAT includes brief stages, during which the child is exposed to a feared stimulus, and response measurements are taken. An example of a five-stage BAT includes an initial adaptation stage, during which the child is placed in an environment and permitted time to adapt to the setting. Next, during the baseline stage, the clinician records the child's level of fear (e.g., by measuring heart rate), then the same procedure is conducted while the child is walking and moving during the walking baseline stage. The child is then exposed to a feared stimulus for a brief period, or until the fear subsides. Finally, the child is placed in the original setting with the feared stimulus absent, and the length of time to reach baseline levels is measured. The stages of a BAT may be conducted in simulated or natural settings. Natural settings are preferred, although external validity in either setting has not been well established.

Observational rating systems primarily measure motor behaviors that are consistent with anxious responding, such as verbalizations, trembling, avoiding, poor eye contact, and body rigidity. The Behavior Assessment System for Children (BASC)—Student Observation System is one example of an observational rating system. However, further research regarding reliability and validity of these measures is required.

Role-play tests are typically used with children diagnosed with social phobia and are administered by asking the child to respond during role plays as if they were in anxiety arousing situations. The test may include videotaping so that the child's responses may be analyzed. However, reliability and validity of role-playing tests have not been demonstrated.

Self-monitoring requires the child to record his or her own anxious responses by keeping diaries of antecedents, anxiety-producing events, and consequences or by recording thoughts that are associated with

the anxiety. However, self-monitoring approaches may be limited by the child's language and cognitive capabilities and his or her willingness to comply with the procedure (Kearney & Wadiak, 1999).

### **Physiological Assessment**

Due to the increase of physiological arousal in individuals with anxiety disorders, physiological assessment may be a particularly informative and useful tool. This relatively new area of study involves the measurement of bodily responses and vital functions in the assessment of anxiety, such as measuring heart rate, blood pressure, respiration, blood volume, skin temperature, and electrical activity in tense muscles (Kearney & Wadiak, 1999). However, to date, methodology regarding the use of physiological assessment of anxiety has not been well established. Normative, reliability, and validity data on physiological assessment have not been established.

### **INSTRUMENTS**

The following section discusses a number of instruments commonly used to measure symptoms of anxiety. Although not a comprehensive list, this section provides examples of general behavior rating scales, anxiety scales, and interviews. The instruments listed vary with respect to format and respondent, and psychometric properties of each are reviewed.

#### **Behavior Assessment System for Children**

The Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) is a comprehensive diagnostic system for use with 4- to 18-year-olds, which includes a Structured Developmental History, Student Observation System, Teacher Rating Scales (TRS), Parent Rating Scales (PRS), and Self-Reports of Personality (SRP). The SDH and SOS do not have norms and cannot be readily integrated with information gathered from the rating scales but are an important inclusion in making differential diagnosis of emotional and behavioral disorders based on multiple sources of information. The TRS, PRS, and SRP provide scores based on national norms, gender within the norm group, or in comparison to a group of seriously emotionally disturbed children. In addition to clinical scales and composite scores for Internalizing and Externalizing behaviors, an Adaptive Scale is provided. The Adaptive Scale was constructed using items in concordance with current literature

and is beneficial in assessing the mentally retarded. Use of the scale in its entirety is

#### 516 VII. ADVANCED TOPICS

helpful in profiling strengths as well as weaknesses.

The BASC was developed using exceptional standardization procedures, as item analyses were carefully conducted to ensure adequate discrimination and assessed for discrepancies in performance by gender or ethnicity. Where the sample did not match the population on ethnicity, maternal education, geographical region, and special education placement, weighting was used to create unbiased norms. While the BASC also includes a clinical sample, the largest component of which included behavior disorders and attention-deficit/hyperactivity disorder, this sample was not adjusted, and white males were overrepresented. Internal consistency estimates were good for all scales with alphas ranging in the .80s and .90s, and the clinical sample has similar internal consistencies. However, alphas for individual scales were lower than for composites. For example, the PRS revealed average coefficients in the mid to upper .70s (Jones & Witt, 1994). Test-retest reliabilities, with a 1-month interval between administrations, were also in the mid-.80s to mid-.90s with the exception of some in the .70s for the PRS for adolescents. Interrater reliabilities for the TRS were somewhat low for the preschool form of the BASC, but composite correlations ranged from .69 to .89. Moderate correlations were found between the PRSs for all age groups. Latent trait analyses and factor analyses were conducted to establish the scales and composites ensuring good construct validity, and the three-factor structure of Internalizing, Externalizing, and Adaptive Skills was supported. Correlations between the TRS and several other teacher rating scales were high. The PRS was also correlated highly with the Child Behavior Checklist and moderately with the Personality Inventory for Children—Revised and the Conners Parent Rating Scales. The SRP correlations with other self-report inventories varied and indicate that the BASC-SRP may measure something different than some of the other scales. Validity of the BASC was also supported by the performance of many groups of children with previously acquired diagnoses of conduct disorder, behavior

disorder, depression, emotional disturbance, attention-deficit/hyperactivity disorder, learning disability, mild mental retardation, and autism. Convergent and discriminant validity was demonstrated. While the correlations were relatively high between the TRS and PRS, with the exception of those at the preschool level, they were only low to moderate between the SRP and PRS and between the SRP and TRS. In summary, the BASC's easy-to-use format, well-constructed manual, and sound psychometric properties make it an excellent choice in assessing a child's behavioral and emotional functioning. While the reliability and validity of the TRS and PRS are highest, the SRP is adequate, but the reading level required of 8- to 11-year-olds may limit its use with this age group (Sandoval, 1998).

#### **Personality Inventory for Children— Revised**

This rating scale, to be completed by the child's guardian, preferably the biological mother, is to be used for children ages 3–16 (PIC; Wirt, Lachar, Klinedinst, & Seat, 1984). The test was designed in the late 1970s to elicit 33 scaled scores, 16 profile scores, and 17 experimental scores describing a child's behavior, emotional and cognitive processes, and family characteristics. While the entire Personality Inventory for Children (PIC) contains 600 true/false items, it is possible to administer only certain sections of the test (e.g., eliminating the supplemental experimental scales). The format of the test is easy to understand and follow, and only a sixth- to seventh-grade reading level is required of the respondent. As the PIC was normed using biological mothers, validity may be affected if someone other than the biological mother completes the scale. Also, the form should be completed in one sitting, if possible, so that answers are not contaminated if the respondent talks to someone else about the items between test sessions. Despite the length of the PIC, clinicians should not be discouraged from administering additional instruments to other family members or to those individuals familiar with the child in order to obtain a comprehensive picture of the child's functioning.

The area of primary concern in administering the PIC relates to the test's outdated norms. The original test was normed be-

tween 1958 and 1962 in a relatively small geographical area. Although new scale items were developed in the 1970s, the old norms were applied. Furthermore, the manual lacks specific age and gender norm differences, which do not coincide with the effects of development on the characteristics that the scales purport to measure. The revised manual lists studies from the 1977 manual to support reliabilities for each scale, which ranged from .46 to .94 for a psychiatric outpatient sample, from .50 to .89 for a sample of normal children, and from .68 to .97 with another sample of normal children. One internal consistency study conducted with a heterogeneous clinic sample estimated alpha between .57 and .86 for the scales. A vast array of validity studies conducted on the PIC indicated excellent results. However, a national restandardization of the PIC is needed to determine its current diagnostic utility (Knoff, 1989).

#### **Piers–Harris Children’s Self-Concept Scale**

The Piers–Harris Children’s Self-Concept Scale (CSCS; Piers & Harris, 1969) was originally designed to measure children and adolescents’ self-concepts. The scale, which can be used with children in grades 4–12, and requires a third-grade reading level, includes 80 yes-or-no questions about how the child feels about him or herself in a self-report format. The Piers–Harris produces six cluster scores (Behavior, Intellectual and School Status, Physical Appearance and Attributes, Anxiety, Popularity, and Happiness and Satisfaction), an overall self-concept scale, a Response Bias Index, and an Inconsistency Index. The comprehensive manual, revised in 1984, is easy to follow and does a good job of cautioning the user against interpreting the scale without support from other instruments and in listing its weaknesses. Although the Piers–Harris has not been renormed since the original standardization sample, which included 1,183 students in Pennsylvania, several recent reliability and validity studies have supported its continued use. Internal consistency estimates for the Total Score ranged from .73 to .81, and test–retest reliabilities ranged from .42 to .96. Validity studies comparing the Piers–Harris with other measures of self-concept, personality, and behavior revealed moderate relationships. Factorial validity studies have resulted in conflicting results. Some studies supported the six-factor structure,

but others found additional factors or indicated factor instability. Therefore, interpretation of cluster scores may not be valid.

The ease of use, brevity, and sound reliability and validity make the Piers–Harris an attractive option for assessing self-concept and provide a base for the further investigation of anxiety related issues. However, the outdated norms and the limited geographical region in which the test was originally normed are areas that require future investigation (Epstein, 1985).

### **State–Trait Anxiety Inventory**

The State–Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1969), for ages 9–16 and adults, assess anxiety on two levels, state and trait anxiety. State anxiety is represented by the authors as an immediate sense of how one is feeling, whereas trait anxiety is a more permanent condition. Each level is measured by 20 items in counterbalanced order. The theory behind the state and trait anxiety distinction is well outlined in the manual, which is also easy to read and interpret.

The STAI was normed on 377 high school juniors, 982 college freshman, 484 introductory psychology course students, 461 male neuropsychiatric patients, 161 general medical and surgical patients, and 212 state prisoners. Separate norms are provided for males and females. Normative means and standard deviations are provided for particular psychiatric diagnoses, medical patients, and prisoners.

Internal consistency coefficients range from .83 to .92 for high school and college students' state scores and from .86 to .92 for trait scores. Test–retest reliabilities were reported as follows: State Anxiety with a 1-hour interval between test administrations was .33 for males and .16 for females, with a 20-day interval was .54 for males and .27 for females, and with a 104-day interval was .33 for males and .31 for females. Trait anxiety was .84 for males and .76 for females after 1 hour, .86 for males and .76 for females after 20 days, and .73 for males and .77 for females after 104 days. Criterion-

### **518 VII. ADVANCED TOPICS**

related validity studies indicated that the STAI college women were highly correlated with other measures. Factor analyses generally supported the distinction between state and trait anxiety (Dreger, 1978).

### **Revised Children's Manifest Anxiety Scale**

The Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Richmond, 1985) was designed to assess the level of children's anxiety across five scales; Physiological Anxiety, Worry/Oversensitivity, Social Concerns/Concentration, Total Anxiety, and Lie. The test can be used with children and adolescents from the ages of 6–19, requires a third-grade reading level, and is formatted into a 37-item self-report. The extremely well-organized manual contains information regarding the theory of anxiety on which the test is based and several case studies that demonstrate how the test was used as part of a larger psychoeducational battery. The RCMAS manual is easy to use and contains the necessary information for interpretation of each of the five scores; however, the authors caution against using the three individual anxiety scaled scores for more than hypothesis generation due to their low reliability. They do recommend using the separate norms provided according to age, sex, and ethnicity (Gresham, 1989). The large standardization sample of 4,972 included 44% white males, 44% white females, 5.8% African American males, and 6% African American females and covered a variety of geographic regions throughout the United States.

Internal consistency studies revealed alpha levels for Caucasian and African American males and females for each age level for the Total Anxiety Score in addition to standard errors of measurement for the Total Anxiety Score. Coefficient alphas ranged from .42 to .84, so interpretation for groups with lower reliability estimates should be made cautiously. Internal consistency for each of the anxiety subscales was lower than optimal for direct interpretation.

Test–retest reliability estimates were reported for the Total Anxiety Score at .68 and for the Lie Scale Score at .58 with an interval of 9 months, but when the interval was 3 weeks, test–retest reliability was reported at .98. No test–retest coefficients were provided for groups older than the seventh grade.

The authors' claim that the RCMAS is a measure of trait, as opposed to state, anxiety was supported by a validity study that indicates that the Total Anxiety Score is more highly correlated with a measure of trait anxiety than state anxiety on the STAIC (Revised Children's Manifest Anxiety Scale; Stewart, 1989). The factor analysis

produced a five factor solution, but theoretically invalid rotation procedures reported in the manual may render the five-factor structure uninterpretable (Grisham, 1989).

### **Child Anxiety Scale**

The Child Anxiety Scale (CAS; Gillis, 1980) is a self-report questionnaire that was derived from the Early School Personality Questionnaire (ESPQ) developed by Coan and Cattell in 1959. The 20-item scale can be administered to children from ages 5–12 and is easy for young children to understand due to its dichotomous response method of choosing either a red or blue circle in response to pictorial and audiotaped items. Young children may still have difficulty understanding the nature of the task, and neither a lie nor social desirability scale is provided. However, the bright and colorful nature of the scale, in addition to its brevity, make it an appealing choice for use with children.

The author gives separate age and grade norms but does not separate norms by gender, claiming that the tendency for females to score higher than males was not strong enough to warrant separate norms. The 2,105 children sample is described in the manual according to geographic region and ethnicity. The one area of concern regarding the norming sample relates to the small number of children from urban areas that were included in the norming sample. Seventy-five percent of the children were sampled from areas with populations less than 50,000, indicating that use of the instrument with children from urban settings should be interpreted cautiously. Test–retest and internal consistency coefficients were good. Test–retest coefficients for firstthrough third-graders ranged from .82 to .92, and a Kuder–Richardson 20 coefficient for internal consistency was .81. However, validity studies are severely lacking, as the manual reports no correlations between the CAS and other measures of anxiety. However, factorial validity is supported by factor analyses conducted on the ESPQ (Maxwell, 1985).

### **Multidimensional Anxiety Scale for Children**

The Multidimensional Anxiety Scale for Children (MASC; March, 1997) is a self-report measure that can be used to assess anxiety in children and adolescents ages 8–19.

The MASC is a brief measure consisting of items rated on a 0- (never true about me) to 4-point (always true about me) Likert scale, with 4 indicating the most severe emotional problems. The MASC produces a total of four factor scales, a total scale, and a validity scale, which amount to 13 subscales, including Physical Symptoms (Tense Symptoms Subscale, Somatic Symptoms Subscale, Total), Harm Avoidance (Perfectionism Subscale, Anxious Coping Subscale, Total), Social Anxiety (Humiliation Fears Subscale, Performance Fears Subscale, Total), Separation/Panic, Total Anxiety, Anxiety Disorders Index, and an Inconsistency Index. Total testing time is 15 minutes (Caruso, 2001). Although reliability and validity of the scale were generally acceptable, the reliability of some specific subscales were questionable. For example, the Perfectionism and Anxious Coping Subscales of the Harm Avoidance Scale, the Performance Fears Subscale, and the Anxiety Disorders Index all had internal consistency reliability coefficients of less than .65 in the normative sample. However, the Anxiety Disorders Index consists of items tapping several types of anxiety disorders, which may account for the heterogeneity of test items. Validity was demonstrated through confirmatory factor analysis, which supported the four factor structure. The four general scores and the Anxiety Disorders Index effectively discriminated between a random subset from the normative sample and a separate group of children with DSM-IV diagnosed anxiety disorders. Caution should be used in diagnosing OCD, however, because children with OCD were not included in the group used to develop the Anxiety Disorder Index Scale. Further validity studies are warranted to determine if the scale can differentiate between different types of anxiety disorders (Christopher, 2001).

The norming sample consisted of 2,698 children and adolescents but lacked an adequate representation of minorities, particularly Hispanic children, in relation to census proportions. The racial distribution of the sample was as follows: 53.3 % Caucasians, 39.2 % African Americans, 7% Hispanic/Latin Americans, 1.4 % Asian Americans, 2.4% Native Americans, and 3 % other. Separate norms are provided for males and females in 4-year intervals but not for racial groups. The test has a fourth-grade reading

level, and items may be read to young test takers, but no validity data regarding the ability of nonnative English speakers to respond to test items is provided (Caruso, 2001).

The MASC-10 is a short version of the original test providing a single score to be used for screening purposes. The validity of this measure was less than optimal, and internal consistency coefficients were also questionable. Overall, the MASC appears to be a useful screening tool. However, further validity studies are required to support its diagnostic value (Caruso, 2001).

#### **Anxiety Scales for Children and Adults**

The Anxiety Scales for Children and Adults (ASCA; Battle, 1993) was developed to measure the presence and level of anxiety in individuals via a self-report form. There are two forms, form Q for children and form M for adults. Both forms list items that, according to the authors, correspond to common symptoms of anxiety. The child form contains 25 dichotomously scored items, while the adult form contains 40 items from 1 (*never*) to 5 (*always*).

No mention of item analysis, internal consistency analysis, or factor analysis is made in the manual. Due to a lack of information that one would expect in the manual in regard to development of the test items, the basis on which the items were chosen (e.g., according to DSM criteria) is not apparent (Merenda, 1995). According to a review by Oehler-Stinnett (1995), no distinction is made by the authors between real or imagined anxiety-producing stimuli, state or trait anxiety symptoms, and fear, worry, or panic. Test-retest reliability is reported at

#### **520 VII. ADVANCED TOPICS**

.82 to .96, but time between testing was 1 week for the child's form and 2 weeks for the adult form. Furthermore, the manual is lacking in empirical evidence for construct validity so it is unclear how closely the test measures anxiety and if there are any specific types of anxiety that the test measures better than others. Predictive validity is not discussed in detail. Studies of concurrent validity revealed that the ASCA Form Q was correlated with the State-Trait Anxiety Inventory for Children at  $r = .64$  and the Nervous Symptoms subtest of the California Test of Personality at  $r = .66$ . Form M was correlated with the Taylor Anxiety Scale for Adults at  $r = .75$  and the Nervous Symptoms

subtest of the California Test of Personality at  $r = .63$  (Oehler-Stinnett, 1995). However, the ASCA was also highly correlated with several depression inventories, calling into question interpretability of the scores as used to measure anxiety. The manual also lacks studies regarding performance on the ASCA by age, socioeconomic status, race, or clinical populations. In summary, one should use caution in using this form for diagnostic or treatment decisions due to the lack of empirical support for the test's validity (Merenda, 1995).

### **Social Phobia and Anxiety Inventory**

The Social Phobia and Anxiety Inventory (SPAI; Turner, Beidel, & Dancu, 1996) is a self-report rating scale that assesses the presence of social phobia characteristics in individuals 14 and over. The test produces three scores, a Social Phobia Score (SP), Agoraphobia Score (Ag), and a Difference score (SP minus Ag), from a 6-point Likert scale assessing how frequently the test taker experiences anxiety in response to given circumstances.

A sixth-grade reading level is required to take the test without assistance. Although the SPAI is easy to score, the rationale for scoring methods is not fully explained in the manual. Furthermore, the authors endorse the use of clinical judgment in interpreting the pattern of responses given by examinees. For example, the test does not delineate a separate score for anxiety in particular situations, such as with strangers, but a trained clinician may glean this information by reviewing the pattern of item scores.

Of the 308 college students used in test development, 182 were used to calculate test-retest reliability, which resulted in the following scores:  $r(173) = .85$  for SP,  $r(173) = .74$  for Ag, and  $r(173) = .86$  for Difference scores, suggesting adequate stability across test administrations. Internal consistency scores were also high, as Cronbach's alpha was estimated at .96 for SP and .85 for Ag scores. Validity studies supported the use of the SPAI for identifying social phobia in individuals. Although factor analysis was performed, factor structures varied from group to group. However, confirmatory factor analyses performed by the authors supported the two-factor (SP and Ag) solution (Walcott, 2001). Furthermore, the authors' claim that the Difference score produces a purer measure of social phobia has been debated.

The apparent redundancy of test items may cause concern over the validity of interpreting scores based on these items. The authors do provide information regarding interpretation of the SPAI for different genders, ages, and races, although sample sizes on which this information was based were relatively small (Engelhard, 2001). Although the SPAI has primarily been used with adults, a newer version for children, the SPAI-C, has been developed (Walcott, 2001).

### Test Anxiety Profile

The Test Anxiety Profile (TAP; Oetting & Deffenbacher, 1980) provides a measure of test anxiety in a variety of different testing situations for individuals in grades nine through college. Two components of anxiety are assessed across six different testing situations, (multiple choice test, time-limit test, "pop" quiz, essay test, giving talk, and math test) providing 12 scores. In addition, two anxiety scores may be used to assess a student's anxiety feelings and thoughts in unique testing situations. The Feeling Anxiety (FA) score purports to measure self-perceived emotions and physiological responses to particular testing situations, and the Thought Interference (TI) score was constructed to measure a person's cognitions that may interfere with thought processes required to perform efficiently on an exam. Although the authors have studied test anxiety extensively, and the TAP appears to have adequate psychometric properties, the manual does not describe in detail the norming sample and the procedures used to construct individual items. It appears that the sample consisted of 600 Colorado State University students with reliability and validity studies based on this sample and a sample of 61 high school students. However, the internal consistency estimates reported for the college student sample ranged from .88 to .96. Furthermore, a test-retest reliability study with a 7- to 10-week interval between tests revealed coefficients ranging from .66 to .81. Criterion validity studies resulted in correlations from .28 to .59. Discriminant validity studies appeared to conform to patterns of perceived thoughts and emotions before and after test taking in the anxiety literature; however, no evidence was provided to support that the FA and TI scales measure what they are intended to

measure (Brown, 1985). In summary, the TAP appears to have adequate psychometric properties to be a useful diagnostic tool, but other factors, such as study skills and test-taking strategies should be examined as possibly influencing scores on this measure.

### **Schedule for Affective Disorders and Schizophrenia, Third Edition**

The SADS (Endicott & Spitzer, 1978) is a semistructured interview developed in the mid-1970s primarily as a research tool to identify groups of patients with the same psychiatric symptoms. The SADS was designed to decrease the unreliability of diagnostic practices by providing questions in a sequential format based on the Research Diagnostic Criteria, a forerunner to DSM-III (RDC; Spitzer, Endicott, & Robins, 1975).

The SADS takes approximately 1½ to 2 hours to administer and is divided into two parts, focusing on the current episode and past history, respectively. However, not all DSM diagnoses can be obtained by using the SADS, most notably PTSD. The interview results in a total of 24 scales including a Global Assessment Scale. The SADS has been reliable in diagnosing many disorders, but the reliability depends on the validity of measurement with each specific disorder. Concurrent validity studies comparing the diagnoses of the SADS with those of the Diagnostic Interview Schedule were weak, indicating that using the SADS for diagnostic purposes rather than as it was intended, as a research instrument, may not be useful. Intensive training and practice is recommended to use the SADS, as the format is somewhat confusing and assumes a certain level of knowledge of the subject area. Furthermore, the terminology may be outdated and translation into current symptom terminology may be necessary.

The Kiddie-Schedule for Affective Disorders (K-SADS; Kaufman, Birmaher, Brent, Ryan, & Rao, 2000) has been modified from the SADS to be used with school-age children (Carmer, 1995). The K-SADS addresses most anxiety disorders through parent and child interviews, and like the SADS, continues to focus on past and present episodes of psychopathology. Due to the unreliability of diagnosing the specific anxiety disorders, several different versions of the K-SADS have also been developed in attempts to address structural inadequacies of the interviews (Silverman, 1994).

## Anxiety Disorders Interview Schedule for DSM-IV: Child Version

The Anxiety Disorders Interview Schedule for DSM-IV: Child Version (ADIS for DSM-IV: C) was revised from the ADIS-C in 1996, which was revised from the original 1983 version of the ADIS (Silverman & Albano, 1996). The interview can be used with children ages 7–17, who are suspected of having any of the anxiety disorders listed in DSM-IV. The semistructured interviews were developed to be consistent with DSM-IV criteria, and include a child version as well as a detailed parent version. In addition to sections devoted to the anxiety disorders, separate sections that assess mood and externalizing disorders, school refusal behavior, and screening sections for substance abuse, psychosis, selective mutism, eating disorders, somatoform disorders, and learning disorders are included. According to the authors, the ADIS for DSM-IV:C is an effective instrument for use with research and clinical populations. It is recommended that both the child and parent interviews be administered to obtain a comprehensive diagnosis.

All diagnostic questions can be answered dichotomously (e.g., yes/no), and interference scores that rate the degree to which symptoms interfere with the child's

life may be given on a 0–8 scale. The child interview begins with an explanation of the “Feelings Thermometer,” which allows the child to visually rank his or her feelings of anxiety by choosing the appropriately filled thermometer.

The manual reports internal reliability estimates of between .64 to 1.00 for each of the specific anxiety disorders with an overall kappa of .75. Test–retest reliability coefficients obtained from a sample of 50 outpatients with a 10–14-day interval between test administrations ranged from .64 to .84 for individual anxiety disorders and .75 overall. The authors report that criterion validity for symptom scale scores and for symptom summary scores is good for both the parent and child interviews (Albano & Silverman, 1996.)

### CONCLUSION

To more fully understand the way in which children and adolescents experience anxiety, further research must investigate the nature of worry in typically developing children versus children who develop symptoms at

clinically significant levels. Thus far, most of the literature regarding the assessment of anxiety has been developed through investigations of adult symptoms. Even the prevalence of specific anxiety disorders and the frequency of occurrence of anxiety symptoms in children are largely unknown, and to date, extensive investigation into the progression of anxiety symptoms from childhood into adolescence and adulthood is lacking (Albano et al., 1996). Consistent with these future studies are continued explorations into optimal diagnostic methods for children and adolescents, which may include alternatives to categorical systems.

In addition to further inquiry into anxiety assessment, advances in treatment methods, including pharmacological developments, will follow new research. Currently, adrenergic, serotonergic, and GABA(4-aminobutyrate) ergic neurotransmitter systems have been explored in regard to anxiety (Bernstein, 1994). Yet once again, a paucity of literature exists that focuses on the use of anxiety medications with children and adolescents. The effectiveness of medications used to treat different childhood anxiety disorders has also not been well studied. These issues are particularly important given the increasingly high rate of anxiety disorders detected in school children. While attention and behavior problems were once thought to be the most serious childhood disturbances in school, recent research has uncovered that internalizing problems, such as anxiety disorders, may be even more prevalent. Thus, school-based approaches will be new areas of practice and research in psychology, and new advances in assessment, as well as treatment, will surely follow.

## REFERENCES

- Albano, A. M., Chorpita, B. F., & Barlow, D. H. (1996). Childhood anxiety disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 196–241). New York: Guilford Press.
- Albano, A. M., & Silverman, W. K. (1996). *Anxiety Disorders Interview Schedule for DSM-IV: Clinician manual*. San Antonio, TX: Psychological Corporation.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Anderson, J. C. (1994). Epidemiological issues. In T. H. Ollendick, N. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 43–65). New York: Plenum Press.
- Battle, J. (1993). *Anxiety scales for children and adults*. Austin, TX: Pro-Ed.
- Beidel, D. C., & Stanley, M. A. (1993). In C. G. Last (Ed.), *Anxiety across the lifespan: A developmental*

perspective (pp. 167–203). New York: Springer.

Bell-Dolan, D., & Brazeal, T. J. (1993). Separation anxiety disorder, overanxious disorder, and school refusal. *Child and Adolescent Psychiatric Clinics of North America*, 2, 583–580.

Bernstein, G. A. (1994). Psychopharmacological interventions. In T. H. Ollendick, J. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 169–186). New York: Plenum Press.

Bernstein, G. A., & Borchardt, C. M. (1991). Anxiety disorders of childhood and adolescence: A critical review. *Journal of the American Academy of Child and Adolescent Psychiatry*, 30, 519–532.

Blagg, N., & Yule, W. (1994). School refusal. In T. H. Ollendick, J. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 169–186). New York: Plenum Press.

Brown, S. D. (1985). Test Anxiety Profile. In *The ninth mental measurements yearbook* (Vol. 2, pp. 1543–1545). Lincoln: University of Nebraska Press.

**23. Assessment of Childhood Anxiety 523**

Carmer, J. C. (1995). Schedule for Affective Disorders and Schizophrenia, Third Edition. In *The twelfth mental measurements yearbook* (pp. 918–919). Lincoln: University of Nebraska Press.

Caruso, J. C. (2001). Multidimensional Anxiety Scale for Children. In *The fourteenth mental measurements yearbook* (pp. 799–800). Lincoln: University of Nebraska Press.

Christopher, R. (2001). Multidimensional Anxiety Scale for Children. In *The fourteenth mental measurements yearbook* (pp. 801–803). Lincoln: University of Nebraska Press.

Coan, R. W., & Cattell, R. B. (1959). The development of the Early School Personality Questionnaire. *Journal of Experimental Education*, 28, 143–152.

Costello, E. J., & Angold, A. (1988). Scales to assess child and adolescent depression: Checklists, screens, and nets. *Journal of the American Academy of Child and Adolescent Psychiatry*, 27, 726–737.

Dreger, R. M. (1978). State–Trait Anxiety Inventory. In *The eighth mental measurements yearbook* (Vol. 1, pp. 1094–1096). Lincoln: University of Nebraska Press.

Endicott, J., & Spitzer, R. L. (1978). *A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia*. New York: Department of Research Assessment and Training, New York State Psychiatric Institute.

Engelhard, G. (2001). Social Phobia and Anxiety Inventory. In *The fourteenth mental measurements yearbook* (pp. 1161–1163). Lincoln: University of Nebraska Press.

Epstein, H. (1985). Piers–Harris Children’s Self-Concept Scale. In *The ninth mental measurements yearbook* (Vol. 2, pp. 1168–1169). Lincoln: University of Nebraska Press.

Gillis, J. S. (1980). *Child Anxiety Scale*. Champaign, IL: Institute for Personality and Ability Testing Inc.

Gresham, F. (1989). Revised Children’s Manifest Anxiety Scale. In *The tenth mental measurements yearbook* (pp. 695–697). Lincoln: University of Nebraska Press.

Herbert, M. (1994). Etiological considerations. In T. H. Ollendick, N. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 3–20). New York: Plenum Press.

- Jones, K. M. & Witt, J. C. (1994). Rating the ratings of raters: A critique of the Behavior Assessment System for Children. *Child Assessment News*, 4, 10–11.
- Kaufman, J., Birmaher, B., Brent, D. A., Ryan, N. D., & Rao, U. (2000). K-SADS-PL. *Journal of the American Academy of Child and Adolescent Psychiatry*, 39, 1208.
- Kearney, C. A., & Allan, W. D. (1995). Panic disorder with or without agoraphobia. In A. R. Eisen, C. A. Kearney, & Schaefer (Eds.), *Clinical handbook of anxiety disorders in children and adolescents* (pp. 251–281). Northvale, NJ: Jason Aronson.
- Kearney, C. A., & Wadiak, C. (1999). In S. D. Netherton, D. Holmes, & C. E. Walker (Eds.), *Child and adolescent psychological disorders: A comprehensive textbook* (pp. 282–303). New York: Oxford University Press.
- Knoff, H. M. (1989). Personality Inventory for Children, Revised Format. In *The tenth mental measurements yearbook* (pp. 625–630). Lincoln: University of Nebraska Press.
- Last, C. G., & Strauss, C. C. (1989). Obsessive–compulsive disorder in childhood. *Journal of Anxiety Disorders*, 3, 295–302.
- Last, C. G., Strauss, C. C., & Francis, G. (1987). Comorbidity among childhood anxiety disorders. *Journal of Nervous and Mental Disease*, 175, 726–730.
- March, J. (1997). *Multidimensional Anxiety Scale for Children*. Tonawanda, NY: Multi-Health Systems.
- Maxwell, S. (1985). Child Anxiety Scale. In *The ninth mental measurements yearbook* (Vol. 1, pp. 297–298). Lincoln: University of Nebraska Press.
- Merenda, P. F. (1995). Anxiety Scales for Children and Adults. In *The twelfth mental measurements yearbook* (pp. 78–79). Lincoln: University of Nebraska Press.
- Oehler-Stinett, J. (1995). Anxiety Scales for Children and Adults. In *The twelfth mental measurements yearbook* (pp. 79–81). Lincoln: University of Nebraska Press.
- Oetting, E. R., & Deffenbacher, J. L. (1980). *Test Anxiety Profile*. Fort Collins, CO: Rocky Mountain Behavioral Science Institute.
- Piers, E. V., & Harris, D. B. (1969). *Piers–Harris Children’s Self-Concept Scale (The Way I Feel About Myself)*. Los Angeles, CA: Western Psychological Services.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children*. Circle Pines, MN: AGS.
- Reynolds, C. R., & Richmond, B. O. (1985). *Revised Children’s Manifest Anxiety Scale*. Los Angeles, CA: Western Psychological Services.
- Sandoval, J. (1998). Behavior Assessment System for Children. In *The thirteenth mental measurements yearbook* (pp. 128–131). Lincoln: University of Nebraska Press.
- Silverman, W. K. (1994). Structured diagnostic interviews. In T. H. Ollendick, N. J., King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 293–315). New York: Plenum Press.
- Silverman, W. K., & Albano, A. M. (1996). *The Anxiety Disorders Interview Schedule for DSM-IV—Child and Parent Versions*. San Antonio, TX: Psychological Corporation.
- Silverman, W. K., & Rabian, B. (1999). Rating Scales for Anxiety and Mood Disorders. In D. Shaffer & C. P. Lucas (Eds.), *Diagnostic assessment in child and*

*adolescent psychopathology* (pp. 127–166). New York: Guilford Press.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. (1970). *State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Spitzer, R. L., Endicott, J., & Robins, E. (1975). *Research Diagnostic Criteria (RDC)*. New York: Biometrics Research, New York State Psychiatric Institute.

Stewart, K. J. (1989). Revised Children's Manifest Anxiety Scale. In *tenth mental measurements year-524 VII. ADVANCED TOPICS book* (pp. 697–699). Lincoln: University of Nebraska Press.

Swedo, S. E., Rapoport, J. L., Leonard, H., Lenane, M., & Cheslow, D. (1989). Obsessive-compulsive disorder in children and adolescents: Clinical phenomenology of 70 consecutive cases. *Archives of General Psychiatry*, *46*, 335–341.

Turner, S. M., Beidel, D. C., & Dancu, C. V. (1996). *Social Phobia and Anxiety Inventory*. Tonawanda, NY: Multi-Health Systems.

Walcott, D. D. (2001). Social Phobia and Anxiety Inventory. In *The fourteenth mental measurements yearbook* (pp. 1163–1164). Lincoln: University of Nebraska Press.

Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1984). *Multidimensional description of child personality: A manual for the Personality Inventory for Children*. Los Angeles, CA: Western Psychological Services.

THE BECK SCALES IN CLINICAL PRACTICE 63  
SUMMARY AND FUTURE DEVELOPMENTS 64  
REFERENCES 64

In addition to his substantial contributions to the development and validation of cognitive theory and therapy (see Dobson & Dozois, 2001), Aaron T. Beck has, over the past 40 years, established himself firmly in the area of test construction.

Along with his colleagues, Beck has developed some of the most well known and frequently utilized self-report instruments available for research and practice. These measures cover depressive (Beck, Rush, Shaw, & Emery, 1979; Beck, Steer, & Brown, 1996; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and anxious (Beck & Steer, 1990) symptomatology, hopelessness (Beck & Steer, 1988), suicidal ideation

*Acknowledgments:* During the preparation of this chapter, David J.A. Dozois was supported by a fellowship from the Ontario Mental Health Foundation, and Roger Covin was assisted by a studentship from the Natural Sciences and Engineering Research Council of Canada. The authors gratefully acknowledge this support.

(Beck & Steer, 1991), dysfunctional attitudes (Weissman & Beck, 1978), self-concept (Beck, Steer, Epstein, & Brown, 1990), and personality (Beck, Epstein, Harrison, & Emery, 1983). More recently, the Beck Youth Inventories, which purport to evaluate emotional and social impairment in youth, have been produced (Beck, Beck, & Jolly, 2001).

This chapter describes the Beck Depression Inventory-Second edition (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). Given that the BDI-II is the most widely used of these measures, coupled with the fact that comprehensive reviews of this revised instrument have yet to appear in the literature, the primary focus of this chapter concerns the examination of the BDI-II.

However, the remaining scales that we review are used widely as well, especially in the assessment of depression. Although we do not review the Beck Anxiety Inventory (BAI), which is another of the most commonly used Beck scales, readers

are directed to some recent review papers (see Steer & Beck 1997; Wilson, de Beurs, Palmer, & Chambless, 1999).

We

begin with a review of the principal features, test development, psychometric characteristics, research status, and applicability of each of these instruments. We also discuss the limitations of these measures, mention age and cross-cultural factors, highlight accommodations made for persons with disabilities, address legal and ethical issues, and summarize each instrument's current research status. Following this examination, we underscore how these measures may be used in clinical practice.

**BECK DEPRESSION INVENTORY-SECOND EDITION (BDI-II)**

**Test Description**

The BDI-II is a 21-item self-administered inventory designed to measure the intensity of depressive symptoms in psychiatric and nonpsychiatric populations of both adults and adolescents (Beck et al., 1996). Each item contains a header that is intended to focus the examinee on the general purpose of the response options. Directly below this label are four

statements listed in order of increasing severity. Respondents are instructed to choose the alternative that best describes how they felt during the “past two weeks, including today.”

A sample item follows:

#### 5. *Guilty Feelings*

0 I don't feel particularly guilty.

1 I feel guilty over many things I have done or should have done.

2 I feel guilty most of the time.

3 I feel guilty all of the time.

Items are rated on a 4-point scale (0 to 3) and total scores are obtained by tallying the ratings for all 21 items. Scores range from 0 to 63, with higher scores reflecting increased depressive severity. For instance, scores ranging between 0 and 13 are indicative of “minimal depression”; scores that fall between 14 and 19 are considered to reflect a “mild” level of depression; scores of 20 to 28 are considered “moderate”; and a score ranging from 29 to 63 is labeled “severe.” Researchers studying dysphoria or depression in analogue samples should consult Dozois, Dobson, and Ahnberg (1998) for recommended cutoff scores for college populations. The BDI-II requires approximately 5 to 10 minutes to complete and may be administered to individuals 13 to 80 years of age. Although this instrument is typically self-administered, it can also be administered orally with only slight modification to the instructions.

#### **Theoretical Basis**

The BDI-II items were specifically selected to evaluate the symptoms and attitudes characteristic of the phenomenology of depression rather than to adhere to any particular theory (Beck et al., 1996). Additionally, although the BDI-II's items are congruent with the criteria outlined in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV;* American Psychiatric Association, 1994), the BDI-II is intended to identify the severity of symptoms and not nosological depression. Thus, the BDI-II should be supplemented with other information for a comprehensive assessment and diagnosis of depression.

#### **Test Development**

The original BDI (Beck et al., 1961) was designed to be administered in an interviewer-assisted fashion by trained professionals (Beck et al., 1996; Katz, Katz, & Shaw, 1999). The BDI items were initially drawn from clinical observations and typical descriptions of symptoms provided by depressed patients. These descriptions were refined and assembled to yield a 21-item measure with response choices ranging from 4 to 7 per item. Each statement was given a weight between 0 and 3 points. The original BDI asked respondents to describe “the way you feel today, that is, right now” (Beck & Steer, 1984). The 1978 revision, which was published as the BDI-IA in Beck et al. (1979), permitted simpler administration and scoring (see Beck & Steer, 1984). For example, the items were standardized so that they would each involve only four possible choices, alternative ways of asking the same questions were eliminated, and the language of the items was clarified (e.g., the use of double negatives was

avoided). The BDI-IA was designed as a self-report index and the temporal focus was on the “past week, including today.”

The BDI-IA demonstrated adequate reliability and validity (see Beck, Steer, & Garbin, 1988, for an elaborate review). However, it became apparent that this instrument did not correspond adequately to current diagnostic symptom criteria, and questions were raised regarding its content validity. For example, the BDI-IA covered only six of the nine symptoms highlighted in *DSM-IV*. In addition, this instrument only permitted the assessment of insomnia and decreases in appetite and weight rather than reversed neurovegetative symptoms (Moran & Lambert, 1983; Vredenburg, Krames, & Flett, 1985). Several changes were made in the BDI-II to increase its correspondence with *DSM-IV*: Four items (i.e., “body image change,” “work difficulty,” “weight loss,” and “somatic preoccupation”) were eliminated, 17 response options were reworded, 2 items were relocated, 4 new items were constructed (i.e., “agitation,” “worthlessness,” “loss of energy,” and “concentration difficulty”), item labels were provided to make the intention of each item more explicit, and the time frame was extended 2 weeks (see Beck, Steer, Ball, & Ranieri, 1996; Beck et al., 1996).

### **Psychometric Characteristics**

The BDI-II represents a significant change to earlier editions of this instrument. Therefore, it is important for researchers and clinicians to be familiar with the psychometric properties of this particular instrument and to be acquainted with how this measure corresponds to its previous editions.

#### **Reliability**

A number of studies have now documented that the BDI-II exhibits high internal consistency. Table 5.1 presents the coefficient alphas across 13 studies. The average coefficient alpha was .91 (range  $\_$  .89 to .94). Regardless of the population investigated, the internal reliability of the BDI-II appears to be excellent.

There is a paucity of information on the test-retest reliability of the BDI-II. A 1-week test-retest reliability coefficient of .93 ( $n$   $\_$  26 outpatients) was reported in the test manual (Beck et al., 1996). Aside from this finding, it is necessary to rely on previous research using the BDI and BDI-IA to address the temporal stability of this instrument. Beck, Steer, and Garbin (1988) reported the findings from 10 different studies. The test-retest reliability estimates ranged from .48 to .86 in psychiatric patients and from .60 to .83 in nonpsychiatric samples. The assessment periods varied widely in this review from a few hours to a few months. However, the higher overall correlation in nonclinical, relative to clinical, samples and the fact that higher correlations are found when shorter test-retest periods are used support the reliability of this measure (Beck, Steer, & Garbin, 1988; Richter, Werner, Heerlein, Kraus, & Sauer, 1998). Across 20 studies, Yin and Fan (2000) found an average test-retest reliability coefficient of .72.

The use of test-retest reliability on a measure that is supposed to measure a construct reliably but also be sensitive to

treatment change is difficult. On one hand, a short temporal period between assessments may overestimate reliability because participants are better able to remember how they responded previously. On the other hand, reliability may be underestimated when using a longer time frame between assessments, because true changes in depressive symptoms may have occurred (e.g., improvement due to treatment; see Richter et al., 1998). Notwithstanding the fact that the BDIII should be both sensitive to clinical change and reasonably stable over time (Boyle, 1985), a number of researchers have reported that BDI/BDI-IA scores decrease significantly upon reassessment even without the introduction of treatment (Ahava, Iannone, Grebstein, & Schirling, 1998; Hatzenbuehler, Parpal, & Matthews, 1983; Yin & Fan, 2000; Zimmerman, 1986). Ahava et al. (1998) tested the stability of the BDI over multiple assessment periods and found a 40% reduction in scores over 8 weeks. These authors argued that this reduction was due to measurement error rather than authentic changes in depressive severity. It is possible that the BDI-II will show greater temporal stability because of the increased time frame in the instructions, but this remains to be demonstrated empirically.

**TABLE 5.1 Internal Consistency Estimates for the Beck Depression Inventory-II**

Reference Sample

Coefficient

Alpha

*Psychiatric*

Beck, Steer, Ball, &

Ranieri (1996)

140 adult outpatients 0.91

Beck, Steer, & Brown

(1996)

500 adult outpatients 0.92

Buckley et al. (2001) 416 substance-abusing males 0.91

Steer, Beck, & Brown

(1997)

210 adult outpatients 0.92

Steer et al. (1999) 210 adult outpatients 0.90

Steer, Clark, Beck, &

Ranieri (1999)

840 adult outpatients 0.92

Steer, Kumar, Ranieri, &

Beck (1998)

210 adolescent outpatients 0.92

*Nonpsychiatric*

Arnau et al. (2001) 340 primary care patients

**Validity**

The content validity of the BDI-II appears to be excellent.

The BDI-II now covers the major content domains of depression, including sadness, pessimism, beliefs of being a failure, loss of pleasure, feelings of guilt, punishment feelings, self-dislike, self-criticalness, suicidal thoughts or wishes, experiences of crying, agitation, anhedonia, indecisiveness, feelings of worthlessness, lack of energy, altered sleep patterns (i.e., hypersomnia and insomnia), irritability, increases or decreases in appetite, concentration difficulties, fatigue,

and loss of interest in sex.

The convergent and divergent validity of the BDI-II also appears to be well supported. The BDI-II correlates significantly with other indices of depression and depression-related constructs, including the BDI-IA ( $r = .93$ ; Beck et al., 1996; Dozois et al., 1998), the Hamilton Rating Scale for Depression ( $r = .71$ ), and the BHS ( $r = .68$ ; Beck et al., 1996).

BDI-II scores also correlate more highly with measures of depression than with measures of anxiety (Beck et al., 1996; Osman et al., 1997; Steer, Ball, Ranieri, & Beck, 1997). For example, Steer et al. (1997) found that the BDI-II was more strongly associated with the Depression subscale of the Symptom Check List-90-Revised ( $r = .89$ ) than with the Anxiety subscale of this same instrument ( $r = .71$ ). The divergent validity of the BDI-II is also upheld by low correlations between this instrument and age, sex, ethnicity, and social desirability (Beck, Steer, Ball, & Ranieri, 1996; Osman et al., 1997; Steer & Clark, 1997). One criticism of the BDI-II that also applies to many depression assessment instruments (see Dozois & Dobson, 2002) is that this measure correlates highly with other measures of anxiety and may not discriminate adequately between depression and other affective states. Lovibond and Lovibond (1995) argued that this problem may be a result of item overlap. However, given the high rates of comorbidity between depression and other emotional disorders, it is difficult to ascertain whether this inability to differentiate between such groups is a function of the measure itself or the construct being assessed (e.g., the heterogeneous nature of depression and the polythetic criteria in the *DSM-IV*).

The BDI-II does appear to differentiate well between depressed and nondepressed persons (Ambrosini, Metz, Bianchi, Rabinovich, & Undie, 1991; Arnau, Meagher, Norris, & Bramson, 2001; Beck et al., 1996; Martinsen, Friis, Hoffart, 1995). Other research has demonstrated that the BDI-II is also able to distinguish among varying levels of depressive severity (Steer, Brown, Beck, & Sanderson, 2001), and between mood disorders and other forms of psychopathology, including anxiety (Beck et al., 1996). Not surprisingly, this instrument does not differentiate among varying types of mood disorders (e.g., major depressive disorder and dysthymia; Richter et al., 1998). Although the BDI-II coincides with *DSM-IV* symptomatology, it was intended to be used as an index of severity, not necessarily as an indicator of nosological depression. However, further study of the diagnostic specificity of this instrument is warranted (see Dori & Overholser, 2000).

Evidence for the construct validity of the BDI-II also stems from factor analytic studies. Table 5.2 displays the number of factors described in the manual and by nine additional studies. As shown in this table, a stable factor structure exists for the BDI-II (Arnau et al., 2001; Beck et al., 1996; Buckley, Parker, & Heggie, 2001; Dozois et al., 1998; Steer, Ball, Ranieri, & Beck, 1999; Steer & Clark, 1997; Steer, Kumar, Ranieri, & Beck, 1998; Whisman, Perez, & Ramel, 2000). With few exceptions, two main factors appear

to emerge consistently in the literature. In clinical samples, the two factors typically represent the somatic-affective and cognitive aspects of depression. A similar factor structure is found in nonclinical samples, but the affective items appear to load more consistently on the cognitive than the somatic factor.

***Comparability of the BDI-II to the BDI***

The BDI-II appears comparable to its earlier versions in terms of reliability, but it is a clearly superior instrument in terms of its validity (Dozois & Dobson, 2002; Dozois et al., 1998). Beck and Steer (1984) found that the original (1961) version and 1978 (BDI-IA) revision yielded coefficient alphas of .88 and .86, respectively (also see Beck, Steer, & Garbin, 1988). The average internal consistency for the BDI-II is somewhat higher (α = .91). Earlier factor analytic studies of the BDI revealed that a three-factor solution (Negative Attitudes Toward Self, Performance Impairment, Somatic Disturbance) was most frequently identified in the literature (Beck, Steer, & Garbin, 1988). However, the number of factors extracted ranged anywhere from one to seven and the average number of factors was 3.96 (*SD* = 1.91) (see Beck, Steer, & Garbin, 1988). Conversely, research conducted on the BDI-II indicates that a stronger and more stable factor structure exists than for the BDI and BDI-IA (Beck et al., 1996; Dozois et al., 1998; Steer et al., 1999; Steer, Kumar, et al., 1998). Beck et al. (1996) noted that “the transition from the usage of the BDI-IA to that of the BDI-II should introduce no meaningful interpretative problems” (p. 596). Although the mean BDI-II score is approximately 2 points higher than the BDI-IA (Beck et al., 1996; Dozois et al., 1998), a similar relationship to other inventories is demonstrated (Beck et al., 1996) and conversions are available in the test manual. Many of the limitations of the BDI appear to have been resolved with the 1996 revision, making the BDI-II an even stronger instrument than its earlier versions.

**54 The BDI-II, BHS, and BSS**  
**TABLE 5.2 Factor Analytic Studies of the Beck Depression Inventory-II**

Reference Sample Method # Factors Factor Labels

*Psychiatric*

Beck, Steer, & Brown (1996) 500 outpatients (53% with mood disorders) Principal factors 2 Somatic-Affective  
Cognitive

Buckley et al. (2001) 416 male substance abusers CFA 3 Cognitive  
Affective

Somatic

Steer et al. (1999) 210 depressed outpatients Principal factors 2 Somatic-Affective  
Cognitive

(same sample) CFA 2 Cognitive

Noncognitive

Steer, Kumar, et al. (1998) 210 outpatient adolescents Principal factors 3 (2)a Cognitive  
Somatic-Affective

*Nonpsychiatric*

Arnau et al. (2001) 340 primary care patients Principal components 2 Somatic-Affective  
Cognitive

Beck, Steer, & Brown (1996) 120 college students Principal factors 2 Cognitive-Affective  
Somatic

Dozois et al. (1998) 511 college students Maximum likelihood 2 Cognitive-Affective  
Somatic-Vegetative

511 college students CFA 2 Cognitive-Affective  
Somatic-Vegetative  
Osman, Downs, et al. (1997) 230 college students Maximum likelihood 3 Negative Attitudes  
Performance Difficulty  
Somatic Elements  
Steer & Clark (1997) 160 college students Principal factors 2 Cognitive-Affective  
Somatic  
Whisman et al. (2000) 576 college students CFA 2 Cognitive-Affective  
Somatic  
aOnly two factors were generalizable; CFA \_ confirmatory factor analysis.

### **Range of Applicability and Limitations**

The BDI-II and its predecessors have been used extensively in research and practice and are among the most frequently used psychological tests to date (Camara, Nathan, & Puente, 2000). A number of populations have been studied over the years using the BDI scales, including different psychiatric groups, nonclinical (analogue, undergraduate, and community) samples, ethnic groups, medical populations, and age groups. This instrument is an extremely useful research tool and is also a clinically sensitive instrument that may be used for determining a baseline level of severity, formulating clinical hypotheses, deriving a case conceptualization, monitoring session-by-session treatment change, and determining treatment outcome (see Dozois & Dobson, 2002).

Nonetheless, there are a number of limitations of this instrument that need to be considered. The liabilities include difficulties at the individual item level, its limited ability to detect deviant response sets and styles, the instability of scores over time, the lack of normative information for different ethnic groups, and the potentially premature development of BDI-II derivatives.

Most of the BDI-II items and response options discriminate among individuals who differ in their severity of depression. In the BDI-II manual, Beck et al. (1996) reported item-response curves for the BDI-II that indicated that the majority of its items map appropriately onto the construct of depression. However, the item weights did not perform as expected for four items (“punishment feelings,” “suicidal thoughts or wishes,” “agitation,” and “loss of interest in sex”). For instance, on some of these items severely depressed individuals are more likely than less depressed persons to endorse statements that represent lower rather than higher a priori item weights (i.e., a score of 1 or 2 rather than 3). Similar findings were reported by Santor, Ramsay, and Zuroff (1994) for the BDI-IA.

Concern has also been expressed that women tend to score higher on the BDI-II than do men (Beck et al., 1996; Steer, Kumar, et al., 1998), which might imply that the items are biased or that different psychometric properties exist across gender. However, item analyses using item-characteristic curves demonstrate that few biases exist. For example, although some BDI-IA items were shown to be somewhat biased by gender (“punishment,” “crying,” and “body image change”), for the most part equally depressed males and females do not respond differently on the BDI-IA (Santor et al., 1994). Moreover, the item concerning body image was

dropped in the most recent revision of the BDI. Because the gender difference in total scores was not present with the BDI-IA, however, research needs to determine whether separate cutoffs and norms are necessary (Steer, Kumar, et al., 1998).

Another limitation of the BDI-II pertains to the lack of established criteria for determining the accuracy of an examinee's response. Some researchers have pointed out that very low BDI scores may reflect social desirability. Clark, Crewdson, and Purdon (1998), for example, found that participants whose total score was 0 or 1 were characterized by a positive impression management style. Although possibly less problematic in clinical settings, Clark et al. suggested that for research purposes extremely low-scoring participants should be excluded from nondepressed control groups (also see Kendall, Hollon, Beck, Hammen, & Ingram, 1987). In addition to low-end specificity (i.e., the degree to which low scores are truly indicative of the absence of psychopathology), Kendall et al. (1987) also raised the issue of high-end specificity (i.e., the degree to which high scores reflect nosological depression as opposed to other clinical conditions).

The BDI-II may differentiate depression from other conditions (Beck et al., 1996), but high scores do not necessarily imply specificity to depression. Thus, accessing other sources of information is essential for appropriate diagnosis.

There are limited data on the test-retest reliability of the BDI-II, but what is available suggests that this instrument is moderately stable over short periods of time yet also sensitive to clinical change. The BDI is an excellent measure for monitoring session-by-session changes in therapy and for the assessment of treatment outcome (see Dozois & Dobson, 2002).

The issue of repeated assessment and the extent to which decreases in BDI-II scores are artifactual (e.g., due to error variance) or genuine has, however, been contested of late (e.g., Ahava et al., 1998; Yin & Fan, 2000). Because this problem is not unique to the BDI-IA and BDI-II (cf. Sharpe & Gilbert, 1998), we discuss this issue further in a subsequent section of the chapter.

Another limitation of the BDI-II is that there are not adequate norms for diverse ethnic groups. The majority of psychometric studies on the BDI-II have used Caucasians. In the BDI-II studies reviewed in this chapter, the average sample composition was 80% Caucasian (range \_ 37% to 97%). Although some population-specific norms have been developed for the BDI-II, appropriate norms for different ethnic groups are needed (Buckley et al., 2001; O'Hara, Sprinkle, & Ricci, 1998).

Finally, a recent criticism of the BDI-II involves the development of controversial spin-offs. There are many derivatives of the BDI-II, including the BDI 13-item short form (Leahy, 1992), the Beck Depression Inventory for Primary Care (BDI-PC), and the BDI for youth. The psychometric properties of these alternative forms appear to be good (e.g., Leahy, 1992; Reynolds & Gould, 1981) but there are limited data on some of these measures and questionable utility for others. For example, there are no published data on the Beck

youth scales aside from what is reported in the test manual (J.S. Beck, personal communication, November 2001). The BDI-PC, which has been marketed by Psychological Corporation as the BDI-Fast Screen, is similarly problematic.

This instrument consists of seven items (sadness, pessimism, past failure, loss of pleasure, self-dislike, self-criticalness, and suicidal thoughts or wishes) from the BDI-II. Beck and his colleagues developed the BDI-Fast Screen in response to the confound that exists in primary care patients between somatic items and physiological problems: "The average rate of specificity may be lower because the somatic and performance symptoms of depression, which are contained in most self-report measures of depression, overlap with the types of somatic and performance symptoms that occur in medical illnesses" (Beck, Guth, Steer, & Ball, 1997, p. 785).

The rationale for developing the BDI-Fast Screen appears reasonable and the test seems to be reliable and valid in primary care populations of adults (Beck, Guth, et al., 1997; Steer, Cavalieri, Leonard, & Beck, 1999) and adolescents (Winter, Steer, Jones-Hicks, & Beck, 1999). However, it is probable that the developers of this derivative were too hasty in excluding somatic-related items in the development of this instrument. Arnau et al. (2001), for instance, argued that there are good reasons not to exclude somatic items when assessing depression in medical patients. These items may overlap with medical problems but nonetheless contribute importantly to the overall prediction of depressive severity. In a sample of primary care patients, Arnau et al. found that the receiver operating characteristics of the BDI-II full scale were excellent. Other researchers have also found that the omission of somatic items does not enhance predictive ability in this population (Aikens et al., 1999; Geisser, Roth, & Robinson, 1997; Levin, Llabre, & Weiner, 1988; Lustman, Clouse, Griffith, Carney, & Freedland, 1997). Deleting items in order to increase specificity often results in a scale's decreased sensitivity (Aikens et al., 1999). Given these criticisms, researchers and clinicians should be cautious and ensure that supplementary data are collected if they choose to use the BDI-Fast Screen.

### **Age and Cross-Cultural Factors**

The BDI appears to be appropriate for research with elderly populations and as a clinical screening instrument (Gallagher, 1986; Gallagher, Nies, & Thompson, 1982; Keane & Sells, 1990). Some of the issues that have been identified in using the BDI with older adults include the readability of the items, possible social desirability (e.g., elderly patients tend to underreport subjective distress), and whether somatic complaints are the best indicators of depression in the elderly because of their ubiquitous nature (Gallagher, 1986). Some patients may find multiple-choice questionnaires such as the BDI-II too cumbersome to complete, particularly if they are severely impaired. Clinicians may therefore opt to use questionnaires that are less complicated (cf. Dozois & Dobson, 2002). Because the BDI-II also contains somatic items, clinicians who choose to use this measure in elderly or medical populations would be prudent to follow up with questions about the etiology

of such complaints (e.g., whether they pertain more to one's physical problem or affective status; see Gallagher, 1986). Conversely, there are some data that suggest that the inclusion of the somatic items does not result in a biased estimate of depression in the elderly. Laprise and Ve'zina (1998) examined the receiver operating characteristic curves of French translations of both the BDI and the Geriatric Depression Scale and found no differences in their diagnostic performance indices. Moreover, as the research we reviewed earlier demonstrated, the exclusion of the BDI's somatic items may not improve the instrument's performance. The BDI scales have been translated into several languages, including Spanish (Bonicatto, Dew, & Soria, 1998), French (Byrne, Baron, & Campbell, 1994), Chinese (Chan, 1991; Shek, 1990), Portuguese (Gorenstein, Andrade, Filho, Tung, & Artes, 1999), Dutch (Bosscher, Koning, & Van Meurs, 1986), Persian (Hojat, Shapurian, & Mehryar, 1986), Arabic (Abdel-Khalek, 1998; West & Al-Kaisi, 1995), and Hmong (Mouanoutoua, Brown, Cappelletty, & Levine, 1991). In general, these translated versions show psychometric properties that are comparable to the untranslated version. A Chinese translation of the BDI-IA, for example, was found to be reliable, valid, and quite commensurate with the English version (Chan, 1991). Nontranslated versions of the BDI also appear to be reliable and valid in English-speaking minority groups (e.g., Gatewood-Colwell, Kaczmarek, & Ames, 1989). Across languages, the BDI has acceptable reliability and validity (Bonicatto et al., 1998). Simply because a measure seems appropriate for some minorities or cultural groups does not, however, necessarily imply that it will be equally valid and reliable in other similar groups. Furthermore, direct translations may not produce similar reliability at the item level. For example, the item "loss of libido" may be correct semantically but has been shown to lack content validity in Chinese cultures (see Zheng & Lin, 1991; Zheng, Wei, Lianggue, Guochen, & Chenggue, 1988). Clinicians and researchers should ensure that the BDI scale used is appropriate for their particular clientele or population, taking into account their level of identification and acculturation with a given culture.

#### **Accommodation for Populations with Disabilities**

The reading difficulty of the BDI-II is quite low, which makes this task easy to understand and use. Assuming that there are no severe language disabilities or thought disorders, the BDI can be reliably administered to educationally and developmentally delayed individuals (Beck, Carlson, Russell, & Brownfield, 1987). The BDI-II may also be administered orally for individuals with reading impediments or severe concentration difficulties (Beck et al., 1996).

#### **Legal and Ethical Considerations**

As mentioned previously, the BDI-II should not be used as the sole determinant of an individual's diagnosis of depression. First, the BDI-II was not intended to diagnose depression but to serve as an index of symptom severity. Second, the BDI-II is not comprehensive enough to provide conclusive diagnostic information. For example, the *DSM-IV* exclusionary criteria are not included in the BDI-II nor is the

requirement that respondents endorse at least one of the sadness and anhedonia questions. As such, the BDI-II should be used to determine a patient's symptom severity, to monitor the efficacy of treatment over time, or to suggest the need for a more thorough assessment.

Clinicians and researchers should be aware that some of the BDI-II items are related to an increased risk of suicidality.

In particular, patients or participants who score higher than 1 point on items 2 ("pessimism") or 9 ("suicidal thoughts or wishes") should be evaluated for suicide potential (see Beck et al., 1996). Moreover, because hopelessness and the risk of suicide are part of the nature of depression, clinicians working in this area are encouraged to be familiar with local laws and procedures regarding involuntary hospitalization.

### **Computerization**

Scannable record forms of the BDI-II are available and a computer-administered version of the BDI-IA has been examined (Steer, Rissmiller, Ranieri, & Beck, 1994). Few empirical studies have evaluated the utility or cost efficiency

of this measure relative to the standard administration. Steer et al. (1994) examined computerized versions of the BDI and BHS in a sample of patients with mixed psychiatric diagnoses.

Both instruments yielded results that concur with the published literature on paper-and-pencil administrations. Research is necessary, however, to directly test the relationship

between computer- and questionnaire-administered measures.

Four of the Beck scales may also be scored and interpreted simultaneously through the Beck InterpreTrak\_. According to information provided by Psychological Corporation, this software program permits clinicians to track patient progress on the BDI-II, BHS, BAI, and BSS. Item responses are analyzed by this program and an interpretive summary is provided.

In addition, this software generates session-by-session graphs of symptom change to allow clinicians to monitor treatment outcome. Given that InterpreTrak\_ was released in 2000, it is presently unclear whether this software corrects for repeated administrations or provides additional information (e.g., suggestions for the timing of treatment termination or for the prevention of relapse) that would not be easily available elsewhere.

### **Current Research Status**

The BDI-II has numerous assets that make it an excellent choice for both research and practice (see Dozois & Dobson, 2002). These strengths include the BDI-II's consistency with *DSM-IV* criteria, its excellent psychometric properties, the ease of administration and scoring, its sensitivity to treatment change, and its large empirical database with which to compare results. Some of the limitations of the BDI-II include its inability to provide conclusive diagnostic evidence, the potentially inaccurate weighting of some item statements (as identified via item-characteristic curves), problems with high and low-end specificity, the instability of scores over time even among nonclinical samples, inadequate normative data for different ethnic groups, and the potentially premature development of BDI-II derivatives. Many of these limitations are not uncommon among self-report instruments and have

to do more with the *use* of the BDI-II than with the instrument per se. As such, this inventory will likely remain within the top ten most frequently used psychological tests (Camara et al., 2000).

Existing research on the BDI-II suggests that this instrument is comparable and yet superior to its earlier editions.

Nonetheless, there are a number of important research directions that should be pursued, including: (1) determining the utility of the short forms of the BDI-II for screening purposes—for example, how do these instruments compare to other screening instruments in terms of their overall correct classification rates?; (2) assessing whether the exclusion of somatic items actually enhances or impairs the predictive utility of the BDI-Fast Screen; (3) developing norms for different ethnic groups; (4) evaluating whether different norms are required for men and women separately; (5) testing the utility and psychometric properties of the Beck Youth scales; (6) investigating the extent to which the Beck InterpretTrak\_ enhances clinical conceptualization, the provision of feedback to patients, the evaluation of treatment outcome, and clinical decision making; and (7) examining how the BDI-II can be used to test clinical hypotheses regarding treatment choice, treatment outcome, and prevention of relapse. As there are a host of other important research questions that may be addressed, the recommendations we provide are not intended to be exhaustive, but rather to serve as a springboard for further empirical investigation.

### **BECK HOPELESSNESS SCALE (BHS)**

#### **Test Description**

The BHS is a 20-item true-false questionnaire that is easy to administer and score. Nine of the items are false keyed to control for acquiescent response styles. Each item is scored either 0 or 1, with total scores ranging from 0 to 20. Higher scores reflect more intense levels of hopelessness. Overall, the content of the items represents negative expectations for the future. For example, item 2 states “I might as well give up because I can’t make things better for myself,” (true-keyed response) and item 13 states “When I look ahead to the future, I expect I will be happier than I am now” (false-keyed response). Generalized pessimism regarding the future is the main focus of the BHS and appears to account for the majority of response variance in factor analytic studies.

#### **Theoretical Basis and Test Development**

Based on previous theory implicating hopelessness with depression and suicide (e.g., Beck, 1963), Beck, Weissman, Lester, and Trexler (1974) developed and tested a measure that would allow the construct of hopelessness to be examined quantitatively. Although hopelessness is somewhat of an abstract concept, Beck et al. (1974) followed the suggestions of Stotland (1969) and defined hopelessness objectively in terms of a person’s negative expectancies for the future. Several stages were involved in the development of this scale. First, 9 of the 20 items were adapted from a previous test of attitudes toward the future (Heimberg, 1961, as cited in Beck & Steer, 1988). The remaining items were derived from pessimistic statements about the future made by psychiatric

patients who were previously rated by clinicians as exhibiting significant hopelessness. Next, the scale was administered to depressed and nondepressed patients and feedback was obtained regarding the appropriateness of each item to the construct. Finally, several clinicians were asked to rate the scale on its face validity and comprehensibility. Changes in wording were made based on the suggestions of patients and clinicians.

### **Psychometric Characteristics**

The reliability of the BHS is well supported, primarily with clinical populations. In adult psychiatric patients, internal reliability coefficients (Kuder-Richardson-20) range from .83 to .93 and are often around .90 (Beck, Steer, & Brown, 1993; Beck, Steer, Beck, & Newman, 1993; Durham, 1982; Mendonca, Holden, Mazmanian, & Dolan, 1983; Steer & Beck, 1997; Steer, Beck, Brown, & Beck, 1993; Young, Halper, Clark, & Scheftner, 1992). Reliability estimates are equally strong with adolescent psychiatric patients (Kumar & Steer, 1985; Steer, Kumar, & Beck, 1993a). In contrast to research on psychiatric samples, the internal reliability of the BHS is typically lower and more variable in nonpsychiatric populations (Durham, 1982; Holden & Fekken, 1988; Rew, Taylor-Seehafer, Thomas, & Yockey, 2001; but see Johns & Holden, 1997). It is also possible that some individual test items may lower the overall reliability of the scale (Holden & Fekken, 1988; Steed, 2001; Steer & Beck, 1997; Young et al., 1992). In a nonpsychiatric sample, Steed (2001) found that items 4, 5, 8, and 13 exhibited especially low item-total correlations and recommended excluding these items when testing this population. Similarly, Young et al. (1992) asserted that while the BHS items seem to tap the construct of hopelessness well, they do so only for persons who exhibit moderate to severe levels of hopelessness. Congruent with the weaker internal reliability coefficients found in nonclinical groups, the BHS appears to be less reliable in individuals who display low levels of hopelessness.

Test-retest reliability of the BHS appears to be high in nonpsychiatric students. Holden and Fekken (1988), for instance, reported a 3-week test-retest reliability coefficient of .85. For clinical groups, these estimates are lower, which may reflect genuine clinical change. In mixed psychiatric samples, Beck and Steer (1988) reported test-retest reliability coefficients of .69 and .66 over 1 and 6 weeks, respectively.

Substantial support exists for the validity of the BHS. This instrument shows robust relationships with related constructs, across all types of populations. Among adult clinical patients, the BHS correlates significantly with depression symptoms (Beck et al., 1993; Dyck, 1991; Strosahl, Chiles, & Linehan, 1992; Wilkinson & Blackburn, 1981; Winefield, 1979), suicidal ideation (Beck et al., 1993; Dyck, 1991, Study 3; Ellis, 1985; Mendonca & Holden, 1998; Wetzel, Margulies, Davis, & Karam, 1980), and suicide intent (Strosahl et al., 1992; Wetzel, 1976; Wetzel et al., 1980). In addition, the BHS correlates with suicide intent even after depressive severity has been partialled out (Beck & Steer, 1988; Wetzel et al., 1980). The BHS also correlates negatively and significantly ( $r$  \_

\_.63) with the Reasons for Living Inventory (RFL; Dyck, 1991), a measure developed to assess people's reasons for staying alive when pondering suicide. Further evidence for the validity of the BHS emerges from findings that this measure distinguishes between psychiatric and nonpsychiatric groups and between ideators and nonideators. For example, suicide attempters and psychiatric inpatients display significantly higher BHS scores than do nonclinical controls (Durham, 1982; Simonds, McMahon, & Armstrong, 1991). Moreover, depressed patients exhibit higher scores than nonpsychiatric controls, recovered depressed patients, and neverdepressed controls (Hamilton & Abramson, 1983; Wilkinson & Blackburn, 1981). After controlling for initial depression severity and length of hospital stay, McCranie and Riley (1992) found that pretreatment hopelessness scores were significant predictors of depressive severity 3 weeks later. Finally, the BHS also differentiates between suicide ideators and nonideators (Beck et al., 1993). Hopelessness, in fact, appears to be a better predictor of suicide intent than is depression (Beck & Steer, 1988; Wetzel, 1976; Wetzel et al., 1980).

The validity of the BHS is also supported in nonclinical groups. In these samples, the BHS correlates significantly with severity of depressive symptoms (Dixon, Heppner, Burnett, & Lips, 1993; Dyck, 1991; Johns & Holden, 1997; Joiner & Rudd, 1995; Prociuk, Breen, & Lussier, 1976; Rudd, 1990; Thackston-Hawkins, Compton, & Kelly, 1994; Weber, Metha, & Nelsen, 1997; Whatley & Clopton, 1992; Wilkinson & Blackburn, 1981), suicidal ideation (Dyck, 1991; Johns & Holden, 1997; Rudd, 1990; Weber et al., 1997; Whatley & Clopton, 1992), and suicide intent (Dyck, 1991), and negatively with measures of hope (Herth, 1991; Obayuwana et al., 1982). It is important to note, however, that the validity of the BHS is limited in this population because of its questionable reliability.

There exists support for the validity of the BHS in adolescent samples. This instrument correlates significantly and positively ( $r = .68$ ) with the BDI (Johnson & McCutcheon, 1981) and negatively ( $r = -.64$ ) with the RFL-A (an adolescent version of the RFL; Gutierrez, Osman, Kopper, & Barrios, 2000). In fact, Kumar and Steer (1995) found that the BHS was a more powerful predictor of suicide ideation than 12 other variables, including history of sexual abuse, current suicide attempt, past suicide attempt, ethnicity, and diagnosis of a mood disorder. The BHS also appears to predict suicide ideation better than both the BDI and BAI in adolescent inpatients (Steer, Kumar, & Beck, 1993a, 1993b).

Although this evidence is encouraging, some studies raise concern about whether the BHS is as valid for adolescents as it is for adults. To illustrate, the BDI and BHS correlated highly ( $r = .75$ ) in a sample of female, adolescent suicide attempters, but neither instrument was significantly associated with suicide intent (Rotheram-Borus & Trautman, 1988). Furthermore, Gutierrez et al. (2000) found that the BHS did not significantly differentiate those individuals who never seriously considered suicide from those who attempted.

The predictive utility of the BHS has been examined across

a number of studies. Beck, Brown, Steer, Dahlsgaard, and Grisham (1999) used receiver operating characteristic (ROC) analyses to identify cutoff scores that provided the best prediction of future suicide. In these analyses high sensitivity was understandably considered more important than high specificity in identifying potential suicide attempters. With a cutoff of 8 and higher representing a high-risk group, the BHS was highly sensitive (90%) but limited in specificity (42%). Individuals scoring at or above this threshold were six times more likely to commit suicide than those scoring below the cutoff. These findings concur with those of Beck, Brown, Berchick, Stewart, and Steer (1990), who found that a cutoff score of 9 yielded 94% sensitivity and 41% specificity. In this study, individuals in the high-risk group were 11 times more likely to commit suicide than those in the low-risk group. Using a more stringent cutoff criterion, Cochrane-Brink, Lofchy, and Sakinofsky (2000) found that scores of 15 or higher yielded a sensitivity rate of 100% and a specificity rate of 71%. Negative predictive power was excellent at 100%, and positive predictive power (45%) was much higher than the 1.3% reported by Beck et al. (1999). Therefore, using a slightly higher cutoff score than the recommended 8 or 9 may increase overall classification rates without necessarily jeopardizing sensitivity. Factor analytic studies have also supported the construct validity of the BHS. The BHS was originally reported to consist of three factors (Beck et al., 1974), but more recent studies have demonstrated that the variance is best explained by a one-factor solution (see Steed, 2001). This finding is replicable across both adult clinical (Dyce, 1996; Mendonca et al., 1983; Young et al., 1992) and nonclinical (Chang, D’Zurilla, & Maydeu-Olivares, 1994; Steed, 2001; Ward & Thomas, 1985) samples. Thus, the BHS is arguably a unidimensional measure best described as general hopelessness or pessimism for the future.

#### **Range of Applicability and Limitations**

The BHS was initially developed and tested on adult clinical patients, and there is a wealth of evidence implying that hopelessness, as measured by the BHS, is an excellent and reliable predictor of suicidal tendencies, including ideation, intent, and attempt. The BHS was not developed for use with nonpsychiatric individuals, yet its validity seems to be supported in this population as well. Because the internal reliability of the BHS is lower and more variable in nonclinical groups, however, researchers may wish to consider Steed’s (2001) recommendation to eliminate items that do not fit well with the rest of the scale. Regardless of the sample composition, researchers or clinicians should be cautious using the BHS when moderate to high levels of hopelessness are not anticipated (see Young et al., 1992).

There has been some debate in the literature over whether the BHS measures hopelessness or high, negative social desirability in patients (see Glanz, Haas, & Sweeney, 1995).

Fogg and Gayton (1976) first reported that the BHS correlated negatively with social desirability (coefficients ranged from  $-.47$  to  $-.64$ ), leading these authors to warn that

scores on the BHS could be contaminated by response sets. This finding has since been replicated by several other researchers (e.g., Ivanoff & Jang, 1991; Linehan & Nielsen, 1981; Mendonca et al., 1983). For example, Mendonca et al. (1983) studied 78 patients who presented at a crisis unit in a psychiatric hospital. BHS scores were significantly different in nonsuicidal individuals, suicidal ideators, and attempters, but these differences disappeared when social desirability was partialled out. Although one would expect that hopelessness is inversely related to social desirability, the BHS items related more to responding undesirably than to the magnitude of suicidality. Thus, even though social desirable response sets should be considered when using the BHS, social undesirability is a genuine feature of many aspects of psychopathology.

#### **Age and Cross-Cultural Factors**

Because of the lack of psychometric information on the BHS in adolescents, coupled with the fact that some studies indicate that this measure may not be as predictive of suicidality as it is in adult clinical groups, it is unclear whether this instrument should be used with adolescent clinical populations. Congruent with this notion, Beck and Steer (1991) stated that this measure is most appropriate for adults aged 17 years and older. Regardless of whether this instrument is to be used in younger individuals, BHS users should supplement their assessment of suicidality with additional indices to provide convergent evidence. Additional research is also necessary to more adequately assess the psychometric characteristics of the BHS in adolescents and to ascertain the relationship between the BHS and suicidality in this population.

There is a similar paucity of data on the use of the BHS with various ethnic groups. As with the BDI-II, the majority of BHS studies have used primarily Caucasian samples. Steer, Iguchi, and Platt (1994) found that Black participants scored significantly lower on the BHS than Caucasian and Hispanic individuals. Whether this decreased hopelessness is a veridical finding or may reflect cultural biases in the BHS is not presently known. Additional research is required to test the utility of the BHS in ethnic minority groups (Ivanoff & Jang, 1991), including ethnic minority adolescents (Steer et al., 1993a), and to determine whether separate normative information is warranted.

Both standard and translated versions of the BHS have been examined in various geographic regions of the world, including Brazil (Feijo, Saueressig, Salazar, & Chaves, 1997), China (Chiles et al., 1989), Finland (Suominen, Isometsa, Henriksson, Ostamo, & Lonnqvist, 1997), Japan (Tanaka, Sakamoto, Ono, Fujihara, & Kitamura, 1996; Tanaka, Sakamoto, Ono, Fujihara, & Kitamura, 1998), and Sweden (Nordstroem, Schalling, & Asberg, 1995). These studies reveal that there may indeed be significant cultural differences in the meaning of BHS scores and its psychometric properties. Studies conducted in European populations report findings similar to those using North American populations (Nordstroem et al., 1995; Suominen et al., 1997). Studies of Asian populations have, however, produced disparate results.

Tanaka et al. (1998) found a two-factor structure for the BHS as opposed to the one-factor solution typically found in North American studies. Also, there were no significant correlations between depression symptoms and the factor scores, nor any notable differences in BHS scores between individuals with a psychiatric history and those without such a history. Similarly, Chiles et al. (1989) found that the BHS was a significant predictor of suicide intent for American, but not Chinese, attempters. The relationships among depression, hopelessness, and suicidal intent (and the extent and manner in which it is manifested) may be quite different across cultures.

#### **Accommodation for Populations with Disabilities**

Some persons might have difficulty completing the BHS on their own, because of extreme fatigue, severe concentration problems, chronic illness, or visual impairments. In such instances, it is possible to administer the questionnaire orally (for instructions, see Beck & Steer, 1988).

#### **Legal and Ethical Considerations**

The BHS is considered to be an indirect measure of suicidal risk (Beck & Steer, 1988) and should not be used as the sole index of suicidality. Rather, suicide potential should be assessed with a thorough clinical interview and the use of more direct measures. Moreover, when deciding which cutoff to use as indicators of high suicidal risk, the concern for high sensitivity must override considerations to reduce the number of false negatives.

#### **Computerization**

Steer et al. (1994) tested a computerized version of the BHS with 330 inpatients. The reliability of the computerized BHS concurs with the printed version. The computerized version correlates highly with the total BDI-IA ( $r = .70$ ) and the BDI-IA pessimism item ( $r = .67$ ). The relationship between the BHS and the BDI-IA was virtually the same ( $r = .68$ ) when the pessimism item was excluded, thereby ruling out the confound that the relationship between these measures was driven by item overlap. Although the computerized BHS was able to significantly differentiate mood-disordered and non-mood-disordered groups it produced mean scores that were significantly lower than those reported previously using the printed version (Steer et al., 1994). Thus, test users should be cognizant of the fact that the cutoff scores recommended for the printed version may not be appropriate for the computerized version. Further research is needed to assess the comparability of the printed and computerized versions.

#### **Current Research Status**

The BHS is an excellent measure and is the most widely used measure of hopelessness available (Glanz et al., 1995). Empirical data indicate that the BHS is a highly reliable and valid measure for use among adult clinical groups. Although the BHS provides only an indirect assessment of suicidality, important information concerning the relative severity of pessimism may also be acquired. As such, the BHS may be an important measure to include in the evaluation of treatment outcome. Although this measure is used most frequently with depressed and suicidal individuals, hopelessness is appropriate to assess in a variety of other populations as well. To

illustrate, the BHS has been used to assess levels of hopelessness in identified carriers of the Huntington's disease gene (Tibben, Timman, Bannink, & Duivenvoorden, 1997), HIV-positive persons (Swindells et al., 1999), and alcohol and heroin-dependent women (Beck, Steer, & Shaw, 1984). The BHS may be used to assess baseline levels of hopelessness and to monitor fluctuations and improvements in hopelessness over time.

Until more consistent data are available, ascertaining whether the BHS is as useful for adolescents as it is for adults is difficult. Rotheram-Borus and Trautman (1988) suggested that hopelessness may be a symptom of depression in adolescents

rather than a separate factor. Clinicians are cautioned to supplement the use of the BHS with another instrument designed specifically for assessing suicidality in adolescents. Use of the BHS with nonclinical groups has been ubiquitous, despite the rather poor reliability reported for this population. Considering that the BHS may be unreliable among groups with low levels of hopelessness (Young et al., 1992), researchers should ensure that at least moderate levels of hopelessness are anticipated among the population they intend to test. Also, users should be aware of the presence of social desirability among certain populations and how this could affect BHS scores.

An important direction for research involves examining the psychometrics and appropriateness of the BHS with ethnic and cross-cultural groups. This research should address whether reliable ethnic differences in BHS scores exist and if separate norms are necessary for different cultures. Also, test users need to be aware of cultural differences in expectations for the future and how this might influence BHS scores (Tanaka et al., 1998).

Finally, the incremental validity of the BHS, relative to other measures of suicidality, is another empirical question worth addressing. For example, it is not presently clear whether the BHS should be used instead of, or in addition to, the BSS or the Suicide Intent Scale (SIS; Beck, Schuyler, & Herman, 1974) for predicting or assessing suicide risk. Schnyder, Valach, Bichsel, and Michel (1999) argued that there has been a tendency to overvalue the role of hopelessness in suicide assessment. If a clinician is concerned only with determining suicide risk, rather than assessing general pessimism that may be related to psychopathology or treatment outcome, the BSS appears to be the measure of choice (Cochrane-Brink et al., 2000), and this measure is discussed next.

### **BECK SCALE FOR SUICIDE IDEATION (BSS)**

#### **Test Description**

Beck et al. (1988) developed a self-report version of suicidal ideation that could be administered by either paper and pencil or computer. The BSS is a 21-item measure of which only the first 19 items are scored. The final two are items used to record information concerning previous suicide attempts. All items consist of three response options, ranging from 0 to 2. Respondents are asked to circle the statement that best describes how they have been feeling over the

past week, including the current day. An example item follows:

0 I have a moderate to strong wish to live.

1 I have a weak wish to live.

2 I have no wish to live.

Total scores, which range from 0 to 38, are obtained by adding the item values. The first five items (i.e., “wish to live,” “wish to die,” “reasons for living or dying,” “active suicide attempt,” and “passive suicide attempt”) serve as an initial screen for suicide ideation. If respondents circle zero on both of the items pertaining to suicide attempts, they are instructed to skip to the end of this scale to complete the last two items.

### **Test Development**

Beck et al. (1972) published recommendations for the design and operation of programs for suicide prevention and amelioration.

Among these recommendations was a classification system for suicidal behaviors, which was comprised of three broad categories: suicide completions, attempts, and ideation.

Although Beck and colleagues had already developed an intent scale for suicide attempters that was also partly applicable to suicide completers (Beck, Kovacs, & Weissman, 1979), there was a need to develop an instrument that would assess suicidality for the third category of suicidal behaviors—suicide ideation.

Suicide ideation can be defined as “the presence of current plans and wishes to commit suicide in individuals who have not made any recent overt suicide attempts” (Beck et al., 1988, p. 499). Beck et al. (1979) initially developed a clinician rating scale called the Scale for Suicide Ideation (SSI).

The items for this rating scale were generated on the basis of clinical observations, interviews with suicidal patients, and previous research in the area of suicide. This measure was originally piloted on suicidal patients and its items were refined or eliminated if they were ambiguous, difficult to rate, or consisted of overlapping content. This test construction phase resulted in a 19-item clinician rating scale, designed to be administered as a semistructured interview. The test developers were later interested in adapting the SSI into a self-report index of suicidal ideation that could be used independently or concurrently with this instrument. This new inventory, the BSS, was then tested on inpatient and outpatient samples.

### **Psychometric Characteristics**

The internal consistency of the BSS is excellent in both adult and adolescent clinical samples (Beck & Steer, 1991; Beck et al., 1988; Kumar & Steer, 1995; Steer, Kumar, & Beck, 1993b). For the printed version of the BSS, coefficient alpha

ranges from .87 (Beck & Steer, 1991) to .93 (Beck et al., 1988). The computer-administered version yields high coefficients that range from .90 (Beck & Steer, 1991) to .96 (Beck et al., 1988). Item total correlations are generally acceptable and range from .20 to .73 ( $M = .52$ ,  $SD = .15$ ). The test-retest reliability of the BSS appears to be moderate. Beck and Steer (1991), for example, reported a 1-week test-retest reliability coefficient of .54 in a sample of 60 adult inpatients.

BSS scores decreased significantly during this time frame, which the authors hypothesized was the result of clinical improvement (Beck & Steer, 1991).

The validity of the BSS has also been tested using both administration formats. The paper-and-pencil and computerized versions of the BSS each correlate highly (e.g.,  $r = .90$ ) with clinician ratings (Beck et al., 1988), self-report (Beck & Steer, 1991; Cochrane-Brink et al., 2000), and related indices of suicidality (e.g., previous attempts, BDI-IA scores, the BDI-IA suicide ideation item; Beck & Steer, 1991). To date, only the printed version has been tested with adolescent inpatients (Kumar & Steer, 1995; Steer et al., 1993). In this population, the BSS correlates well with past history of suicide attempts, the BDI-IA (excluding suicide ideation and hopelessness items), the presence of a mood disorder, the BHS, and the BAI (Kumar & Steer, 1995; Steer et al., 1993).

The predictive utility of the BSS appears promising, but additional research is needed in this area. A cutoff score of 24 yields excellent sensitivity (100%) and specificity (90%), as well as impressive positive (100%) and negative (71%) predictive power (Cochrane-Brink et al., 2000). The BSS also significantly predicts the decision to admit a patient because of risk for suicide (Cochrane-Brink et al., 2000).

Given that only two studies have factor analyzed the BSS (Beck & Steer, 1991; Steer, Rissmiller, Ranieri, & Beck, 1994), it is not presently known whether a stable factor structure exists for this measure. Three main factors (active suicidal desire, suicidal ideation, and preparation for suicide) emerged in a fairly clear and consistent manner in these studies (although Beck & Steer, 1991, actually reported a fivefactor solution). Visual inspection of the factor loadings reported in these studies points to discrepancies that appear to limit the generalizability of the factor solutions reported. However, further investigation of this factorial validity of the BSS is needed.

#### **Range of Applicability and Limitations**

The BSS was developed as a self-report version of the SSI and has been appraised using adult clinical patients. Therefore, the BSS is best used as a measure of suicidal ideation in this population. However, it should be noted that the BSS should never be used alone to clinically assess suicidality and should be used in conjunction with a professional clinical assessment (Beck & Steer, 1991).

Limited data are available on the use of the BSS with adolescent patients (see Kumar & Steer, 1995; and Steer et al., 1993, for exceptions). Given that the BSS was designed to detect the severity of suicidal thoughts in adults and adolescents (Beck & Steer, 1991), further research should focus on the psychometric properties and the operating characteristics of the BSS in adolescents. Currently, it is most appropriate to use the BSS in adult psychiatric patients older than 17 years (Beck & Steer, 1991). Similarly, no research has examined the psychometrics of the BSS in nonpsychiatric groups. Such studies are important to increase the generalizability of this scale to community samples. In addition, the

development of nonclinical norms would permit normative comparisons and enhance the evaluation of clinical significance.

#### **Accommodation for Populations with Disabilities**

As with the BDI-II and BHS, the BSS may be administered orally if test takers experience difficulties with the standard format (see Beck & Steer, 1991).

#### **Legal and Ethical Considerations**

The BSS was not developed to replace clinical assessments (Beck & Steer, 1991) and should only be used as an adjunct to professional evaluations conducted by trained clinicians. Given the low base rate of actual suicide attempts, and consequently the increased difficulty in predicting suicidal behavior, it might be quite difficult to arrive at a cutoff score for the BSS that produces high specificity, while at the same time maintaining high sensitivity (but see Cochrane-Brink et al., 2000). Therefore, despite the possibility of increased false positives, it is in the best interest of clients to utilize lower cutoff scores to maintain high sensitivity. The BSS was developed to assist in the assessment of suicide risk as opposed to actual suicide prediction (Beck & Steer, 1991), but it is important to remember the gravity of false negatives. Finally, test users should be familiar with the protocol as well as the legal and ethical issues surrounding involuntary hospitalization.

#### **Current Research Status**

Given its good psychometric properties and the ease with which the BSS is administered and scored, it is expected that the BSS will become more widely used than it has been in

studies examining suicide ideation in adult clinical groups. Compared to the clinician-scored SSI, the BSS has the potential to be administered to large groups of research participants at a time, making it a convenient measure for investigating suicidality. Clinically, it has been described as the “clinical scale of choice” for the assessment of suicidality (Cochrane-Brink et al., 2000, p. 450). The BSS also has the potential to be an excellent measure of suicidal risk in adolescents, but more research is needed to determine its psychometric properties in younger populations. There is a paucity of information available concerning the use of the BSS with nonclinical groups. Consequently, except for research purposes, the BSS should not be used with nonclinical groups until adequate psychometric data are available. Finally, additional research is needed to evaluate the utility of the BSS in ethnic minority and cross-cultural groups.

#### **THE BECK SCALES IN CLINICAL PRACTICE**

The Beck scales have been used extensively in research and practice. In this section, we highlight some of the uses of the BDI-II, BHS, and BSS in clinical practice and raise practical issues for clinicians who use these measures. In general, we highly recommend the use of these Beck scales for the assessment of depressive severity, hopelessness, and suicidality and for the ongoing evaluation of psychotherapy outcome.

As this chapter has demonstrated, these scales are reliable and valid and provide clinically meaningful information. Numerous other depression symptom scales are available (Dozois & Dobson, 2002; Nezu, Ronan, Meadows, & McClure, 2000), but they often focus on slightly different themes than the

BDI-II and include items that are not directly pertinent to diagnostic criteria, thereby decreasing their specificity. The BDI-II is most congruent with the *DSM-IV* criteria, exhibits excellent psychometric properties, and emphasizes the cognitive and attitudinal symptoms of depression more than other popular self-report measures. The BDI-II is also the most frequently used self-report measure of depression, which affords practitioners the opportunity to compare their clinical results to the literature.

If clinicians are interested in assessing suicidal risk, we recommend using the BSS over the BHS because of its more direct link with suicidality. This does not mean the BHS should not be used with suicidal patients. Hopelessness, as measured by the BHS, is consistently predictive of suicidal tendencies. The BHS is also an excellent instrument for assessing treatment outcome and for monitoring a patient's feelings of hopelessness over the course of treatment. The BHS deals primarily with general pessimism about the future and may also be used to assess a patient's motivation for therapy and expectations for treatment change. Westra, Dozois, and Boardman (2002), for example, found that pretreatment hopelessness was significantly higher in therapy dropouts than in individuals who completed treatment. Hopelessness about symptom control was also related to fewer reductions in dysfunctional attitudes and to poorer overall treatment response.

The BSS is an excellent tool for assessing suicide risk and for tracking fluctuations in a patient's suicidal thoughts over the course of therapy. The BSS is also a valuable tool for initial assessment as it can be used to assess imminent risk for suicide. Although there exist no published recommendations for BSS cutoff scores, Cochrane-Brink et al. (2000) used a cutoff of 24, which produced excellent predictive values.

The Beck scales may also be helpful for treatment planning. Patients often present to clinics with comorbid conditions and it is often difficult for clinicians to know which condition to target as an initial strategy for intervention. When a patient exhibits comorbid depression and anxiety, for instance, understanding the severity of his or her depressive symptoms, degree of hopelessness, and risk for suicide may differentially guide one's approach to intervention. If the depression, hopelessness, and suicidality are very severe, it would be important to deal with these issues first so that the patient will have the resources, motivation, and energy to manage exposure-based approaches for treating anxiety. Conversely, if the patient does not present a high risk for suicide and does not show a high level of depression or hopelessness, then the clinician may opt to target the anxiety, with the hypothesis being that once the anxiety has improved, the depression will dissipate as well. Thus, having accurate data from these self-report instruments can facilitate case conceptualization and treatment planning (see Dozois & Dobson, 2002).

There are a number of other uses of the Beck scales in practice. These instruments may be used to (1) ensure that

one's approach to treatment is effective; (2) monitor problems (e.g., motivational issues) that may arise during the course of treatment; (3) encourage patients by using data and demonstrating that the practitioner is confident in his or her approach and that he or she respects accountability; (4) illustrate to patients the amount of progress being made (e.g., depressed patients are notorious for disqualifying the positives and believing that they are not making significant progress when they in fact are); (5) examine the stability of the treatment response (e.g., minimizing the chances that a patient's

change simply reflects a flight into health); (6) indicate when treatment has been successful and may safely be terminated; (7) determine the clinical significance of treatment change; and (8) prevent relapse (see Dozois & Dobson, 2002).

The clinical significance of symptom change is important in both psychotherapy outcome trials and in clinical practice. One strategy for determining clinical significance is to use normative comparisons. Normative comparisons allow clinicians to determine whether a patient's functioning on a given measure has shifted from being within the dysfunctional range to being within a nomothetically average range.

Nonclinical norms have been developed for both the BDI-II (Kendall & Sheldrick, 2000) and the BHS (Dozois, Covin, & Brinker, 2003). We recommend the use of these norms for the assessment of clinical significance.

One issue that was raised earlier in this chapter pertains to the use of the Beck scales in repeated assessment. Although we recommend using the Beck scales frequently in practice to, among other things, gauge efficacy, researchers and practitioners need to be aware of the issues that surface when administering a task repeatedly to clients. As previously noted, a number of studies have found that self-report symptom scores drop with repeated assessments. Ahava et al. (1998) administered the BDI-IA over 7 weekly administrations and found that scores decreased substantially over time in nonclinical participants with no intervention. In the BDI-II manual, Beck et al. (1996) mentioned that the effects of memory and response sets need to be examined but that they should be the same with the BDI-II as they were with the earlier versions of the BDI. This is not very encouraging given the findings from Ahava et al. (1998) and others (e.g., Sharpe & Gilbert, 1998; Yin & Fan, 2000). Clinicians who decide to use the Beck scales recurrently in practice should consider the effects that repeated assessment have on obtained scores. One implication is that it is important to ensure that the decrease in depression, hopelessness, or suicidality scores are in fact related to treatment change rather than due exclusively to repeated assessment. Although this would deviate from standardized administration, clinicians may also consider randomizing response options to minimize the effects of earlier exposure to these tests. Researchers using any of these instruments for prescreening may wish to use alternative measures for the initial screen and the indicated measure for determining whether participants meet criteria for their study. Although regression to the mean may account for some of the findings, they are instructive and should serve as

a caution when conducting repeated assessments (see also Yin & Fan, 2000).

#### **SUMMARY AND FUTURE DEVELOPMENTS**

This chapter provided a comprehensive review of the BDIII, BHS, and BSS. As our review has documented, these scales exhibit excellent psychometric properties and are extremely useful for research and practice. In terms of research, these Beck scales are used widely. The BDI-II, in particular, has been cited in numerous psychotherapy outcome studies as one of the core dependent variables. This instrument is also used frequently for determining inclusion and exclusion criteria in myriad studies of clinical and analogue depression. Although the BHS and BSS have not been as dominant as the BDI-II, these measures have been utilized to test pessimism and suicide risk in a variety of populations. Clinically, these three scales appear to be excellent measures of the constructs they purport to measure and are useful for case conceptualization, treatment planning, monitoring patient change over time, and evaluating treatment outcome and the clinical significance of therapeutic change. Throughout this chapter, we have highlighted some of the limitations of each of these measures and provided a number of suggestions for further empirical work. Rather than reiterating these recommendations, we simply conclude that there are many exciting avenues for future research that we hope researchers will investigate. The BDI-II, BHS, and BSS are highly useful instruments and we anticipate that a future review of this research will confirm this generally positive review.

#### **REFERENCES**

- Abdel-Khalek, A.M. (1998). Internal consistency of an Arabic adaptation of the Beck Depression Inventory in four Arab countries. *Psychological Reports, 82*, 264–266.
- Ahava, G.W., Iannone, C., Grebstein, L., & Schirling, J. (1998). Is the Beck Depression Inventory reliable over time? An evaluation of multiple test-retest reliability in a nonclinical college student sample. *Journal of Personality Assessment, 70*, 222–231.
- Aikens, J.E., Reinecke, M.A., Pliskin, N.H., Fischer, J.S., Wiebe, J.S., McCracken, L.M., & Taylor, J.L. (1999). Assessing depressive symptoms in multiple sclerosis: Is it necessary to omit items from the original Beck Depression Inventory? *Journal of Behavioral Medicine, 22*, 127–142.
- Ambrosini, P.J., Metz, C., Bianchi, M.D., Rabinovich, H., & Undie, A. (1991). Concurrent validity and psychometric properties of the Beck Depression Inventory in outpatient adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*, 51–57.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Arnau, R.C., Meagher, M.W., Norris, M.P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychology, 20*, 112–119.
- Beck, A.T. (1963). Thinking and depression. *Archives of General Psychiatry, 9*, 324–333.
- Beck, A.T., Brown, G., Berchick, R.J., Stewart, B.L., & Steer, R.A. (1990). Relationship between hopelessness and ultimate suicide:

A replication with psychiatric outpatients. *American Journal of Psychiatry*, 147, 190–195.

Beck, A.T., Brown, G.K., Steer, R.A., Dahlsgaard, K.K., & Grisham, J.R. (1999). Suicide ideation at its worst point: A predictor of eventual suicide in psychiatric outpatients. *Suicide and Life-Threatening Behavior*, 29, 1–9.

Beck, A.T., Davis, J.H., Frederick, C.J., Perlin, S., Pokorny, A.D., Schulman, R.E. et al. (1972). Classification and nomenclature. In H.L.P. Resnik & B.C. Hathorne (Eds.), *Suicide prevention in the 70s* (pp. 7–12; DHEW Publication No. HSM 72–9054). Washington, DC: U.S. Government Printing Office.

Beck, A.T., Epstein, N., Brown, G., & Steer, R.A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, 56, 893–897.

Beck, A.T., Epstein, N., Harrison, R.P., & Emery, G. (1983). *Development of the Sociotropy-Autonomy Scale: A measure of personality factors in psychopathology*. Unpublished manuscript, Center for Cognitive Therapy, University of Pennsylvania Medical School, Philadelphia.

Beck, A.T., Guth, D., Steer, R.A., & Ball, R. (1997). Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. *Behaviour Research and Therapy*, 35, 785–791.

# The Assessment of Attention via Continuous Performance Tests

CYNTHIA A. RICCIO CECIL R. REYNOLDS

Problems with attention or concentration in one form or another are common concerns, if not the primary concern, when children and youth are referred for comprehensive evaluation; indeed, such problems are common among children generally. Attention is a key component in cognitive functioning (Siegel, 2000) and the neural traces of attention form the foundation of memory (Cohen, 1993b; Sohlberg & Mateer, 1989); furthermore, attention can be viewed as the process that controls the flow of information processing in the brain (Siegel, 2000). Consistent with this view, Broadbent (1953, 1957) proposed that the individual's capacity to take in information is limited, and, therefore, information that is not relevant needs to be filtered out. He further proposed that characteristics of the stimulus (i.e., intensity, importance, and novelty) determined whether specific information would be filtered out or attended to. With its presumed relation to cognition, information processing, and memory, it is not surprising that attentional problems are the most frequent type of cognitive impairments resulting from neurological disorders or dysfunction (Cohen, Malloy, & Jenkins, 1999; Mapou, 1999).

When parents and teachers indicate that a child's attention or concentration is a concern, however, defining or measuring what is meant by attention is not a simple task. As used in common discourse, attention is not unitary but has been described as a highly complex construct (Mirsky, Fantie, & Tatman, 1995). Although many theories of attention have focused on arousal (Hebb, 1958; Moruzzi & Magoun, 1949; van Zomeran & Brouwer, 1994), current conceptualizations of attention are concerned with more than the individual's arousal level (e.g., Posner & Peterson, 1990; Voeller, 1991). It has been suggested that attention should be conceptualized as having multiple components or elements. These components include (1) the initiating or focusing of attention; (2) sustaining attention or vigilance; (3) inhibiting responses to irrelevant stimuli or selective attention; and (4) shifting attention (Mirsky, 1989; Mirsky, Anthony,

Duncan, Ahearn, & Kellam, 1991; Sohlberg & Mateer, 1989; Zubin, 1975). Due to the need for attentional shifts and attentional flexibility in the self-regulatory and inhibition processes, attentional processes held a central role in Luria's (1966) model of normal and abnormal brain function and are closely linked to executive control. Both executive control and attention are also necessary for initiation or generation of the response to a specific stimulus, for maintenance of the response or shifting of the response, and for the flexibility needed to meet changing task demands (Cohen, 1993c; Whyte, 1992).

The multifaceted nature of attention and executive control makes it difficult to map these behaviors onto specific neurological structures. Current functional system models include cortical and subcortical structures as well as connecting pathways and projections with increasing levels of complexity and are reviewed elsewhere (see Cohen & O'Donnell, 1993; Riccio, Reynolds, & Lowe, 2001; van Zomeran & Brouwer, 1994). The complex functional system models that integrate both neuroanatomical and neurochemical influences on behavior provide a plausible explanation for the myriad manifestations of deficits in attention and executive control across disorders of disparate etiology (Voeller, 1991). Using these complex models, interference at any level of the functional system could lead to a cluster of clinically similar, and yet potentially different, behaviors depending not only on the location of the damage or dysfunction but also on the extent of that damage or dysfunction. Because compromise anywhere within the system may lead to attentional deficits regardless of etiology, the use of attention problems alone as a diagnostic consideration is suspect. At the same time, the assessment of attention may provide useful information for treatment planning across disorders. To this end, a number of tests have been developed for the purpose of measuring the many facets of attention (Cohen, 1993a, 1993c; Riccio & Reynolds, 2001). To rule out other confounds (e.g., cognitive ability), the task demands of measures of attention need to involve limited cognitive processing and, thus, need to be relatively simple. Similarly, to eliminate confounds with memory limitations, the memory load must be minimal as well. The specific

components of attention that are of interest also will dictate task parameters. When sustained attention is the construct to be measured, repetition and duration of the task are key features. When focus or selective attention is of interest, the characteristics of the relevant and nonrelevant stimuli become more important. Parameters of task complexity, temporal demands, processing speed requirements, and task or target salience also need to be considered in the assessment of attention (Cohen et al., 1999).

Computerized assessment of attention and executive control facilitates the accommodation of tasks to match these constraints.

The continuous performance test (CPT) is one group of paradigms for the study of attention. Although not all CPTs are computerized, technology allows for virtually infinite variations to task parameters across differing CPTs. As we review the CPT literature and the multitude of CPT variations, we discuss the extent to which the parameters can change the task and influence the results obtained.

#### THE CPT PARADIGM

The CPT is one group of paradigms for the evaluation of attention as well as response inhibition (or disinhibition). Most often, CPTs are used to obtain quantitative information regarding an individual's ability to sustain attention over time. Of the myriad measures of attention available, the CPT is the most frequently used (DuPaul, Anastopoulos, Shelton, Guevremont, & Metevia, 1992). The original CPT was developed by Rosvold, Mirsky, Sarason, Bransome, and Beck in 1956 as a research tool to study vigilance. In the original version, letters were presented visually one at a time, at a fixed rate (interstimulus interval, or ISI) of 920 milliseconds. The individual was required to respond by pressing a lever whenever the letter "X," designated as the target stimulus, appeared. At the same time, the individual was to inhibit responding when any other letter appeared; this is referred to as the X-type CPT. A variation of this task in which the target was the letter "X," but only if the "X" was immediately preceded by the letter "A" (or an AX-type CPT) was used as well. Rosvold and colleagues found group differences to occur when comparing individuals with brain damage to controls on the X-type CPT; these between-group differences were magnified with the increased

difficulty level of the AX-type CPT.

#### 292 IV. SPECIFIC SYNDROMES AND SYMPTOMS

##### CPT Variations and Effects

Since 1956, the CPT has continued to be used in the study of attention as well as executive control, with multiple variations in the components of the task; today, the majority of CPTs are computer administered.

The basic paradigm for the CPT consists of rapid presentation of continuously changing stimuli with a designated “target” stimulus or “target” pattern such that the individual is to respond (or inhibit responding) based on the stimulus presented. As such, it requires selective attention or vigilance for an infrequently occurring target or relevant stimulus; at the same time, the duration of the task is intended to be sufficient to measure sustained attention. Despite these general similarities, Halperin (1991) commented that there were as many versions of the CPT available as there were clinicians who used them, while Greenberg and Waldman (1993) suggested that there are more than 100 different versions of the CPT in use. Different CPTs include variations in the basic task or CPT type, the characteristics of the target, variations in the ISI, presence or absence of distractors, presence or absence of feedback following responses, modality of presentation, duration of the target presentation, duration of task, and so on (see Table 13.1). The effects of some of these possible variations and modifications to the CPT on performance have been reviewed elsewhere (see Ballard, 1996; Corkum & Siegel, 1993; Riccio et al., 2001) and are summarized here.

##### Task Type

The task may be the simpler X-type CPT, or an AX-type CPT, or a further modification of the AX such that the target must be preceded by itself (XX-type: e.g., Fitzpatrick, Klorman, Brumaghim, & Borgstedt, 1992) or where color and letter are critical features (e.g., orange T followed by blue S: Garfinkel & Klee, 1983). Another modification involves a change in the directions to respond *except* when the target is presented (not-X type) as in the Conners CPT (Conners, 1992, 1995) and the CPT II (Conners & MHS Staff, 2000). A number of studies have compared performance on the X-CPT to performance on the AX-CPT with consistent findings that the AX-CPT is the more difficult task (e.g., Alexander, 1973; Goldstein,

Rosenbaum, & Taylor, 1997). Similarly, when the X-CPT has been compared to the XX-CPT, findings supported the premise that the X-CPT is the easier task (Friedman, Vaughan, & Erlenmeyer-Kimling, 1978, 1981). Schachar, Logan, Wachsmuth, and Chajczyk (1988) compared the performance of 114 boys (7–11 years of age; clinic referrals) on an X-, XX-, and AX-CPT. Keeping all other parameters the same and counterbalancing to control for order effects, results indicated that there was a greater deterioration of performance over time on the XX-CPT and AX-CPT as compared to the X-CPT. In particular, Schachar and colleagues found decreased correct hits with the AX- and XX-CPT as compared to the X-CPT. In addition, results revealed faster reaction times for the AX-CPT and increased commission errors (false alarms) on the XX-CPT (Schachar et al., 1988).

#### Stimulus Characteristics

Other differences include the type of stimulus (e.g., letters, numbers, shapes, words, and objects), or how the target is defined (e.g., position of small square within large square: Greenberg, 1988–1999). Drawing from Broadbent (1953, 1957), it would be expected that the characteristics of the target (i.e., novelty and semantic content) could affect the likelihood that irrelevant (nontarget) stimuli would be filtered out. Consistent with this, characteristics of the target (i.e., animal vs. letter, nonword vs. word, and verbal vs. nonverbal) have been found to have an impact on CPT performance (e.g., Earle-Boyer, Serper, Davidson, & Harvey, 1991; Harper & Ottinger, 1992; Shapiro, Morris, Morris, Flowers, & Jones, 1999).

Another characteristic of the stimulus is the quality of the stimulus itself. In some studies, the stimulus is degraded in some manner. Evans (1988) compared performance on a standard X-CPT, the X-CPT with degraded (i.e., blurred) stimuli, and other variations. Findings revealed that the condition with degraded stimuli was the most difficult, while the standard X-CPT was the easiest. Clinicians need to keep in

13. Continuous Performance Tests 293

#### 294 IV. SPECIFIC SYNDROMES AND SYMPTOMS

##### TABLE 13.1. Variations in Continuous Performance Tests

Variation Examples Possible effect(s)

Task type X-type, AX-type, XX-type, Difference in level of difficulty (e.g., Friedman

not-X type, identical pairs et al., 1981; Schachar et al., 1988). Difference between initiation of response (e.g., on X-type) and inhibition of dominant response on not-X has not been studied.

Stimulus Letters, numbers, shapes, Novelty and semantic content believed to affect the characteristics words, objects, ear likelihood that irrelevant stimuli (nontargets) would be filtered out (e.g., Earle-Boyer et al., 1991; Shapiro et al., 1999).

Position on screen (or ear if No research available. auditory), change in position (or ear) within task

Standard presentation Blurred or degraded presentation increased the degraded or blurred difficulty level of the task (e.g., Evans, 1988).

Modality of Visual, auditory, or variable Auditory presentation more difficult than visual presentation within task (e.g., Baker et al., 1995; Sandford et al., 1995a, 1995b). No research available to compare single modality (visual or auditory) to variable within task modality.

Target frequency High or low frequency High frequency is believed to increase the likelihood of impulsive responding, while low frequency is believed to increase the likelihood of inattention (e.g., Beale et al., 1987).

Interstimulus Shorter, longer, variable, ISIs that are either too short or too long have been interval (ISI) adaptive found to yield increased error rates (e.g., Chee et al., 1989; Sykes et al., 1971)

Stimulus duration Shorter or longer Duration of presentation was found to impact results (Chee et al., 1989).

Distractor Present or absent, same or Distractors (auditory or visual) for visual CPTs conditions different modality as resulted in increased errors (e.g., Crosby, 1972). stimulus Visual distractors resulted in poorer performance in other studies as well (e.g., Golier et al., 1997).

Task duration Range of current CPTs is No comparative research available (i.e., how many 3–30+ minutes minutes are sufficient? 9, 14, 23?).

Examiner Present or absent A decline in performance is associated with presence examiner absence (e.g., Draeger et al., 1986; Power, 1992)

Instructional set Emphasis on speed, Instructional set has been found to affect accuracy, or both performance (e.g., Lark et al., 1999; Sergeant & Scholten, 1985).

Provision of Feedback on performance, Research findings are equivocal (e.g., Corkum et feedback or tangible reinforcers, al., 1996; Levy & Hobbes, 1988). reinforcement response cost contingency

Equipment used Size of screen, spacebar or No research is available on these differences. mouse, type of mouse (if

mouse), laptop or PC, headphones or speakers

mind, however, that blurring may increase the information-processing component rather than, or in addition to, increasing attentional demands.

### Modality of Presentation

Another variation involves changing the modality such that presentation may be visual as in the initial version, or it may be auditory (e.g., Earle-Boyer et al., 1991; Keith, 1994), or it may vary within the same task from auditory to visual stimuli (e.g., Sandford & Turner, 1994–1999). Modality differences have been investigated with associated differences found in CPT performance

(e.g., Baker, Taylor, & Leyva, 1995; Draeger, Prior, & Sanson, 1986; Driscoll, 1994; Lark, Dupuy, Greenberg, Corman, & Kindschi, 1996; Sandford, Fine, & Goldman, 1995a, 1995b; Sandford & Turner, 1995). For example, Sandford and Turner (1995) reported better performance on visual than auditory tasks in their study, with children with attention-deficit/hyperactivity disorder (ADHD) making more commission errors when the stimuli were presented in the auditory format. Lark and colleagues (1996) reported comparisons between visual and auditory test formats. Results indicated that children made twice as many omission errors on the auditory task, but significantly more commission errors and faster reaction times on the visual task. At the same time, children demonstrated greater variability in response time on the auditory task (Lark et al., 1996). Finally, auditory CPTs have yielded higher correlations with verbal ability, language, and achievement as compared to visual CPTs (Kupietz & Richardson, 1978; Swanson & Cooney, 1989).

#### **Target Frequency**

Studies have varied the frequency of the target to provide a higher frequency or a lower frequency (e.g., Beale, Matthew, Oliver, & Corballis, 1987). Under the frequent target condition, a response set is created such that the dominant response is to press the mouse button, and, therefore, an individual who is impulsive is more likely to respond and to make commission errors. In contrast, the infrequent target condition requires vigilance and response inhibition. It is believed that the stimulus–infrequent condition creates a situation in which a client who is inattentive is less likely to respond and, therefore, will make omission errors. If vigilance is not maintained during this segment of the test, the probability of omission errors will increase. The individual's "readiness to respond" or arousal level may impact the number of omission errors as well (Beale et al., 1987; Sandford & Turner, 1995).

#### **ISI Differences**

The time lapse between presentations of the stimuli or the ISI has been varied with studies using a shorter interval or longer interval (e.g., Rueckert & Grafman, 1996; Sykes, Douglas, Weiss, & Minde, 1971); in at least some instances, the ISI has been found to affect CPT performance. For example, Sykes

and colleagues (1971) assessed 40 children with hyperactivity and 19 control children (5–12 years of age) on multiple versions of the X- and AX-CPT with two ISIs (1,000 and 1,500 milliseconds). Results indicated a main effect for ISI for total correct responses and error scores such that both groups of children performed better when the ISI was longer regardless of other conditions. Related to ISI, yet discussed less frequently, is the stimulus onset asynchrony (SOA) which is the time between the onset of one stimulus and the next stimulus. The SOA is the sum of the duration of the stimulus presentation and the ISI, and, therefore, is sensitive to the effects on differences in duration of presentation as well as differences in ISI (e.g., Chee, Logan, Schachar, Lindsay, & Wachsuth, 1989). Consistent with the findings for ISI, performance improved with longer SOAs in the Chee study.

The ISI may vary within the task as well. Variable intervals may be preset such that for some blocks of trials the ISI is at one rate, while for other blocks the ISI is either longer or shorter (e.g., Conners, 1992, 1995; Conners & MHS Staff, 2000). In contrast, with an “adaptive” variable rate, the computer program automatically increases or decreases the ISI by 5% based on the accuracy of the last response (e.g., Brumm, 1994; Girardi et al., 1995; Rapoport et al., 1980). For CPTs with this format, the mean ISI (indicative of optimal performance for that person) can be determined and is believed to reflect the information-processing speed of the individual.

#### Distractor Conditions

Other studies have included distractor conditions (e.g., Crosby, 1972; Golier et al., 1997; Hoy, Weiss, Minde, & Cohen, 1978). Crosby (1972) used visual X- and AX-CPTs with children under three conditions—no distractor, with auditory distractors, and with visual distractors. Results revealed deterioration in performance for all groups under the distractor conditions as compared to the no-distractor condition. Similarly, other studies (e.g., Golier et al., 1997; Hoy et al., 1978) consistently found a decline in performance for both the clinical and control groups under the distractor conditions.

#### Feedback/Reinforcement

Still other studies examined the effects of providing feedback or reinforcement. As

noted in Corkum and Siegel's (1993) review, the provision of feedback (e.g., O'Dougherty, Nuechterlein, & Drew, 1984) and provision of tangible reinforcers (e.g., Levy & Hobbes, 1988) have been found to affect the overall performance of individuals. In at least one study, provision of feedback resulted in a higher hit rate (O'Dougherty et al., 1984). In another study, Levy and Hobbes (1988) found that a response cost contingency resulted in fewer errors of omission than either a no-feedback or reward contingency condition. In contrast, however, Corkum, Schachar, and Siegel (1996) found that the provision of incentives did not significantly affect CPT performance and concluded that motivational issues did not account for the vigilance decrement evidenced.

#### **Task Duration**

Another variable that differs across CPTs is the duration of the task itself. The condensed version of the MINI-CPT (Bremer, 1989) is probably one of the shortest CPTs with a duration of only 3 minutes. On the other hand, at least one CPT lasts over 30 minutes (Mansour, Haier, & Buchsbaum, 1996). Clearly, the extent to which one can interpret the results as indicative of sustained attention is directly linked to duration of the task.

#### **Examiner Presence or Absence**

Another factor found to affect CPT performance is examiner presence or absence during the administration of the CPT (e.g., Draeger et al., 1986; Power, 1992). Draeger and colleagues (1986) compared the CPT results for children under two conditions: examiner present or examiner absent. Results indicated that for children with ADHD as well as normal controls, a decline in performance occurred when the experimenter was absent; the deterioration in performance was greater for the children with ADHD. Results of the Power (1992) study supported this finding with some indications that children who were aggressive were more likely to demonstrate the deterioration of performance in the examinerabsent condition. Moreover, Power found that the deterioration in performance in the examiner absent condition was greater if the absent condition occurred second as opposed to first in the counterbalanced sequence.

#### **Instructional Set**

In addition to examiner presence or absence,

differences in the instructional set or directions given to the individual have been investigated (e.g., Sergeant & Scholten, 1985; Tupler, 1989). Sergeant and Scholten (1985) administered the CPT under three different instructional set conditions—emphasis on speed, emphasis on accuracy, and emphasis on both speed and accuracy. Results indicated that CPT results were affected by the instructional set with decreased errors in the accuracy condition and increased errors in the speed condition. More recently, Lark, Dixon, Hoffman, and Huynh (1999) replicated these findings. They concluded that differences in performance could be elicited by changing the instructions given at the start of the CPT.

#### Summary

As can be seen from the effects of parameter differences on CPT performance, it is critical to ensure that the normative data and clinical data are obtained under the same parameters with strict adherence to standardization procedures. Still other differences (e.g., the size of the computer screen, use of laptop vs. PC, use of mouse vs. space bar) remain to be investigated as possible confounds to be considered in the use of the CPT. Notably, in addition to the actual administration parameters, differing CPTs also provide different variables for interpretation of performance.

#### CPT Variables for Interpretation

CPT variables reported often include correct hits (number or percent of correct responses to targets). In some cases, omission errors (i.e., number or percent of targets not responded to) are reported. Both correct hits and omission errors are interpreted as indicative of selective attention (e.g., Allen, 1993; Carter, Krener, Chaderjian, Northcutt, & Wolfe, 1995) as well as alertness (Gordon Systems, 1986). Commission errors (number or percent of responses to stimuli other than the target) are reported as an index of response inhibition. In some studies, relative accuracy (percent of correct responses of all responses made including incorrect responses) or total errors (combining omission and commission errors) may be reported (e.g., Bock, 1982).

A number of researchers have investigated the types of commission errors that the individual makes on the CPT. One broad category of commission errors is referred to

as target-related errors such that at least one item in the sequence is relevant to the correct target (e.g., X or A in the AX-CPT). Target-related commission errors are believed to be suggestive of anticipation or impulsivity (Gordon Systems, 1986; Halperin, Sharma, Greenblatt, & Schwartz, 1991). In contrast to target-related errors, random errors would be those responses to a sequence where neither item in the sequence was relevant to the target sequence. Random commission errors are not believed to be associated directly with inattention, impulsivity, or hyperactivity (Gordon Systems, 1986; Halperin, Wolf, Greenblatt, & Young, 1991) but, rather, to be reflective of dyscontrol (Halperin, Sharma, et al., 1991) or to be associated with problems in the individual's level of arousal, or motivational level (Gordon Systems, 1986). Some programs provide a measure of multiple responses to the same stimulus (i.e., the individual responds more than once to the same stimulus). Multiple responses are believed to reflect hyperresponsivity (Sandford & Turner, 1995) and may be more consistent with motor disinhibition as opposed to impulsivity. Reaction time is another measure frequently reported with CPTs. Reaction time or response latency is believed to reflect the speed of processing as well as the speed of motor responding. For example, a child may demonstrate increased omission errors and a slower rate of responding without an associated increase in commission errors; this type of pattern may be interpreted as supporting a hypothesis of difficulty with the allocation of information-processing resources (Eliason & Richman, 1987). Reaction time is also important when considering the consistency or variability of the individual's performance over time. Some CPT programs generate the standard deviation of the reaction time across blocks as a measure of consistency in responding and the ability to sustain attention over time. Alternatively, some researchers report the standard error of the reaction time (e.g., Conners, 1992, 1995; Conners & MHS Staff, 2000) as an indicator of the consistency or variability of responding over time. Still others use the standard deviation of the standard error over time as an indication of consistency (Levin et al., 1996). Although the use of variability of reaction time may be confounded to a large extent with IQ (see

Jensen, 1982; Vernon, 1989), this variability statistic is nonetheless sensitive to central nervous system trauma and deserves considerably more study.

Related to consistency, the vigilance decrement or the extent to which the individual's accuracy declines over the course of the task is of interest as well (Parasuraman, 1984a, 1984b) and is considered an index of sustained attention (Cohen et al., 1999). Rather than using differences in the reaction time as a measure of consistency, some clinicians focus on comparisons of correct and incorrect responses over differing blocks of time within the same administration. To this end, some CPTs provide comparative information from blocks at the beginning of the task and at the end of the task either directly (e.g., Conners, 1992, 1995) or via derived quotients (e.g., Allen, 1993; Levav, 1991; Sandford & Turner, 1995; Slicker, 1991).

### Signal Detection Theory

As an alternative to direct performance scores, some clinicians incorporate signal detection theory (SDT) in generating performance indexes for interpretation (e.g., Klorman, Brumaghim, Fitzpatrick, & Borgstedt, 1991; Liu, Hwu, & Chen, 1997). The basic premise of SDT is that the decision to respond is based on the child's setting a certain standard or criterion for responding. SDT variables of sensitivity (also referred to as  $d'$  or  $d$ -prime) and response bias (also referred to as Beta) are based on signal to noise (i.e., target to nontarget) ratios (Cohen, 1993a).

The calculation of these scores is based on the distributions of the child's responses to both signal (targets) and noise (nontargets). For the CPT, sensitivity is derived from the mean distribution of responses to both targets and nontargets; it is equivalent to the difference between the sum of the distributions for both targets and nontargets, and nontargets alone (see Swets, 1973, 1984). Sensitivity is believed to represent the likelihood that the individual will detect the signal (respond to the target) when it is presented or the ability to discriminate targets from nontargets. As such, sensitivity is dependent on both the characteristics of the stimulus and the sensitivity of the individual. Related to sensitivity, response bias or response style is believed to reflect the extent to which the individual is being conservative

or impulsive in responding. Response bias is presumed to relate to the strategy used in making the decision to respond or the individual's response style. For example, if the individual tends to be cautious and conservative in responding, the individual is more likely to miss signals (increased omission errors) but to have fewer commission errors. On the other hand, if the response style is less cautious, the individual may have decreased omission errors but increased commission errors.

It has been argued that the sensitivity and bias indexes may be more sensitive to differences in performance on the CPT than omission or commission errors (Lam & Beale, 1991) and that SDT procedures may be particularly useful in neuropsychological assessment (Cohen, 1993c). SDT allows for the determination of response characteristics based on the frequency of the targets and is easily converted into indices for assessing the vigilance decrement and response inconsistency (Cohen et al., 1999).

Although many clinicians and researchers use the measures of sensitivity and response bias in reporting CPT scores, concerns have been raised by others as to the applicability of SDT to the CPT paradigm (e.g., Jerison, 1967; Parasuraman, 1979). In particular, it has been argued that the CPT does not meet the suggested criterion for SDT application in that it does not require successive discrimination of stimuli, but relies on sequential processing (Parasuraman & Davies, 1977). Parasuraman (1984a, 1984b) also argued that for a decrement in sensitivity to occur, the task must be of sufficient duration (i.e., 30–45 minutes). Alternatively, R. A. Cohen (1993c) pointed out that those CPTs with an increased emphasis on reaction time and test-generated changes in the ISI are more likely to result in vigilance decrements and should approximate the traditional SDT task.

### **Commercially Available CPTs**

As can be seen by the number of variations to the CPT paradigm and variables reported, the potential number of CPTs is limited only by one's imagination. Not all the CPTs reported in the literature, however, are commercially available. From the commercially available CPTs, four have been selected for inclusion in this chapter based on the differences in CPT parameters among the four tasks. The four tasks are the Conners CPT-II

(Conners & MHS Staff, 2000), the Gordon Diagnostic System (GDS; Gordon, 1983), the Integrated Visual and Auditory CPT (IVA; Sanford & Turner, 1994–1999), and the Test of Variables of Attention (TOVA®; Greenberg, 1988–1999) and TOVA®—Auditory (TOVA-A; Greenberg, 1996–1999).

Of these, only the GDS is not computerized. Some of these measures have been reviewed in more detail elsewhere (see Dumont, Tamborra, & Stone, 1995; Lowe, Reynolds, Riccio, & Moore, 1999; Riccio et al., 2001). Although some of these programs provide an option for customization, description of the standard paradigm and scores generated is the focus here.

#### **Conners Continuous Performance Test—II**

The standard version of the CPT-II, like the earlier version (Conners, 1992, 1995) is a 14-minute visual CPT. A not-X CPT, the individual is required to respond when a letter appears on the computer screen, except for when the letter “X” (i.e., the target) appears; when the letter “X” appears, the individual is to inhibit the response. This format is intended to ensure a greater number of responses and therefore decreased chance error (Conners, 1992, 1995; Conners & MHS Staff, 2000). The targeted and nontargeted stimuli are randomly shown for 250 milliseconds. The CPT-II standard paradigm consists of six blocks, with each block divided into three subblocks. For the three subblocks within a block, the ISI may be 1,000, 2,000, or 4,000 milliseconds; the order of the three different ISI subblocks varies from block to block. There is no distractor condition on the CPT-II, and the stimuli are not blurred or degraded in any way.

The CPT-II can be used for individuals from age 4 through adulthood. A short practice test is available to help the client become familiar with the paradigm (Conners & MHS Staff, 2000). Instructions are presented to the examinee on screen. Directions for task administration include an additional prompt to respond both quickly and accurately and the provision of a single prompt to redirect the examinee if needed. The manual clearly states that the examiner is to remain present during task administration. The CPT-II manual (Conners & MHS Staff, 2000) provides a description of the normative data ( $N = 1920$ ) for the standard

paradigm. The normative or general data are described as coming from 30 sites across the United States and including three provinces of Canada. A breakdown by gender for the general sample indicates that males under age 18 comprised 52.5% of the sample. Ethnicity of the general sample by age level is not provided; for the complete sample of adults and children, 47% of the sample were white, 27% were black, and 21.4% were other. Additional information specific to the size of the population of the cities or type of setting (i.e., rural, suburban, and urban) where the data were collected, educational level, and socioeconomic status were not provided. The limited information on the sample by age level hinders the extent to which results can be generalized to populations of differing educational levels, socioeconomic status (SES), or ethnicity. In addition to the general sample, clinical data were collected for 271 children with ADHD ranging in age from 6 to 17 years. For this sample, 75.3% of the sample was male. Ethnicity, SES, or educational level are not provided for this clinical sample. A total of 13 indices are generated from the CPT-II for use in interpretation; due to the skewed distributions, computation of all variables use logarithmic transformations. Variables include the more commonly reported correct hits, omission errors, commission errors, and reaction time. The CPT-II also provides the variables from SDT of sensitivity ( $d'$ ) and response bias (Beta). Several measures relate specifically to the consistency or variability of the individual's performance across blocks and subblocks (e.g., standard error of reaction time by block, by ISI, and for the entire test). Finally, the CPT-II provides an overall index for the individual's performance that is a weighted score derived from all the other indices as well as a Confidence Index that indicates the likelihood of attention problems. The default calculation of the Confidence Index (compared to the clinical sample) is designed to optimize the overall hit rate and assumes a base rate of 50% for clinical status (Conners & MHS Staff, 2000). The option is available, however, to select a classification criteria that minimizes either false positives or false negatives. The CPT-II indices are presented in raw score, percentile, and T-score formats. All variables are presented such that high scores indicate

a problem in order to eliminate possible confusion. Data are clustered into eight age groups for both samples, with 2-year age intervals for children and youth. For the normative data, subjects were not excluded based on prescreening, and at some sites, no prescreening was conducted.

### 13. Continuous Performance Tests 299

Thus, some cases in the general sample may have been clinically significant; this is believed to provide more accurate representation of the general population. Because of the potential inclusion of some clinical cases in the general sample, it is recommended that a T-score of 60 (or a percentile of 90) be used as the cutoff for determining impaired performances (Conners & MHS Staff, 2000).

#### Gordon Diagnostic System

Of the commercially available CPTs, the GDS (Gordon, 1983) has been available for the longest amount of time and is probably the most frequently used CPT in research studies. The GDS is a microprocessor unit, as opposed to a computer software program, that generates 11 separate tasks. Of these tasks, there are three basic paradigms—the delay task, distractibility task, and vigilance task. Both the vigilance and distractibility tasks are CPTs and there is more than one version of the distractibility and vigilance tasks available (Gordon, 1986a, 1986b; Gordon & Mettelman, 1988; Gordon Systems, 1991).

The vigilance task is believed to tap into cognitive skills, behavioral regulation, and motivation (Gordon Systems, 1986). The prototype for the vigilance task is the AXCPT in a visual format with numbers as stimuli (Gordon & Mettelman, 1988; Gordon Systems, 1991). The children's standard version of the vigilance task lasts 9 minutes and requires the child to press a button every time a two-number target combination (a "1" followed by a "9" or in the alternate version a "3" followed by a "5") is presented. For both versions, the numerals are displayed for 200 milliseconds, with a 1,000-millisecond ISI (Gordon, 1986b; Gordon & Mettelman, 1988; Gordon Systems, 1991). Both of these versions are for use with children ages 6–16 years. Another version of the vigilance task is for use with preschoolers (i.e., 4- and 5-year-old children) and is shorter (6 minutes). In the preschool version, the vigilance task is an

X-CPT. As with the children's version, the stimuli are displayed for 200 milliseconds; the ISI for the preschool version is increased to 2,000 milliseconds (Gordon Systems, 1991). Regardless of the version (i.e., children's standard and alternate versions and preschool version), the vigilance task is divided into blocks, so that a child's performance may be monitored every 3 minutes, thus providing insight into the vigilance decrement over time (Gordon & Mettelman, 1988; Gordon Systems, 1991).

The second CPT of the GDS is the distractibility task. The distractibility task incorporates the AX-CPT of the vigilance task. At the same time that the stimuli are presented, however, digits are displayed simultaneously on either side of the target stimulus (i.e., distractors). As such, the distractibility task assesses the extent to which the individual can selectively attend to the target stimuli (Gordon Systems, 1991). Stimulus duration, task duration, target frequency, and ISI are the same on the children's distractibility task as for the children's vigilance task. There is no preschool version of the distractibility task.

In addition to the two CPT-type tasks (vigilance and distractibility), the GDS includes one other task—the delay task. The delay task is administered prior to the vigilance and distractibility tasks. The delay task serves as a “warmup” for the vigilance and distractibility tasks. The delay task is not a CPT paradigm but is designed as a measure of impulse control. On this task the child can earn points by inhibiting a response. The delay task takes less than 9 minutes. There is no separate preschool version for the delay task.

The manual for the GDS includes instructions for administration, tables (Gordon Systems, 1991), an interpretive guide (Gordon Systems, 1986), and a technical guide (Gordon Systems, 1987). The order in which the GDS tasks are to be administered and the standard instructional set that is to be used (Gordon & Mettelman, 1988) are provided. The GDS provides practice trials for the vigilance and distractibility tasks that are slower versions of the tasks. During administration of the GDS, directions clearly state that the examiner will remain present (Gordon Systems, 1991).

Normative data for the preschool and children's versions of the GDS can be found

in the work of Gordon and Mettelman (1988) as well as in the manual (Gordon Systems, 1991) and is limited in scope. The normative sample for the GDS vigilance task (standard children's version) consists of 1,019 nonreferred children, ranging in age from 6 to 16 years of age; when the preschool sample of 4- and 5-year-olds is included, it increases the normative sample to 1,266 children; children with psychological, learning, neurological, attention, and impulse control problems were excluded from the normative sample (Gordon & Mettelman, 1988). Unfortunately, the geographic region of the normative sample is restricted to Syracuse, New York (91% of the sample) and Charlottesville, Virginia (9% of the sample) areas (Gordon Systems, 1987). Information on the ethnic composition of the sample was not provided.

For the vigilance and distractibility tasks, variables of performance include correct hits, omission and commission errors, and reaction time (latency) for each block as well as the average latency for the entire task. The slope of the reaction time across blocks (block variability) is used to assess consistency of performance (Gordon Systems, 1991). The examiner completes the test form, records the results from the microprocessor, and computes the block variances themselves. Tables include raw score, percentile, and thresholds. The threshold tables were derived based on the percentile information with scores divided into three categories: abnormal (5th percentile or lower), borderline (6th to 25th percentile), and normal (26th percentile and above). At the preschool level, the norms are provided in 1-year age intervals for 4- and 5-year-olds. For the children's version, norms are in 2-year age intervals for 6- to 11-year-olds, and in a single 5-year age interval for 12-16-year-olds. As gender effects were reported to account for only 2% of the variance, separate gender-based norms were not deemed necessary (Gordon & Mettelman, 1988; Gordon Systems, 1987).

**Integrated (or Intermediate) Visual and Auditory Continuous Performance Test**  
The IVA is a 13-minute X-CPT, and unlike other CPTs discussed here, the IVA incorporates both auditory and visual modalities within the same task. As such, the IVA requires the individual to shift attentional

modalities within the same task. On the IVA, the child is required to respond to a visual or auditory target stimulus (i.e., the number “1”) and to refrain from responding when the nontarget stimulus (i.e., the number “2”) is presented in either a visual or auditory format. The order of auditory or visual presentation of stimuli is fixed and the target and nontarget stimuli are presented in a pseudo-random pattern with a 1,500 milliseconds ISI (Sandford & Turner, 1995). Because there is only a single nontarget stimulus, the discrimination task is simpler than that involved in the CPT-II or GDS. In addition to including auditory presentation, the IVA provides differing target frequency conditions. The target frequency changes by blocks in order to elicit omission and commission errors such that for alternating blocks of the task, there is a high frequency of targets; for the other half of the task, there is a low frequency of targets. The target to nontarget ratios as well as the visual and auditory stimuli changes occur in a counterbalanced design in order to reduce fatigue and to control for learning effects (Sandford & Turner, 1995).

The normative sample for the IVA consists of 781 individuals (358 males and 423 females) ranging in age from 5 years through adulthood. The sample was comprised of individuals with no known learning, psychological, neurological, attention, or impulse control problems (Sandford & Turner, 1995; Turner & Sandford, 1995a, 1995b). Normative data can only be accessed through a “read only” file. Normative data are divided into 2–10-year intervals depending on the age. Information on the geographic region(s) where the normative data were collected, education level, SES, and ethnicity of the sample are not provided in the manual.

In describing administration of the IVA when normative data were collected, the manual notes that the same type of mouse, 14-inch screen, and headphones were used for all subjects (Sandford & Turner, 1995). The manual includes specific directions with regard to examinee distance from the screen, center of the screen relative to examinee eye level, and so on. The extent to which variation from any of these conditions would affect resulting performance is not indicated. Directions are provided by following computer prompts and directions

by the examiner. There is no specific state-  
13. Continuous Performance Tests 301  
ment relating to examiner presence or absence  
during the test but the provision of  
additional prompts during the task would  
suggest that the examiner remains in the  
room during the test administration. A  
warm-up session is provided followed by a  
practice test to allow the child to become familiar  
with the operation of the mouse button.  
Following the practice test, the regular  
test is administered. A cool-down session  
follows the regular test. Reaction times are  
recorded during both the warm-up and  
cool-down sessions (Sandford et al., 1995a,  
1995b; Sandford & Turner, 1995; Seckler,  
Burns, Montgomery, & Sandford, 1995;  
Turner & Sandford, 1995a, 1995b).  
The IVA has 11 raw score scales, 6 quotient  
scales, and 2 composite scores which  
are provided separately for visual and auditory  
portions presentation (Sandford &  
Turner, 1995; Seckler et al., 1995; Turner &  
Sandford, 1995a, 1995b). The two composite  
scores (Response Control Quotient, Attention  
Quotient) are combined to yield a  
Full Scale score as well. Each of the two  
composite scores is derived from three of  
the six quotient scores. The Response Control  
Quotient (RCQ) incorporates the Prudence  
(avoidance of commission errors),  
Consistency (minimal variability in response  
time), and Stamina (response time maintained  
across testing) quotients. In contrast,  
the Attention Quotient is derived from the  
Vigilance (avoidance of omission errors),  
Focus (number of response time outliers  
across testing), and Speed quotients (Sandford  
et al., 1995a, 1995b; Seckler et al.,  
1995; Turner & Sandford, 1995a, 1995b).  
Fine motor/hyperactivity is an additional  
scale and incorporates anticipatory responses  
or continued pressure on the mouse (i.e.,  
failure to release). Additional variables include  
Comprehension (a measure of random  
errors), Persistence (decrement in performance  
related to motivation or arousal),  
and Sensory/Motor (slow reaction time that  
may impair performance); these are considered  
validity measures for the IVA. Finally,  
performance differences based on target frequency  
are reflected in the Readiness score,  
whereas differences in performance based  
on modality are reflected in the Balance  
score. Results of the IVA are presented in  
raw scores as well as in standard scores with

a mean of 100 ( $SD = 15$ ) with numeric as well as graphic formats (Sandford & Turner, 1995).

#### Test of Variables of Attention

The TOVA is an X-CPT that is available in separate visual (TOVA) and auditory (TOVA-A; Greenberg, 1996–1999) versions (Cenedela, 1996; Greenberg & Kindschi, 1996; Greenberg, Kindschi, & Corman, 1999; Lark et al., 1996). The TOVA includes both a clinical version and a briefer screening, or preschool version. Both the TOVA and TOVA-A require an individual to respond whenever the target stimulus is presented. For the TOVA, the target stimulus is a colored square with a smaller square contained within and adjacent to the top edge of the larger square while the nontarget stimulus has the smaller inscribed square adjacent to the bottom edge of the larger square. Thus, the stimulus feature of importance is that of position or placement. For the TOVA-A, two audible tones are used as stimuli, one as the target, and one as the nontarget. As with the IVA, because there are only two stimuli (one target, one nontarget), the difficulty level for discrimination required is minimal.

The stimuli on both the visual and auditory tasks differ from other CPTs in that the stimuli are non-language-based; this is considered important in that otherwise language-based learning problems could confound interpretation of CPT performance (Greenberg & Waldman, 1993). The clinical versions of the TOVA and TOVA-A are approximately 22 minutes in duration. As with other CPTs, the preschool version (for 4- and 5-year-olds) is shorter (11 minutes) in duration. The clinical versions are composed of four intervals whereas the preschool version has two intervals. For both the TOVA and TOVA-A, the stimuli are presented for 100 milliseconds with an ISI of 2,000 milliseconds. As with the IVA, target frequency is manipulated on the TOVA and TOVA-A. For the first half of the test, the target stimulus is randomly displayed on 22.5% of the trials (stimulus infrequent condition). In the second half of the test, the target stimulus is shown on 77.5% of the trials (Greenberg & Waldman, 1993; Lark et al., 1996). The three manuals (Cenedela, 1996; Greenberg et al., 1999; Lark et al., 1996) include standardized

instructions to be given to the examinee (Leark et al., 1996). The examiner remains in the room during the administration of the task. A 3-minute practice test is administered prior to the beginning of the regular test (Leark et al., 1996).

The normative sample for the TOVA is composed of two samples, the original sample and a second sample that was later added to the original sample. Taken together, the TOVA combined normative sample consists of 1,590 individuals, 712 males and 884 females, ranging in age from 4 years through adulthood. Separate norms are available by gender; however, norms stratified on the basis of SES and ethnicity are not provided. Greenberg and Crosby (1992) recommended the use of caution in interpreting the performance of individuals who are members of minority groups or who are from varying levels of SES. The current normative sample for the TOVA-A is limited to 2,551 children ages 6 to 19 years recruited from public schools in Minneapolis, Minnesota; the sample is described as predominantly (99%) Caucasian. As with the TOVA, SES and educational level are not reported. The norms for both the TOVA and TOVA-A are divided into 1-year intervals through age 19 (Greenberg et al., 1999; Leark et al., 1996).

The TOVA and TOVA-A provide multiple variables of performance including omission and commission errors, response time, the standard deviation of the response time, and sensitivity ( $d'$ ). It is important to note that "omission errors" on the TOVA is a measure of relative accuracy derived from the number of correct responses to the difference between the number of targets possible minus the number of anticipatory errors (Greenberg et al., 1999; Leark et al., 1996). The commission error score is based on the number of responses to nontargets relative to the total number of nontargets minus the number of anticipatory responses. For these computations, anticipatory errors are defined as those responses made within 200 milliseconds of the appearance of the stimulus regardless of whether the stimulus is a target or nontarget (Greenberg et al., 1999; Leark et al., 1996); the number of anticipatory errors are reported as well. The program also provides the number of multiple responses and postcommission error response time (Cenedela,

1996; Greenberg et al., 1999; Greenberg & Waldman, 1993; Lark et al., 1996). Finally, the ADHD scale provides a comparison of the examinee's scores to persons of the same age and gender in the database who had a diagnosis of ADHD (Cenedela, 1996). Results are reported as raw scores, percentages, standard scores (mean of 100, *SD* = 15), and standard deviations. Group statistics (means and standard deviations) are provided by age for both the TOVA and TOVA-A with separate norms provided for each age group by gender (Greenberg et al., 1999; Lark et al., 1996). An additional feature of both the TOVA and TOVA-A software is that the results can be saved and subsequent administrations compared for a given individual over to facilitate monitoring of treatment (Cenedela, 1996).

#### **THE CPT AS A DIAGNOSTIC TOOL**

The increased use and the plethora of CPTs available suggest that the basic CPT paradigm has been accepted by clinicians and researchers as a measure of attention and executive control. Not only is some version of the CPT the most frequently used laboratory measure of attention (DuPaul et al., 1992), but the CPT has been described as the "gold standard" for measuring sustained attention (Fleming, Goldberg, & Gold, 1994, p. 205). In recent years, many third-party payers (e.g., Blue Cross/Blue Shield of Texas) seem to think the CPT is a form of "gold standard" or, at the very least, a powerful tool for use in diagnosis of psychopathology, and ADHD in particular. The extent to which the use of the CPT in diagnosis is appropriate will be explored next.

#### **Sensitivity and Specificity**

Over the past 40 years, considerable research has been conducted using CPTs with children diagnosed with ADHD (e.g., August & Garfinkel, 1989; Barkley, Grodzinsky, & DuPaul, 1992; Chee et al., 1989; Halperin, Matier, Bedi, Sharma, & Newcorn, 1992), learning disabilities (e.g., Beale et al., 1987), traumatic brain injury (e.g., Kaufmann, Fletcher, Levin, & Miner, 1993), metabolic disorders (e.g., Anderson, Siegel, Fisch, & Wirt, 1969), autism (Garretson, Fein, & Waterhouse, 1990), low birth weight (Katz et al., 1996), conduct disorder (Chee et al., 1989), and various other disorders.

When considering the usefulness of a

measure or a family of measures in the diagnostic process, it is important to consider the sensitivity and specificity of the measure. *Sensitivity* refers to the ability of a diagnostic tool or procedure to detect a disorder when the disorder is present. Sensitivity can be expressed mathematically as the likelihood that a given method (e.g., the CPT) will detect a disorder when the individual engaging in the task in fact has a disorder (i.e., a true positive). For example, if the question is specific to ADHD, individuals with ADHD may have high rates of commission errors as compared to individuals who do not have ADHD. If sensitivity levels are too high, overdiagnosis (a high rate of false positives) is likely to occur; if sensitivity levels are too low, underdiagnosis (a high rate of false negatives) is likely to occur. The complement to sensitivity, *specificity* is the ability of a procedure to detect the absence of a specific disorder when it is not present. Mathematically, it is the likelihood that a given method will indicate accurately that a disorder is not present (i.e., a true negative). Specificity depends on the capability of the procedure (e.g., a CPT) to differentiate among disorders with overlapping symptoms. Many tests and procedures demonstrate high sensitivity but lack in specificity. This is a concern because in clinical practice the question is not usually as simple as “does this person have a disorder or is this person normal?” but to identify which of many disorders an individual may have. Unfortunately, most of the research on the diagnostic efficacy of CPTs has been conducted comparing a clinical group (e.g., children with ADHD) to a control, nonpsychiatric group. These studies may not provide information on specificity, but they do allow appraisal of the CPT as providing a criterion for the diagnosis of various disorders.

#### **ADHD and CPT Performance**

Given that the symptoms associated with ADHD in childhood include inattention, hyperactivity, and impulse control problems, the notion of the CPT as a diagnostic tool for ADHD appears plausible on the surface. Of the studies reporting on CPT performance of children with ADHD, or attention deficit disorder (ADD), the majority compare children with ADHD to presumably normal children or nonclinical groups. In most of these studies, results fairly consistently yield significant between group differences

in CPT performance on all or nearly all variables. Notably, on the CPT-II, the ADHD group was not found to make more commission errors than a nonclinical group (Conners & MHS Staff, 2000). When demographic variables (age, SES) and verbal ability were controlled for, however, the likelihood of finding differences between groups decreased somewhat (e.g., Koriath, Gualtieri, van Bourgondien, Quade, & Werry, 1985; Werry, Elkind, & Reeves, 1987). Few of the studies included classification rates (accuracy in identifying children as normal or ADHD based on CPT performance) or sensitivity and specificity coefficients. Importantly, when classification rates are reported, the samples are relatively small and this may result in inflated estimates of diagnostic accuracy (Willson & Reynolds, 1982). For example, using the Conners CPT (Conners, 1992, 1995), children with ADHD were compared to the general population (normative sample); results indicated a 13.0% false-negative rate and a 12.9% false-positive rate (Conners, 1992, 1995). Using the GDS with children with ADHD, children with other disorders, and a control group, results indicated approximately 70% agreement with parent and teacher ratings of children with ADHD depending on the age, the rater, and the scale (Gordon Diagnostic Systems, 1987). Although 70% agreement is better than chance, it should be noted that in 30% of the cases, results of the GDS were not consistent with diagnosis based on behavior ratings. Additional studies indicate that for children diagnosed with ADHD in comparison to a normal control group, the IVA was in agreement with group membership in 92% of the cases (Sandford & Turner, 1995). With the TOVA, depending on the cutoff used for predicted group membership, and children with either ADHD or no diagnosis, sensitivity quotients ranged from .60 to .76 (Leark et al., 1996). Using the Conners CPT, 70–75% classification accuracy has been reported for ADHD and other clinical groups (Czerny, O’Laughlin, & Griffioen, 1999). Specificity and sensitivity for the Conners CPT was reported to be 83% and 82%, respectively, for children and youth with ADHD as compared to a nonclinical sample (Conners & MHS Staff, 2000). Across studies, CPTs identified normal

(non-ADHD) children with greater accuracy than they identified children with ADHD (Anastopoulos & Costabile, 1994; Barkley & Grodzinsky, 1994; Harper & Ottinger, 1992).

Notably, few studies compared the performance of children with ADHD by subtypes using either the subtype system of the third edition of *Diagnostic and Statistical Manual for Mental Disorders* (DSM-III; American Psychiatric Association, 1980) or the DSM-IV (American Psychiatric Association, 1994). Results of these studies (e.g., Barkley, DuPaul, & McMurray, 1990; Barkley & Grodzinsky, 1994; Forbes, 1998; García-Sánchez, Estérez-González, Suárez-Romero, & Junqué, 1997; Holcomb, Ackerman, & Dykman, 1985; Johnson, 1993) are equivocal with regard to the ability of CPTs to aid in subtype differentiation. Further, the majority of studies included children with a mean age between 6 and 12 years. Fischer, Newby, and Gordon (1995) compared the performance of younger and older children with ADHD relative to a normal control group. Results indicated that although there was a correct classification rate of 81% for the younger age group, this rate decreased to 20% for children ages 12–17 years. These findings have not been replicated; however, they suggest that the sensitivity of the CPT to ADHD may decline with increased age.

The differentiation of ADHD from other disorders is complicated by the high level of co-occurrence of ADHD with other disorders such as learning disabilities and conduct disorder (e.g., Hynd et al., 1995; Riccio & Jemison, 1998). In addition to studies comparing children with ADHD to a nonclinical control group, a minority of studies have compared the performance of children with ADHD to other clinical groups including children with learning disabilities (e.g., Barkley et al., 1990, 1992; Barkley & Grodzinsky, 1994; Chee et al., 1989; Riccio, Cohen, Hynd, & Keith, 1996; Richards, Samuels, Turnure, & Ysseldyke, 1990; Schachar et al., 1988; Tarnowski, Prinz, & Nay, 1986), conduct disorder (Chee et al., 1989; Halperin et al., 1990, 1993, 1995; Koriath et al., 1985), Tourette syndrome (Harris et al., 1995), schizophrenia (Erickson, Yellin, Hopwood, Realmuto, & Greenberg, 1984), and hypoxia (O'Dougherty et al., 1984). Thus far, results of studies comparing

the CPT performance of children with ADHD to those with learning disabilities, for example, are highly inconsistent. This is not surprising in that many studies comparing children with learning disabilities to normal controls suggest that impaired CPT performance is associated with learning disabilities as well as ADHD with no synergistic effects of comorbid ADHD and learning disability (Chee et al., 1989; Robins, 1992). In that children with conduct disorder generally perform comparably to normal control groups (e.g., Koriath et al., 1985), CPTs are somewhat better in differentiating ADHD and conduct disorder; however, inattention and aggression could not be dissociated based on CPT performance (Halperin et al., 1990). Moreover, in other studies, children with conduct disorder were impaired on the CPT (e.g., Schachar et al., 1988). In comparison to other groups (e.g., Tourette syndrome and schizophrenia), children with disorders other than ADHD sometimes performed worse than or equally as poorly as children with ADHD. Children with Tourette syndrome, for example, demonstrate significantly slowed reaction times (e.g., Harris et al., 1995; Shucard, Benedict, Tekok-Kilic, & Lichter, 1997); this finding has potential heuristic value and should be studied further. Thus, although children with ADHD can be distinguished from some samples with reasonable consistency based on CPT performance, it is more difficult, if not impossible, to differentiate them from other diagnostic groups on a reliable basis.

#### **Other Disorders and CPT Performance**

Studies comparing the CPT performance of children with numerous disorders other than ADHD relative to normal controls or to children with other disorders are less numerous,

#### **13. Continuous Performance Tests 305**

but add to what is known about the sensitivity and specificity of the CPTs. For example, in three out of four available studies on children with a diagnosis of schizophrenia, schizophrenia was found to be associated with impaired CPT performance (Erickson et al., 1984; Rund, Zeiner, Sundet, Oie, & Bryhn, 1998; Strandburg et al., 1990, 1994). With less consistency, children with a parent diagnosed with schizophrenia also have been found to demonstrate impaired CPT performance (Asarnow, Steffy, MacCrimmon, & Cleghorn, 1977; Erlenmeyer-Kimling &

Cornblatt, 1978; Nuechterlein, 1983). Generally speaking, although a diagnosis of ADHD tends to be associated with increased commission errors, a diagnosis or family history of schizophrenia tends to be associated with increased omission errors. Although promising, these differences are not sufficiently consistent for diagnostic certainty (Riccio et al., 2001).

Among clinical groups whose performance on the CPT have been studied, research points to significant performance decrements for children with traumatic brain injury (e.g., Crosby, 1972; Katz et al., 1996, Kaufmann et al., 1993) as well as for children with seizure disorders (e.g., Levav, 1991; Miller, 1996). Children with neurofibromatosis also demonstrate deficits in CPT performance, predominantly in the form of slowed reaction time (Eliason, 1988) as with Tourette syndrome. In fact, significant evidence exists that any direct central nervous system (CNS) compromise will result in impaired performance on the CPT (Riccio et al., in press). Children with metabolic disorders such as phenylketonuria (PKU) have been found to perform poorly on CPTs and to evidence a vigilance decrement over time (Anderson et al., 1969; Brunner & Berry, 1987). Of the populations included in the extant literature, CPT performance generally was affected by most externalizing disorders, autism, and schizophrenia as well as most types of CNS dysfunction (see Table 13.2). Negative effects on CPT performance

#### 306 IV. SPECIFIC SYNDROMES AND SYMPTOMS

##### **TABLE 13.2. Summary of Disorders Associated with Impaired Performance on Continuous Performance Tests**

###### Disorder Relevant Studies

ADHD (all subtypes, with or without a comorbid disorder) Anastopoulos & Costabile (1994); August & Garfinkel (1989); Barkley & Grodzinsky (1994); Barkley et al. (1992); Chee et al. (1989); Conners (1992, 1995); Fischer et al. (1995); Forbes (1998); García-Sánchez et al. (1997); Halperin et al. (1992); Holcomb et al. (1985); Johnson (1993); Koriath et al. (1985); Robins (1992); Sykes et al. (1971); Teicher et al. (1996); Trommer et al. (1988); Werry et al. (1987)

Autism Garretson et al. (1990)

Central auditory processing disorder Riccio et al. (1996)

Hearing impaired/deaf Mitchell & Quittner (1996)

Hypoxic/anoxic injuries O'Dougherty et al. (1984)

Intrauterine toxic exposure Hickey, Suess, Newlin, & Spurgeon (1995); Walker (1993)

Low birth weight Katz et al. (1996)

Mental retardation Crosby (1972); Kintslinger (1987)

Neurofibromatosis Eliason (1988)

Phenylketonuria Anderson et al. (1969); Brunner & Berry (1987)

Schizophrenia Erickson et al. (1984); Rund et al. (1998); Strandburg et al. (1990, 1994)

Seizure disorder Levav (1991); Miller (1996)

Tourette syndrome Harris et al. (1993); Shucard et al. (1997)  
Traumatic brain injury Crosby (1972); Katz et al. (1996); Kaufmann et al. (1993)  
Turner's syndrome Romans et al. (1997)

were associated with metabolic disorders and some medical conditions as well. Notably, CPTs were not found to be particularly sensitive to disorders of mood or affect, except during manic episodes or in the presence of psychotic features (Riccio et al., 2001). Equivocal findings exist for samples of children with learning disability, oppositional defiant disorder, and conduct disorder. Although further study is needed in this area, the obvious conclusion from the studies reviewed is that any disorder that results in a compromise of CNS integrity or function is likely to produce decrements in CPT performance. In fact, for this age range (5–17 years), the CPT appears quite sensitive to CNS compromise of varying etiology. If the diagnostic question is whether there is CNS compromise or not, then the CPT seems to do quite well. In practice, however, the question is more often which of a myriad of disorders or syndromes the child has. Across the literature available to date, when studies have included children with a variety of disorders, classification rates dropped precipitously. For example, Kintslinger (1987) found that CPT results resulted in 47.5% classification accuracy for children previously identified as either mentally retarded, learning disabled, behaviorally disordered, or normal. Notably, the best classification accuracy was for the normal children (19 of 20). Other studies with multiple clinical groups report similar findings (e.g., Halperin et al., 1992). In another study, CPT variables misclassified 28% of the non-ADHD group as ADHD and 20% of the ADHD sample as not ADHD (Forbes, 1998)

As noted previously, multiple disorders include similar symptoms, and it would not be surprising to find that CPTs are sensitive to many such disorders. What may be apparent from this review, however, is that virtually any disorder of childhood that disrupts or compromises the integrity of the CNS is likely to result in impaired performance on the CPT (see Table 13.2). In fact, among disorders usually first identified in childhood and adolescence, the CPT is quite sensitive to problems of many varieties. Because of this sensitivity, and as noted by others (Forbes, 1998; Halperin et al., 1992;

Riccio et al., 2001), the use of CPT performance decrements as indicative of ADHD in particular would likely result in children with a variety of psychiatric (and medical) conditions being misdiagnosed if the comparison group is a normal control or nonclinical group. In reality, the more common task facing the clinician is to determine which of a number of similar disorders is the appropriate diagnosis for a client. The complexity of syndromes precludes the definition of *any* disorder based on one or two dimensions or a single measure without other tests and clinical and historical information. Although CPTs lack sufficient sensitivity and specificity for differential diagnosis, they demonstrate sensitivity to disorders that include behavioral symptoms associated with inattention and poor self-regulation (Riccio et al., 2001). CPT performance decrements are not disorder specific, but, rather, CPTs are sensitive to the presence of impulsivity and to the attentional deficits associated with brain dysfunction. As such, although lacking sufficient sensitivity and specificity for differential diagnosis, CPTs possess high degrees of usefulness for objective documentation of symptoms (including treatment monitoring) associated with multiple disorders as well as in ruling out disorders that involve brain dysfunction.

### Validity and Reliability Issues

#### Construct Validity

If CPTs are to be interpreted as measures of attention or executive function, the results generated by the CPT would be expected to correlate significantly with results from other accepted measures; this is one method for establishing the construct validity of a measure's suggested interpretation (Anastasi, 1988; Cohen & Swerdlik, 1999). At the same time, due to the multifaceted nature of attention, a one-to-one correspondence between CPT performance and other measures or methods would not be expected.

Barkley (1998) asserted that the optimal method for assessing attention and self-regulation continues to be direct observations.

Although direct observation may be the most ecologically sound means of assessing attention and executive control, observations in natural settings can be time-consuming.

As a result, in practice, laboratory

#### 13. Continuous Performance Tests 307

measures of attention and executive control are used frequently. Another alternative to

direct observation involves the use of behavior rating scales. If the CPT paradigm is measuring attention and executive control, it would be expected that CPT performance would be significantly correlated with results of direct observation, laboratory measures, and behavior scales.

The majority of studies investigating the relation between CPT scores and results of direct behavioral observations yielded moderate to high correlations (Barkley, 1991; Garretson et al., 1990; Gordon, DiNiro, Mettelman, & Tallmadge, 1989; Harper & Ottinger, 1992; Kupietz & Richardson, 1978). Notably, higher correlations emerged in those studies with clinical samples of children regardless of diagnostic category (e.g., Barkley, 1991; Garretson et al., 1990; Harper & Ottinger, 1992; Kupietz & Richardson, 1978). The research further suggests that CPT performance is correlated with scores from other laboratory measures assessing shifts in attention (e.g., Allen, 1993; Suslow & Arolt, 1997), focused attention (e.g., Burg, Burright, & Donovan, 1995; Das, Snyder, & Mishra, 1992), selective attention (e.g., Das et al., 1992), motor activity (e.g., Allen, 1993; Trommer, Hoepfner, Lorber, & Armstrong, 1988), and impulsive behavior (Kardell, 1994; Slicker, 1991). The degree of association between CPT variables and a number of other measures varies depending on the CPT parameters and population. For example, omission errors on a visual X-CPT with typically developing children correlated moderately ( $r = .52$ ) with the Stroop (Das et al., 1992). With a different visual X-CPT and a combined clinical and control group, however, the correlation between omission errors and the Stroop was negligible ( $r = .05$ ). Across studies, however, the relative consistency of moderate correlations of CPT variables to direct behavioral observations and other measures of attention would support the notion that the CPT paradigm is measuring some aspect of attention.

A number of studies have examined the association between CPT performance and various behavior rating scales (e.g., Allen, 1993; Das et al., 1992; DuPaul et al., 1992; Garretson et al., 1990; Gordon et al., 1989; Halperin, Wolf, et al., 1991; Harper & Ottinger, 1992; Lam & Beale, 1991; Sandford et al., 1995b; Slicker, 1991; Teicher, Ito, Glod, & Barber, 1996; Wherry et al., 1993).

Across studies, the relation between CPT performance and behavior scales is variable depending on the subscale, the rater, and CPT parameters.

The degree to which hyperactivity scales or subscales were associated with CPT performance tended to be stronger than attention (or inattention) or impulsivity. There was a moderate degree of association found between commission errors and teacher ratings of hyperactivity (Barkley, 1991; Kupietz & Richardson, 1978). When comparing modality of the CPT administration, the correlation between hyperactivity and commission errors was stronger for the visual CPT format as compared to the auditory format (Kupietz & Richardson, 1978). Studies that included correlations of subscales of inattention or distractibility with CPT scores obtained correlation coefficients generally in the  $\pm 0.45$  range (Lowe et al., 1999). In contrast, correlation of CPT results to scales of impulsivity or dyscontrol tended to be lower. Coefficients were higher when impulsivity was combined in the scale with inattention or hyperactivity but were not as high as when "pure" attention or hyperactivity scales were used (e.g., Slicker, 1991; Teicher et al., 1996).

Using the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992), Floyd's (1999) results indicated substantial correlations of the Hyperactivity and Attention Problems subscales of the Parent Rating Scale (PRS) with omission errors. Attention Problems correlated moderately with commission errors as well. For the BASC Teacher Rating Scale ((TRS), correlation coefficients were moderate for Hyperactivity and Attention Problems with commission errors. Consistent with the findings of sensitivity to a myriad of disorders, commission and omission error scores also correlated moderately with other scales, including aggression and depression (Floyd, 1999).

At the same time, a number of studies have examined the relationship between cognitive ability and other functional domains with CPT performance (e.g., Aylward, Gordon, & Verhulst, 1997; Chae, 1999; Hoerig, D'Amato, Raggio, & Martin, 1998; Kardell, 1994). Overall, CPT performance has not been found to be correlated significantly with verbal IQ, performance

IQ, or Full Scale IQ. Campbell, D'Amato, Raggio, and Stephens (1991) concluded that CPTs measure attention or strategies associated with achievement to some extent; however, the level of association between CPT results and achievement test results varies depending in part on the modality of the CPT. Regardless of modality, the correlations are lower than expected for prediction of achievement from CPT performance. Moderate to strong levels of association were found with some motor tasks (e.g., Allen, 1993), suggesting that any motor difficulties may need to be considered when interpreting CPT results. Thus, the relationship between CPT performance and direct observations, rating scales, and laboratory measures of attention support the ecological validity of the CPT paradigm; at the same time, the level of association of CPT performance with other measures suggests that CPTs have some unique characteristics not shared with other measures and, thus, may provide information not otherwise available from traditional tests. One of the reasons so much emphasis is placed on attention is the belief that attention is the precursor to memory (Cohen, 1993b; Sohlberg & Mateer, 1989). Because of the presumed association between memory and attention, the degree of association between CPT variables and measures of memory would seem important, but to this point, there is a paucity of research in this area.

#### **Temporal Stability**

Independent researchers have conducted studies on the temporal stability of the scores derived from different CPT paradigms (e.g., Finkelstein, Cannon, Gur, Gur, & Moberg, 1997; Halperin, Sharma, et al., 1991; Harper & Ottinger, 1992; Rosvold et al., 1956). Reliability coefficients vary considerably depending on the CPT parameters, the CPT variable, and the test–retest interval. Of the variables of interest, the highest temporal stability has been found for omission errors or correct hits. For example, with a test–retest interval of at least 1 week, Harper and Ottinger (1992) reported test–retest reliability estimates for omission errors of .55 for an X-CPT and .80 for an AX-CPT. Halperin, Sharma, and colleagues (1991) reported moderate test–retest reliabilities for an AX-CPT with 34 nonreferred boys (7–11 years old) and an interval between test and retest of approximately

4.8 months. The test–retest reliability estimates for correct hits, reaction times, omission errors, inattention scores, and impulsivity scores ranged from 0.65 to 0.74, suggesting moderate temporal reliability. Across studies, test–retest reliability estimates for commission errors have been found to be lower than those for omission errors or correct hits (Halperin, Sharma, et al., 1991; Harper & Ottinger, 1992). Study of the temporal stability of the GDS children’s standard version is provided in the technical manual (Gordon Systems, 1987) and in Gordon and Mettelman (1988). The test–retest reliabilities were reported for two groups of children (nonreferred and referred). Results indicated that the GDS scores for vigilance and distractibility tasks have moderate to high test–retest reliability over the short term (2–22 days) and the vigilance task score has moderate to high reliability long term (1 year). Comparable data for the preschool versions were not available. For the CPT-II, temporal reliability coefficients ranged from low (e.g., hit reaction time by block,  $r = .28$ ) to high (e.g., omission errors,  $r = .84$ ) (Conners & MHS Staff, 2000)

A temporal reliability study with 70 volunteers (mean age of 21.8 years) and test–retest intervals ranging from 1 to 4 weeks resulted in quite inconsistent results depending on the variable (Seckler et al., 1995), suggesting that the temporal stability of some scores is adequate while the stability of other scores is below expected levels. The extent to which these estimates apply for children is unknown. The TOVA scores have been found to have moderate internal consistency with reliability coefficients varying depending on the variable of interest. Similar results were found for the internal consistency of the TOVA-A scores (Lark et al., 1996). In an independent study, Llorente and colleagues (2000) investigated the internal consistency of the TOVA scores with a group of children ( $n = 63$ ) with ADHD. Each half of the test was compared to the other half as well as the whole test;

13. Continuous Performance Tests 309

results suggested strong internal consistency. Moderate temporal stability of the TOVA scores is reported in the manual for 24 presumed normal children with a test–retest interval of 90 minutes, an interval useful for same day testing of medication

effects (Leark et al., 1996). With a longer (4-month) test–retest interval, Llorente and colleagues found temporal reliability estimates for the TOVA scores to range from adequate to moderate depending on the variable of interest.

### Potential Moderator or Mediator Variables

#### Developmental Issues and the CPT

Because attention and executive control are subserved by neurological structures and systems, some understanding of the developmental trajectory of the CNS is critical.

Although the primary cortical areas generally mature by birth (Luria, 1980), secondary and tertiary areas continue to develop postnatally. These include the functional systems involved in learning, memory, emotion, cognition, and language, as well as those systems involved in attention and executive control. Not only is there continued development of these areas, but primary areas as well as the pathways that connect structures within the primary areas are likely to change over time as myelination and cell specialization continue to occur (Merola & Leiderman, 1985; Rutter, 1981; Vygotsky, 1980). Although the knowledge base regarding neurodevelopment has increased in recent years, the majority of this knowledge is grounded on observations and informal assessment of individuals with identified brain damage as opposed to typically developing children (Reynolds, 1997).

Developmental differences have been found to be associated with CPT performance in a number of studies across the lifespan (e.g., Conners, 1995; Greenberg & Crosby, 1992; Halperin, Sharma, et al., 1991; Holcomb et al., 1985; Klorman et al., 1991; Marks, Himelstein, Newcorn, & Halperin, 1999; Romans, Roeltgen, Kushner, & Ross, 1997; Sandford & Turner, 1995; Turner & Sandford, 1995a, 1995b). Developmental differences in children's performance on CPTs have been reported by a number of researchers using normative samples (e.g., Gordon & Mettelman, 1988; Greenberg & Waldman, 1993; Turner & Sandford, 1995a) as well as with clinical samples (e.g., Raggio & Whitten, 1994). Based on the available research base, after age 5, there is a gradual improvement in performance as a function of age which stabilizes in young adulthood to be followed by a decline in performance in later adulthood (e.g., Greenberg & Waldman, 1993;

Raggio & Whitten, 1994; Sandford & Turner, 1995). Because of the developmental impact on CPT performance, it is important to incorporate developmental theory into the interpretation process. To meet this end, it has been asserted that normative data need to be provided in relatively small intervals (4- to 6-month intervals) for children and increasing to 1-year intervals in adulthood (Lowe et al., 1999).

#### **Gender and CPT Performance**

Gender, as a moderating variable, has been investigated minimally and research findings do not suggest substantial differences based on gender; however, the results are equivocal (e.g., Driscoll, 1994; Goldstein et al., 1997; Greenberg & Waldman, 1993; Levy, 1980; Wagner, 1987). For example, Levy (1980) administered the X-CPT to 120 boys and 110 girls (3 to 7 years old) and found no significant gender differences between the boys and the girls for omission errors, commission errors, or mean reaction times scores. In contrast, Wagner (1987) examined gender differences in CPT performance for a sample of 83 boys and 34 girls on the X-CPT and AX-CPT. The sample consisted children with ADHD and a control group of presumably typically developing children. Consistent with the findings of other studies (Greenberg & Crosby, 1992; Sandford & Turner, 1995), Wagner's results revealed that boys made more errors of commission than girls. In other studies (e.g., Sandford & Turner, 1995), boys were found to exhibit faster reaction times than girls as well. Based on the potential for age and gender differences to affect CPT performance, the use of appropriate gender and age normative groups has been suggested (Sandford & Turner, 1995).

#### **310 IV. SPECIFIC SYNDROMES AND SYMPTOMS Socioeconomic Status and the CPT**

A number of studies have been conducted to assess the relation between CPT scores and SES in the child population; however, as with gender, the findings are equivocal (Evans, 1988; Gordon & Mettelman, 1988; Levy, 1980; Levy & Hobbes, 1988). Levy (1980) found that the child's ability to complete the CPT was dependent on both the child's age and parental SES with children from the upper class able to complete the CPT at a younger age than those in the lower class. Similarly, Gordon and Mettelman (1988) found a significant relation between

mother's SES and total correct scores on the vigilance task whereas father's SES was significantly related to commission error scores on the distractibility task. Although the correlations between CPT scores and parental SES were significant, SES accounted for a small percentage of the variance. In contrast, Evans (1988) found a nonsignificant relation between CPT performance and SES for 164 children (6 to 14 years old) when age effects were partialled out.

#### **Culture and the CPT**

Bauermeister, Berrios, Jimenez, Acevedo, and Gordon (1990) compared a large Puerto Rican sample to the GDS normative data (Gordon Systems, 1987). The Bauermeister sample consisted of 433 children and adolescents, ranging in age from 6 to 16 years in Puerto Rico. For purposes of comparison, the Puerto Rican children were then matched to randomly selected children from the U.S. normative sample on the basis of age and gender. The performance of the children from the two on the delay and vigilance tasks were compared; the distractibility task was not included. In comparison to the Puerto Rican children, U.S. children had higher efficiency ratios and total correct scores and committed fewer commission errors. Further analyses by age groups revealed statistically significant cultural differences in the 6- to 9-year-old age groups, with the U.S. children making more correct responses and committing fewer commission errors in comparison to the Puerto Rican children. Thus, if U.S. norms were to be applied to the Puerto Rican sample, then a larger proportion of the Puerto Rican children would be identified as having potential attention and impulse control problems (Bauermeister et al., 1990; Lowe et al., 1999).

#### **DISCUSSION AND CONCLUSIONS**

The complexity of the constructs of attention and executive control suggests that no single measure of the behavior will provide information that is sufficient to address all facets. Concerns with attention and executive control occur with high frequency regardless of the type of practice or setting or age of the client. The very nature of attentional and executive control functional systems makes it a formidable task to assess adequately and accurately the integrity of these systems. Further, it has been suggested that when one component of attentional

processes is disrupted, other components of attention or executive control are more likely to be affected due to the interconnectedness of these components (Cohen, 1993a). Assessment of attention and executive control must be multifaceted, paralleling the complexity of the functional systems involved.

The basic paradigm for the CPT was initially designed almost 50 years ago for the assessment of attention and, to a lesser degree, response inhibition or dyscontrol. The best any single measure, including the CPT, can provide are data on specific aspects of attention and executive control that, in conjunction with other measures, can be considered in diagnostic hypothesis generation and in the monitoring of the treatment and rehabilitation process. Since Rosvold and colleagues (1956) introduced the original CPT over 40 years ago, a plethora of CPTs have been developed. Today, there are innumerable variations of the CPT used in clinical and research settings. At the same time, differing CPTs and clinicians report a myriad of scores for use in interpretation. Thus, different CPTs may place different demands on an individual's attention, executive, and memory systems. When the possible numbers of differences is considered, it is clear that CPTs are not a unitary measure (Lowe et al., 1999) but a family of measures (Conners, 1992, 1995) with different parameters and scoring indices. The extent to which the differences in parameters and measures across these four CPTs affect diagnostic considerations is unknown as there is no comparative research of the various CPTs across populations available.

In using the CPT, as with any formal assessment measure, it is important for administration to be standardized consistent with the collection of normative data. As such, it is imperative for the CPT manuals to clearly state the conditions for standardized administration. Differences related to use of space bar versus mouse, sensitivity and response times of various mouse devices, different screen sizes, use of headphones as opposed to speakers, use of laptop versus standard PC, and so on have not been investigated; there is no evidence that these differences do not have an effect on the results obtained. Ballard (1996) reviewed a number of experimenter manipulated

variables that may influence CPT scores and recommended that manuals address these issues (e.g., examiner presence or absence or instructional set). The potential for these changes to be confounds to test performance, possibly invalidating the interpretation of the results, requires adherence to standardization procedures in the administration of any CPT. In addition, given differences in task demands and parameters across CPTs, performance on any one of these CPTs can be interpreted based only on the normative data for those specific task parameters. For these reasons, although the capacity to customize the task may be beneficial for research purposes, clinical use of customized versions of CPTs is not recommended due to the lack of normative data.

Ethical and professional standards demand that psychologists use measures that are technically adequate. As with the developers and publishers of other assessment tools, CPT developers and publishers need to ensure that the measures meet the standards of technical adequacy established by the joint committees and published by the American Educational Research Association (AERA; 1999). In keeping with these standards, the adequacy of the normative data need to be evaluated relative to the population for whom the measure is intended in order to ensure generalization of the interpretation (e.g., Anastasi, 1988; Lowe et al., 1999). As such, the normative sample must be representative, of sufficient size to produce stable values and reduce potential sources of error, stratified across a number of demographic variables, and described in detail (e.g., Anastasi, 1988; Cohen & Swerdlik, 1999).

In the past, limited research has been conducted on the relation between demographic variables and neuropsychological test scores (e.g., Lowe & Reynolds, 1999; Lowe et al., 1999; Reynolds, 1997, in press). Only a small number of CPT studies have been published with individuals from different racial or cultural backgrounds (e.g., Bauermeister et al., 1990). The results of the Bauermeister and colleagues (1990) study suggest that there may be a cultural confound to CPT performance. The use of the normative data (United States) with children from diverse cultures may be problematic and further study of the differences in performance that may be attributable to

cultural differences, with appropriate development of separate norms for differing ethnic groups as necessary, seems warranted. At the same time, the research related to gender as a possible confound has been equivocal. An understanding of the relation between demographic variables and test scores is needed if test results are to be interpreted accurately (e.g., Lowe & Reynolds, 1999; Lowe et al., 1999; Reynolds, 1997). As is common among neuropsychological tests (Reynolds, 1997), a great weakness of CPTs generally is that studies of demographic variables have not been done. Studies of ethnic and of gender bias are sorely lacking for these tests, particularly with regard to specificity and sensitivity of CPT scores, and are desperately needed, particularly given their widespread use (see Reynolds, in press, for methods for such research). Another shortfall of the CPT is the limited number of reliability studies. Overall, moderate to high temporal stability and internal consistency were noted when reliability studies had been completed; however, as different versions of the CPT may be conceptualized as separate measures, and these various CPT paradigms may not measure the exact same construct, reliability estimates from studies using one CPT cannot

#### 312 IV. SPECIFIC SYNDROMES AND SYMPTOMS

be presumed to apply to other CPTs. Studies addressing the comparative reliability of CPT scores across demographic variables and diagnostic groupings are needed if CPT scores are to be used and interpreted with accuracy (Lowe et al., 1999). Without evidence of both internal consistency and temporal stability, any conclusions related to diagnostic considerations or treatment effectiveness become spurious.

The fact that the manuals suggest the use of CPTs in the diagnostic process specific to ADHD as well as other disorders should not be interpreted as indicating high levels of sensitivity or specificity. Clinicians need to be cautious in their interpretation of CPT results. The CPT is only one measure, and multiple sources of information and multiple measures should be used when assessing attention and impulse control problems in order to corroborate CPT findings. Taken together with direct observation, behavior rating scales, and other psychometric tests, the CPT may provide useful information. The CPT is an objective measure that is not

subject to rater bias or observer drift. Based on our review, the level of performance on CPTs may be helpful in ruling out or identifying attentional problems. However, reliance on CPTs as a primary diagnostic tool in determining the presence of a specific disorder (e.g., ADHD) is not warranted and will result in an unacceptably high number of false-positive errors or overdiagnosis of ADHD.

In clinical practice, it is unusual for one to be asked simply to distinguish normal from abnormal, yet this has been the focus of most of the research with the CPT. The real issue is the need to adequately differentiate among the various psychopathologies in order to make appropriate diagnosis and treatment decisions. Given the myriad possible variations to the CPT paradigm, it is possible that some combination(s) will prove more helpful in differential diagnosis than others. Determining if diagnostic algorithms with CPTs exist, however, requires multisite research that includes multiple groups with relevant psychopathologies and at differing developmental levels using different CPT paradigms in conjunction with other measures. This research has yet to be conducted but is needed. In the meantime, the use of CPTs as one component of the assessment process as well as a tool for monitoring attentional functioning holds tremendous promise.

## REFERENCES

- Alexander, D. D. (1973). Attention dysfunction in senile dementia. *Psychological Reports, 32*, 229–230.
- Allen, L. F. (1993). *Developmental delay of frontal lobe functioning: A possible cause of attention deficits in children*. Unpublished doctoral dissertation, Texas A & M University, College Station, TX.
- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Anastopoulos, A. D., & Costabile, A. A. (1994). The Conners' continuous performance test: A preliminary examination of its diagnostic utility. *The ADHD Report, 2*(5), 7–8.
- Anderson, V. E., Siegel, F. S., Fisch, R. O., & Wirt, D. (1969). Response of phenylketonuric children on a continuous performance test. *Journal of Abnormal Psychology, 74*, 358–362.
- Asarnow, R. F., Steffy, R., MacCrimmon, D. J., & Cleghorn, J. M. (1977). An attentional assessment

- of foster children at risk for schizophrenia. *Journal of Abnormal Child Psychology*, 86, 267–275.
- August, G. J., & Garfinkel, B. D. (1989). Behavioral and cognitive subtypes of ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, 28, 739–748.
- Aylward, G., Gordon, M., & Verhulst, S. (1997). Relationships between continuous performance task scores and other cognitive measures: Causality or commonality? *Assessment*, 4, 325–336.
- Baker, D. B., Taylor, C. J., & Leyva, C. (1995). Continuous performance tests: A comparison of modalities. *Journal of Clinical Psychology*, 51, 548–551.
- Ballard, J. C. (1996). Computerized assessment of sustained attention: A review of factors affecting vigilance performance. *Journal of Clinical and Experimental Neuropsychology*, 18, 843–863.
- Barkley, R. A. (1991). The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. *Journal of Abnormal Child Psychology*, 19, 149–178.
- Barkley, R. A. (1998). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (2nd ed.). New York: Guilford Press.
- Barkley, R. A., DuPaul, G. J., & McMurray, M. B. (1990). Comprehensive evaluation of attention deficit disorder with and without hyperactivity as defined by research criteria. *Journal of Consulting and Clinical Psychology*, 58, 775–789.
- 13. Continuous Performance Tests 313**
- Barkley, R. A., & Grodzinsky, G. M. (1994). Are tests of frontal lobe functions useful in the diagnosis of attention deficit disorders? *The Clinical Neuropsychologist*, 8, 121–139.
- Barkley, R. A., Grodzinsky, G. M., & DuPaul, G. (1992). Frontal lobe functions in attention deficit disorder with and without hyperactivity: A review and research report. *Journal of Abnormal Child Psychology*, 20, 163–188.
- Bauermeister, J. J., Berrios, V., Jimenez, A. L., Acevedo, L., & Gordon, M. (1990). Some issues and instruments for the assessment of attention deficit hyperactivity disorder in Puerto Rican children. *Journal of Clinical Child Psychology*, 19(1), 9–16.
- Beale, I. L., Matthew, P. J., Oliver, S., & Corballis, M. C. (1987). Performance of disabled and normal readers on the continuous performance test. *Journal of Abnormal Child Psychology*, 15, 229–238.
- Bock, R. D. (1982). *The role of arousal in Tourette syndrome*. Unpublished doctoral dissertation, New York University, New York.
- Bremer, D. A. (1989). Mini-CPT: A continuous performance test program for the Tandy PC–8 pocket computer. *Behavior Research Methods, Instruments, and Computers*, 21, 11–14.
- Broadbent, D. E. (1953). Noise, paced performance, and vigilance tasks. *British Journal of Psychology*, 44, 295–303.
- Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, 64, 205–215.
- Brumm, V. L. (1994). *Neuropsychological and psychological correlates of marital violence in a clinical sample*. Unpublished doctoral dissertation, University of Southern California, San Diego.
- Brunner, R. L., & Berry, H. K. (1987). Phenylketonuria and sustained attention: The continuous performance test. *International Journal of Clinical Neuropsychology*, 9, 68–70.
- Burg, J. S., Burreight, R. G., & Donovick, P. J. (1995).

Performance data for traumatic brain injured subjects on the Gordon Diagnostic System (GDS) tests of attention. *Brain Injury*, 9, 395–403.

Campbell, J. W., D'Amato, R. C., Raggio, D. J., & Stephens, K. D. (1991). Construct validity of the computerized continuous performance test with measures of intelligence, achievement, and behavior. *Journal of School Psychology*, 29, 143–150.

Carter, C. S., Krener, P., Chaderjian, M., Northcutt, C., & Wolfe, V. (1995). Abnormal processing of irrelevant information in attention deficit hyperactivity disorder. *Psychiatry Research*, 56, 59–70.

Cenedela, M. (1996). *TOVA: Test of Variables of Attention user's manual IBM/PC version 7.0*. Los Alamitos, CA: Universal Attention Disorders.

Chae, P. K. (1999). Correlation study between WISCIII scores and TOVA performance. *Psychology in the Schools*, 36, 179–185.

Chee, P., Logan, G., Schachar, R., Lindsay, P., & Wachsmuth, R. (1989). Effects of event rate and display time on sustained attention in hyperactive, normal, and control children. *Journal of Abnormal Child Psychology*, 17, 371–391.

Cohen, R. A. (1993a). Attentional control: Subcortical and frontal lobe influences. In R. A. Cohen (Ed.), *The neuropsychology of attention* (pp. 219–254). New York: Plenum Press.

Cohen, R. A. (1993b). Introduction. In R. A. Cohen (Ed.), *The neuropsychology of attention* (pp. 3–10). New York: Plenum Press.

Cohen, R. A. (1993c). Neuropsychological assessment of attention. In R. A. Cohen (Ed.), *The neuropsychology of attention* (pp. 307–328). New York: Plenum Press.

Cohen, R. A., Malloy, P. F., & Jenkins, M. A. (1999). Disorders of attention. In P. J. Snyder & P. D. Nussbaum (Eds.), *Clinical neuropsychology* (pp. 541–572). Washington, DC: American Psychological Association.

Cohen, R. A., & O'Donnell, B. F. (1993). Models and mechanisms of attention: A summary. In R. A. Cohen (Ed.), *The neuropsychology of attention* (pp. 177–188). New York: Plenum Press.

Cohen, R. J., & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurement* (4th ed.). Mountain View, CA: Mayfield.

Conners, C. K. (1992). *Conner's continuous performance test user's manual*. Toronto: Multi-Health Systems.

Conners, C. K. (1995). *Conners' continuous performance test user's manual*. Toronto: Multi-Health Systems.

Conners, C. K., & MHS Staff. (2000). *Conners' continuous performance test—II user's manual*. Toronto: Multi-Health Systems.

Corkum, P. V., Schachar, R. J., & Siegel, L. S. (1996). Performance on the continuous performance task and the impact of reward. *Journal of Attention Disorders*

### **Global or Big-Five Structure**

From the beginning, Cattell's theory of personality was hierarchical in structure (Cattell, 1946). Generally, the focus was on the primary-level traits because they provide a more finegrain definition of the individual's personality and are more powerful in predicting behavior (Ashton, 1998; Goldberg, 1972; Mershon & Gorsuch, 1988). However, Cattell found originally that when he factor analyzed the 16 primary traits themselves to find the underlying organizing influences among them, the "second-order" or global factors emerged. These five global factors, which describe personality at the most general, broad-brush level, have changed very little since Cattell's 1957 book. The same five have been scoreable from the 16PF Questionnaire since the release of the fourth edition in 1967 (Cattell, Eber, & Tatsuoka, 1970).

In fact, research leading to the development of the recent Big-Five factors of personality was based on Cattell's original scales (e.g., Costa & McCrae, 1976, 1985; Norman, 1963; Tupes & Christal, 1961). Comparisons between the five 16PF global factors and other Big-Five measures such as the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) indicate a high level of concordance (Cattell, 1996). However, the five 16PF global factor definitions have an important distinction: While other Big-Five systems arbitrarily forced orthogonal factor locations on the data because of statistical convenience, the 16PF oblique rotation methods allowed the data to determine the factors, enhancing the definitions of the factors.

### **16PF Fifth Edition**

The test has continued to be refined since 1949, resulting in four revisions, in 1956, 1962, 1968, and the Fifth Edition in 1993. This most recent edition was developed by first finding the best items from all previous 16PF forms, and then rewriting these to improve, simplify, and modernize them. Additionally, new items were written by the authors and other 16PF experts. All items were analyzed for gender, race, and disability bias. Cross-cultural translatability was also evaluated. Final items were selected through a series of factor analyses based on diverse national samples of 1,204, 646, 872, and 3,498 participants. A more detailed description of this process can be found in the *16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994).

### **PSYCHOMETRIC CHARACTERISTICS**

Because of the scientific nature of the test's origins, it shows strong psychometric characteristics supported by a long history of international research.

#### **Reliability**

The reliabilities of the 16PF Fifth Edition Questionnaire primary and global scales are comparable to other personality measures, despite the fact that 16PF scales are fairly short (10 to 15 items). These reliability coefficients are presented in Table 4.3, and further information can be found in the *16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994). Internal consistency reliability (how highly the items in a scale correlate with each other) for the 16PF primary scales averages .75 (ranging from .66 to .86 over the 16 scales)

summarized across two general population samples and one university student sample with a total of 4,660 participants. Test-retest reliabilities (or estimates of the consistency of scores over time) for a 2-week interval ranged from .69 to .87 with a median of .80. Two-month test-retest reliabilities ranged from .56 to .79 with a median of .69. The 16PF global scales have even higher reliabilities, with 2-week test-retest estimates ranging from .84 to .91 with a mean of .87, and 2-month test-retest estimates ranging from .70 to .82 with a median of .80.

### **Validity**

Because 16PF scales were developed through factor analytic methods, the results of these methods provide evidence about the construct validity of the 16 primary scales and the five global factors. A summary of these analyses is provided in the *16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994). The 16PF personality structure has been replicated in studies based on samples differing in language, culture, and education (Bolton, 1977; Boyle, 1989; Cattell, 1946, 1973; Cattell, et al., 1970; Cattell & Krug, 1986; Chernyshenko, Stark, & Chan, 2001; Gerbing & Tuley, 1991; Gorsuch & Cattell, 1967; Hofer, Horn, & Eber, 1997; Krug & Johns, 1986; Mershon & Gorsuch, 1988; Mogenet & Rolland, 1995; Motegi, 1982; Schneewind & Graf, 1998).

Other studies have provided evidence of a high degree of correspondence between the five 16PF global factors and other Big-Five measures (Cattell, 1996). Not only do the NEO PI-R and 16PF five-factor models line up well in factor analyses, but the five 16PF global scales were found to correlate as highly with the NEO PI-R five factors as the two main Big-Five models (NEO-PI-R and Goldberg's Big-Five) correlate with each other.

Many studies provide evidence for the meaning of the 16PF scales by comparing them to similar scales in various other measures of normal personality. Conn and Rieke (1994), for example, provide correlations with all scales from the California Psychological Inventory, the NEO-PI-R, the Personality Research Form, and the Myers-Briggs Type Indicator.

The comprehensiveness of the test has been supported by findings that all dimensions on other major personality tests overlap substantially with the 16PF scales in regression and factor analytic studies (Conn & Rieke, 1994; Cattell, 1996). A substantial body of criterion validity data has accumulated demonstrating the utility of the 16PF scales in predicting a wide range of behavior. For example, these criteria include self-esteem, leadership style, self-discipline, interpersonal needs, empathy, marital compatibility, coping patterns, decision-making style, frustration tolerance, cognitive processing style, compatible career choices, and job performance in a wide range of occupations (e.g., Cattell, et al., 1970; Cattell & Krug, 1986; Cattell, et al., in press; Conn & Rieke, 1994; Guastello & Rieke, 1993b; Krug & Johns, 1990).

In addition, the 16PF Questionnaire has been found to be more powerful than other popular personality inventories in predicting real-life behavior. In a recent study, Goldberg (in

press) compared the ability of several popular personality inventories to predict six clusters of behavioral criteria and found that the 16PF International Personality Item Pool

**The Sixteen Personality Factor (16PF) Questionnaire**  
**TABLE 4.3 Reliability Estimates for 16PF Fifth Edition Scales**

Internal Consistency Test-Retest

Sample

Primary Scales

1

N\_820

2

N\_2,500

3

N\_1,340

Average

N\_4,660

2-week

N\_204

2-month

N\_159

Warmth (A) .69 .69 .74 .70 (0.82) .83 .77

Reasoning (B) .76 .77 .68 .74 (0.75) .69 .65

Emotional Stability (C) .78 .78 .77 .78 (0.91) .75 .67

Dominance (E) .71 .66 .70 .68 (0.81) .77 .69

Liveliness (F) .73 .72 .70 .72 (0.68) .82 .69

Rule-Consciousness (G) .74 .75 .77 .75 (0.97) .80 .76

Social Boldness (H) .86 .85 .87 .86 (0.74) .87 .79

Sensitivity (I) .79 .77 .79 .78 (0.98) .82 .76

Vigilance (L) .74 .74 .70 .73 (1.02) .76 .56

Abstractedness (M) .74 .74 .75 .74 (1.01) .84 .67

Privateness (N) .77 .75 .78 .76 (0.87) .77 .70

Apprehension (O) .78 .78 .79 .78 (0.94) .79 .64

Openness to Change (Q1) .71 .64 .68 .66 (1.11) .83 .70

Self-Reliance (Q2) .78 .78 .78 .78 (0.89) .86 .69

Perfectionism (Q3) .73 .71 .76 .73 (0.96) .80 .77

Tension (Q4) .75 .76 .74 .75 (0.93) .78 .68

Global Scales<sup>b</sup>

Extraversion .91 .80

Anxiety .84 .70

Tough-Mindedness .87 .82

Independence .84 .81

Self-Control .87 .79

<sup>a</sup>Average internal consistency values were weighted with respect to sample size. Standard error of measurement estimates,

using weighted standard deviations, are presented in parentheses.

<sup>b</sup>Internal consistency values are not available for the global factor scales because their scores are derived from combinations

of the 16 primary factor scores.

Adapted with permission from Conn, S. R., & Rieke, M. L. (1994). *The 16PF Fifth Edition technical manual*. Champaign,

IL: Institute for Personality and Ability Testing, Inc.

(IPIP) version (Goldberg, 1999) had the highest predictive validity coefficients.

**RANGE OF APPLICABILITY AND LIMITATIONS**

The 16PF Questionnaire is a comprehensive measure of normal adult personality, and thus it has a wide range of useful applications in varied counseling and clinical settings (including

career counseling and other personal development settings); a range of industrial/organizational settings (employee selection, development, training, coaching, outplacement, and team building); as well as school and research settings. Because of its comprehensive nature and long history of research, the 16PF Questionnaire can be used to understand a wide range of behavioral patterns from self-esteem, coping patterns, and interpersonal needs to leadership, social skills, and compatible career choices (see Validity section). The 16PF test was developed to measure normal behavior and cannot be used to measure pathological dimensions or diagnose mental disorders. For these purposes, a clinical measure should be administered in addition. Nonetheless, normal personality dimensions are important in clinical settings; for example, in quickly gathering a comprehensive picture of the individual's whole personality, facilitating empathy and rapport in the therapeutic relationship, choosing therapeutic approaches, planning developmental goals, identifying relevant adjustment issues, and helping the client develop greater self-awareness (Cattell, 1989; Karson et al., 1997).

As with all tests, 16PF results should be treated as hypotheses to be combined with other sources of collateral information (interview data, history information, other psychological measures such as those of pathological dimensions or cognitive functioning, etc.) to arrive at a prediction of behavior.

#### **CROSS-CULTURAL FACTORS**

Because of its international origins, the 16PF primary factor structure has been evaluated and confirmed across many languages and cultures; for example, France (Mogenet & Rolland, 1995), Spain (Prieto, Gouveia, & Fernandez, 1996), Germany (Schneewind & Graf, 1998), the United Kingdom (Smith, 1994), Japan (Motegi, 1982), Latin America (Krug, 1971), and Italy (Barbaranelli & Caprara, 1996). Currently, the 16PF Questionnaire has been translated into over 30 languages worldwide and can be administered via [www.16PFworld.com](http://www.16PFworld.com) in over a dozen different languages. It is used worldwide, for example, in the selection of managers (Bartram, 1992; Chakrabarti & Kundu, 1984), salespersons (Coyne Didsbury Pty. Ltd., 1998), pilots (Bartram and Baxter, 1996), and police officers (Cooper, Robertson, & Sharman, 1986). It is also used internationally for research into diverse topics such as occupational choice (Arbeo, 1994), self-actualization (Kapoor & Shankhla, 1994), student giftedness (Drabkova & Drabkova, 1998), work team roles (Dulewicz, 1995), and adaptation to renal failure treatment (Carbonell, Hernandez, & Ramos, 1992).

#### **ACCOMMODATION FOR POPULATIONS WITH DISABILITIES**

Braille, American Sign Language, and audiotape formats are available for previous editions of the 16PF Questionnaire, but have not been developed and validated for the Fifth Edition. However, the issue of low literacy has been addressed by lowering the reading grade level of the 16PF Fifth Edition to a fifth-grade level. Using the computerized administration can accommodate some disabilities. Generally, for individuals with greater visual or other physical impairments, an

objective person can read the test questions to the test taker and/or record answers. Answers can also be spoken into a tape recorder and later transcribed onto an answer sheet. Since the 16PF Questionnaire is not timed, many individual accommodations can be made that would not interfere with the test administration. Because the test measures normal personality dimensions, test users do not have to be concerned with accommodating individuals with mental disabilities.

### **LEGAL AND ETHICAL CONSIDERATIONS**

As with all tests, the validity of 16PF results depends on proper test selection, administration, scoring, and interpretation by the professional test user. Therefore, it is incumbent on all test users to read *The 16PF Fifth Edition Administrator's Manual* (Russell & Karol, 1994) and *The 16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994) and to gain proper training in the use of the 16PF test. Professional workshops are given each year by the publisher (the Institute for Personality and Ability Testing [IPAT]: (800) 225-4728; [www.ipat.com/workshop.html](http://www.ipat.com/workshop.html)) and other international publishers.

All test users should read the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and be aware of the meaning of measurement error and the limitations of test scores. The 16PF Questionnaire was developed and documented in accordance with these standards. As with all test scores, 16PF results should be combined with information from other sources in drawing conclusions.

### **COMPUTERIZATION**

Computer administration, scoring, and report generation is available on a PC using the OnSite software system (which also allows for hand entering or scanning of pencil-marked answer sheets). This system also provides a disk that can be used to administer the test at several different locations before returning to the office to score the accumulated data and generate reports. The professional can import, export, and organize test data, and the archive function allows the professional to compress and back up data to a network, hard drive, or floppy disk when space or safekeeping are a concern.

Online administration of the 16PF Questionnaire is also available via NetAssess. After the client completes the test online using a pass code, the test is scored and reports are e-mailed to the professional. Online multilanguage testing is available using [www.16PFworld.com](http://www.16PFworld.com), which can administer the test in over a dozen different languages and score the test with national norms for that language group. The narrative report can be produced in the same language as the test or in one or more of the other available languages. This provides the professional with results based on the same personality dimensions across different language groups.

### **CURRENT RESEARCH STATUS**

The extensive body of research findings, accumulated over half a century, that link 16PF traits to important real-life criteria enhances use of the 16PF Questionnaire. A conservative

estimate of 16PF research since 1974 includes upwards of 2,000 publications (Hofer & Eber, 2001). Because the test has been used in varied settings, research has generated a wide range of profiles and prediction equations for criteria such as creativity, emotional intelligence, self-esteem, marital

satisfaction, coping patterns, leadership style, interpersonal skills, cognitive processing style, career preferences, job performance, and academic achievement, as well as dozens of occupational profiles (e.g., Cattell, et al., in press; Cattell, et al., 1970; Conn & Rieke, 1994; Karol & Russell, 1995; Krug & Johns, 1990).

### **USE IN CLINICAL OR ORGANIZATIONAL PRACTICE**

The 16PF Questionnaire has a long history of use in a wide variety of applied settings. The two most general ones are described here.

#### **Clinical and Counseling Settings**

Although the 16PF Questionnaire was not constructed to measure pathology or to make a clinical diagnosis, it has been found to be valuable in many aspects of clinical and counseling practice because of its ability to provide a rich, comprehensive, objective picture of the individual's enduring personality makeup. Expert clinicians (Karson et al., 1997) emphasize the importance of getting a picture of the whole person as a context in which to place information gleaned from other tests about symptoms and pathology. These authors demonstrate how 16PF results provide an understanding of basic interpersonal, experiential, emotional, and motivational dynamics and lead to effective treatment planning, especially in time-limited settings.

The Karsons also stress the importance of 16PF scores in increasing the client's self-awareness of both strengths and weaknesses, as well as facilitating early empathy and rapport in the therapy relationship. Unlike pathological dimensions, 16PF scores can be discussed openly and easily with clients to stimulate interaction and enable an early alliance in the assessment and therapy process. The Karson Clinical Report (Karson & Karson, 1995) discusses a range of topics including self-esteem, patterns of coping, frustration tolerance, ability to form and maintain gratifying relationships, cognitive processing style, decision making, and motivational dynamics. These and other experienced clinicians (Cattell, 1989, 1995) have shown the usefulness of 16PF scores in understanding a range of clinically relevant issues such as an individual's capacity for insight and introspection, bases for self-evaluation, internalization of societal standards, quality of attachments, interpersonal needs, capacity for intimacy, and openness to change. The *16PF Cattell Comprehensive Personality Interpretation Manual* (Cattell & Cattell, 1997) demonstrates the value of 16PF results in planning a range of developmental and therapeutic goals, for instance, strategies for developing a good working alliance, power dynamics, overcoming resistance to change, facilitating termination, and choosing between modalities such as group versus individual therapy or directive versus nondirective therapies.

16PF scores have also proven useful in marital or couple's counseling. Karol and Russell (1995) summarize research in

the area of marital satisfaction and discuss how 16PF results help the counselor to better understand both individuals, how their traits might combine and interact, and which therapeutic goals and approaches might be helpful.

### **Industrial and Organizational Settings**

The 16PF Questionnaire is commonly used to help organizations in a variety of functions, including hiring, promotion, development, training, outplacement, team building, and coaching. Meta-analytic reviews of the literature (e.g., Barrick & Mount, 1991) have shown measures of normal, adult personality to be valid predictors of job performance in a wide range of occupations. Although other sources of information are also important, 16PF scores provide an objective, comprehensive, and efficient source of information for the employer.

An extensive literature has accumulated demonstrating the ability of 16PF dimensions to predict important aspects of job performance and occupational satisfaction (e.g., Cattell, et al., in press; Cattell, et al., 1970; Conn & Rieke; 1994; Krug & Johns, 1990; Schuerger & Watterson, 1998). For example, 16PF scores have been used with such diverse occupations as salesperson (Guastello & Rieke, 1993a; Tucker, 1991), manager or executive (Guastello & Rieke, 1993b; Johns, Schuerger, & Watterson, 1980), accountants (Anonsen, 1985), police and security personnel (Fabricatore, Azen, Shoentgen, & Snibbe, 1978; Scholl, 2000), programmer/systems analyst (Schuerger, Johns, & Stazyk, 1985), entrepreneur (Aldridge, 1997), and factory worker (Schuerger, Ekeberg, & Kustis, 1994).

In addition, the 16PF Select (Kelly & Mead, 1999) was developed specifically for use in personnel selection. It is a shorter form of the test that includes the scales that are most predictive of job performance across a wide variety of jobs. The 16PF Select is not just a personality measure but a process in which the professional defines the personality characteristics that are most important for effective job performance via regression weights or an optimal-range score approach. It provides concise feedback comparing an applicant's personality and behavioral strengths to desired personality dimensions for the job. The extensive history of 16PF research has also produced equations for a wide range of occupationally relevant dimensions, many of which appear in the interpretive reports listed above. These include such dimensions as different types of leadership styles, leadership effectiveness, cognitive processing style, creativity and imagination, problem-solving style, risk taking, independence and initiative taking, emotional intelligence, social skills, conscientiousness, learning style, academic achievement, potential for on-the-job learning, and Holland Occupational Themes (e.g., Cattell, et al., 1970; Conn & Rieke, 1994; Guastello & Rieke, 1993b; Schuerger & Watterson, 1998; Walter, 2000).

Experts on consulting with the 16PF Questionnaire in industrial/organizational settings (e.g., Cattell, 1989; Edelman, 2001; Lord, 1997, 1999; Walter, 2000; Watterson, Aldridge, & Seelback, 2002) provide practical advice about using 16PF results in a range of industrial/organizational consulting situations, such as executive selection, employee development

or coaching, team building, outplacement, and giving feedback from test results. They demonstrate how to use 16PF scores to answer questions such as, Will the person function effectively in jobs that require a strong technical orientation? Can he or she be counted on to finish things he or she starts? Will he or she be an effective leader? Is this the kind of person who is likely to handle high-stress situations well?

The 16PF test is widely used for career development purposes both inside and outside organizations. There are many resources for identifying the degree of fit between an individual's 16PF scores and those scores characteristic of persons in a range of occupations or occupational types (Cattell, et al., 1970; Conn & Rieke, 1994; Guastello & Rieke, 1993b; Lord, 1997, 1999; Schuerger & Watterson, 1998). For example, the Personal Career Development Profile interpretive report (Walter, 2000) compares an individual's 16PF score patterns to those of about 90 different occupational interest groups. Additionally, 16PF feedback increases self-awareness and allows the individual to understand his or her strengths and weaknesses relative to a particular career path and to plan self-development goals.

#### **FUTURE DEVELOPMENTS**

There will soon be a new tool for 16PF interpretation—a volume in Wiley's series of Essentials in Psychological Assessment called *Essentials of the 16PF Questionnaire*. This book will provide step-by-step guidance in the knowledge and practical skills necessary for test administration, scoring, and interpretation, useful to professionals or graduate students. Additionally, a Spanish-American version of the 16PF Fifth Edition Questionnaire will soon be released. A thorough translation process, involving translators from many countries (e.g., Mexico, Cuba, Puerto Rico, Nicaragua, Peru, Argentina, Colombia, the Dominican Republic) and item response theory analyses, make this a quality assessment for the growing U.S. Spanish-speaking population. Online, multilanguage testing via [www.16PFworld.com](http://www.16PFworld.com) is continually expanding the number of languages available for reliable administration, scoring, and reports for multinational employers.

A revision of the Children's Personality Questionnaire, a 16PF version for 8- to 12-year-olds, is currently under development.

An in-depth 16PF leadership development report for use with executives, managers, and supervisors has just been released (Watterson, Aldridge, & Seelback, 2002).

#### **REFERENCES**

- Aldridge, J.H. (1997). *An occupational personality profile of the male entrepreneur as assessed by the 16PF Questionnaire*. Unpublished doctoral thesis, Cleveland State University, Cleveland, OH.
- Allport, G.W., & Odbert, H.S. (1936). Traitnames. A psycho-lexical study. *Psychological Monographs*, 47(211), 171.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anonsen, M.K. (1985). *Personality and interests of artists and accountants*. Unpublished master's thesis, Cleveland State University,

Cleveland, OH.

Arbeo, B.J.G. (1994). Analisis correlacional entre variables de personalidad y la indecision vocacional compleja. *Informacion Psicologica*, 55, 11–17.

Ashton, M.C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior*, 19, 289–303.

Barbaranelli, C., & Caprara, G.B. (1996). How many dimensions to describe personality: A comparison of Cattell, Comrey, and the Big Five taxonomies of personality traits. *European Review of Applied Psychology*, 46, 15–24.

Barrick, M.R., & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26.

Bartram, D. (1992). The personality of UK managers: 16PF norms for short-listed applicants. *Journal of Occupational & Organizational Psychology*, 65, 159–172.

Bartram, D., & Baxter, P. (1996). Validation of the Cathay Pacific Airways pilot selection. *International Journal of Aviation*, 6, 149–169.

Bolton, B. (1977). Evidence for the 16PF primary and secondary factors. *Multivariate Experimental Clinical Research*, 3, 1–15

Boyle, B.J. (1989). Re-examination of the major personality type

### **16PF® Fifth Edition Cattell Comprehensive Personality Interpretation (CCPI)**

by Heather Birkett Cattell, Ph.D.

with Heather E. P. Cattell, Ph.D.

Name: Ima Leader

Date: May 07, 2012

NetAssess UID: 90e9061010

This confidential report is designed for use by appropriately qualified professionals. The presentation of information is compact and the language of the report is technical. It was not intended to be used for client feedback.

This report is based on the 16PF Questionnaire, which is a measure of normal personality. The report's narrative interprets the 16PF scales in light of issues relevant for counselors and clinicians. However, the report should not be used to diagnose pathology, which requires pathology-oriented measures and/or psychodiagnostic interview.

**The report is intended to be used in conjunction with professional judgment. The statements it contains should be viewed as hypotheses to be validated against other sources of data. Personality test data should be regarded with caution when the data are over one year old or after the occurrence of a major life event. In these instances, it is recommended that the client be retested. All information in this report should be treated confidentially and responsibly.**

For additional information about the report and its contents, please refer to the "16PF Cattell Comprehensive Personality Interpretation Manual," available through IPAT.

Copyright © 1997, 1999 by the Institute for Personality and Ability Testing, Inc., P.O. Box 1188, Champaign, IL 61824-1188. All rights reserved. 16PF is a registered trademark of IPAT.

Cattell Comprehensive Personality Interpretation Ima Leader

May 07, 2012

### **VALIDITY INDICES**

This profile's Impression Management index is within the average range. Thus, Ms. Leader's responses don't appear to either greatly exaggerate or deny socially desirable attributes.

## GLOBAL PERSPECTIVE

Ms. Leader's overall level of Extraversion is average.

Ms. Leader's overall level of Anxiety appears to be high. In this area, she tends to be emotionally reactive, worried, and tense. She would be even higher on Anxiety if she were not also trusting.

Ms. Leader's general level of Tough-Mindedness is low-average. This is a result of her being emotionally sensitive and open to new ideas. She would be even lower on Tough-Mindedness if she were not also pragmatic.

Ms. Leader's general level of Independence is low-average. This is a result of her being unassertive, timid, and trusting. She would be even lower on Independence if she were not also innovative.

Ms. Leader's general level of Self-Control is average.

## COGNITIVE AND PERCEPTUAL FUNCTIONING

**Capacity for Insight:** Ms. Leader's high score on the abstract reasoning ability scale suggests that she has the intellectual capacity necessary for insight into her behavior. However, she may not often use this capacity to gain insight into herself since her attention tends to be more external and concrete in focus rather than internal or abstract. It may be hard for her to step back from her focus on practicalities to see the overall patterns in her behavior. Additionally, since she may be feeling insecure at present, she may be taking an exaggerated view of her shortcomings, which may affect her capacity for accurate self-assessment.

**Standards for Self-Evaluation:** Some of the standards against which Ms. Leader judges herself are moral ones. She has a strong conscience that sets conventional rules for her conduct. About as much as the average person, she has personal ideals that are oriented around culturally admired traits such as self-discipline, organization and goal-directedness. Thus, overall, this second source of self-evaluation may be less potent in her judgments of herself than her standards of right and wrong. Her moral standards may be high, and thus they may be a source of her current insecurity. For example, she may be unforgiving of herself for some real or imagined wrongdoing or be disappointed over a personal failure or about not achieving her goals.

**Perceptual Style:** Since Ms. Leader may tend to overestimate possible risks in situations, she may exaggerate dangers. She is also attentive to subtle social cues signaling negative shifts in people's reactions toward her, and thus may read in more adverse reactions than actually exist.

**Information Processing:** Ms. Leader tends to focus on the concrete, practical aspects of her surroundings. She is usually content to stay with "common sense" explanations of what she

2

Cattell Comprehensive Personality Interpretation Ima Leader

May 07, 2012

observes. Contemplating the underlying or theoretical meaning behind things does not usually interest her unless it leads to some practical information.

Being emotionally sensitive, Ms. Leader's judgments tend to be based more on her personal feelings and intuitions. At times, her judgments may be subjective or lacking in some factual or objective elements.

**Core Values:** Ms. Leader tends to think in conventional moral terms about how things ought to be. She has difficulty accepting anything that she sees as improper. Furthermore, many of her values are based on her emotional sensitivity and her finely tuned aesthetic sensibility. She is sympathetic and refined, and anything crude or inhumane may seem distasteful to her.

**Openness to Influence and Change:** On some matters, Ms. Leader may be easily influenced by other people's behavior and opinions. However, Ms. Leader doesn't always rely on established methods or approaches. Therefore, she may be open to being influenced by avant-garde or progressive ideas, when those around her hold these views. This trait makes her receptive to new ideas and perspectives, and good at coping with change.

## INTERPERSONAL STYLE

**Social Orientation:** Ms. Leader derives an average amount of gratification from interacting with people as an end in itself. She is also sensitive, sympathetic, and able to empathize with others' feelings. Her assessments of people are mainly based on her emotional reactions, which are reasonably warm and kind. Her general assumption when dealing with people is that they can be trusted and are well intentioned.

Ms. Leader is generally timid and shy, and she may be self-conscious about meeting new people

or being singled out for attention in a group. She tends to feel intimidated in challenging social situations. She is discreet about sharing personal information. She tends to keep her reactions to herself and to carefully consider how much to say before speaking. If she has well-developed social skills, she may be able to hide her shyness in many situations. Thus, overall, others may find her somewhat elusive or hard to really get to know.

**Quality of Attachments:** Being moderately people-oriented, Ms. Leader probably has formed some relatively strong social attachments. Ms. Leader takes her social obligations seriously and strives to do what is correct in her relationships. She may not recognize that the average person is not as conscientious as she. If she does recognize this, it doesn't seem to have made her distrusting or cynical about human nature.

Since she is the kind of person who is capable of making large shifts in her life, she may have lost contact with some people who were once significant to her. For example, she may have moved several times and not kept in touch with old friends.

**Power Dynamics:** Ms. Leader is below average on assertiveness and tries to avoid conflict. Since Ms. Leader sometimes relies on other people's ideas and support, she may be easily swayed by others.

**Compatible and Incompatible Personalities:** Overall, Ms. Leader probably sees eye-to-eye with other open-minded individuals like herself, who are in favor of finding better ways of doing

3

Cattell Comprehensive Personality Interpretation Ima Leader

May 07, 2012

things rather than just sticking to traditional ways. She probably also enjoys the company of other bright people like herself. Since Ms. Leader is rule-conscious and tender-minded, she has a lot in common with other principled and emotionally sensitive types. She does not relate well to people who are harsh and insensitive to her feelings or to those whom she views as being immoral and not following established rules. Since her focus is concrete and down-to-earth, she has much in common with other pragmatic individuals. She usually does not try to fathom abstract or theoretical ideas, and therefore may not understand imaginative, absentminded people who "have their heads in the clouds." Her negative reactions are likely to be triggered by people who are unavailable or unsupportive when she needs someone to rely on. Because she is shy and easily embarrassed, she may resent others who place her in the social spotlight or in other demanding social positions that make her feel uncomfortable.

**Impact on Others:** Since Ms. Leader is shy and timid with new people and situations, she initially may give the impression that she is unfriendly unless she pushes herself to overcome her shyness. As others get to know her and she becomes comfortable with them, most will discover that she is somewhat warmer and more talkative than they had first supposed.

Ms. Leader tends to be discreet about revealing personal information.

**Interpersonal Stress:** Ms. Leader tends to be compliant, and anyone who is negatively affected by her lack of assertiveness may wish that she would take firmer stands. Although she prefers to avoid unpleasant conflicts and confrontations, she may be drawn into them sometimes by her inability to control her emotions. Ms. Leader may be too trusting or accepting to be an astute judge of character, and thus she may be deceived or manipulated, at times. Generally, though, most people with whom she interacts will appreciate her trusting, cooperative nature and her tendency to look for the best in people.

Overall, Ms. Leader may be emotionally reactive and become easily frustrated or upset.

Therefore, her relationships may be strained, and she probably experiences relationships as stressful at times.

### **INTIMATE RELATIONSHIPS**

**Capacity for Intimacy:** Given Ms. Leader's trusting and emotionally sensitive nature, she has some of the basic traits necessary for developing emotional intimacy--that is, a relationship in which she and her partner are attuned to each other's feelings.

**Balance of Power:** If Ms. Leader behaves toward her partner as she does toward people generally, she is likely to assume an accommodating role in their relationship.

**Coping With Problems:** Ms. Leader's tendency to blame herself at times may cause her to accept more than her fair share of responsibility when things go wrong. This self-doubt may be exacerbated if it is exploited by a partner who is manipulative or tends to deny blame. Ms.

Leader's rule-consciousness adds strength to her relationships by making her take commitments seriously, but it may also put stress on a relationship with a partner who does not share her values.

These reactions may be intensified by her tendency to overreact to the natural ups and downs and frustrations that any couple experiences. Thus, her temperamental nature may put strain on her

4

Cattell Comprehensive Personality Interpretation Ima Leader

May 07, 2012

relationships.

**Compatibility Issues:** Opposites attract, but the differences that initially fascinated couples or endeared them to each other may later become reframed as sources of disappointment and disillusionment. If Ms. Leader is complaining about aspects of her partner's personality, she likely viewed these same traits in a positive light earlier in their relationship. Therefore, if Ms. Leader feels that her partner is insensitive or emotionally unavailable, chances are that she originally experienced this same trait as emotional strength and self-reliance. Similarly, if she now complains that her partner is domineering, it is likely that she once experienced this same quality as decisiveness. In addition, a partner who was valued in the first stages of the romance as traditional, as imaginative, and as self-sufficient may now be regarded negatively as too conservative, as impractical or absentminded, and as aloof or withholding.

In the long run, Ms. Leader may get along best with a partner who is emotionally sensitive like she is. In addition, since she prefers doing things together, her partner should be willing to provide a good deal of companionship and not have a strong need for personal space. The couple's compatibility would be further heightened if they shared the same sharp intellectual skills and held the same conventional morals.

#### **OCCUPATIONAL CONSIDERATIONS**

**Overall Occupational Compatibilities:** Ms. Leader is most likely to be attracted to pragmatic occupations that require using her emotional or aesthetic sensitivity and her common sense. Fields that, in the past, were considered traditionally feminine may appeal to her. She might find satisfaction in the applied arts, such as interior design, or, alternatively, in a helping profession. Ms. Leader may not be well-suited to jobs that require looking deeply into the abstract meaning behind the facts. She would probably find this kind of work too abstract and theoretical. She tends to prefer applied work in which she can be useful in a concrete way.

**Aptitudes:** Ms. Leader should be able to tackle even the most intellectually challenging problems that arise in her work. Furthermore, she may have a talent for innovative work that requires thinking outside the usual cultural paradigms. However, she may not be as imaginative as might be expected for someone who is open to new ideas. Because she tends to focus on immediate, pragmatic concerns, she may not step back and reflect on the "big picture" and imagine the possibilities. Because a good memory, focus on immediate surroundings, and alertness to practical matters are among her positive attributes, her intelligence would be best used in work with an applied, down-to-earth focus.

**Working Alone or with Others:** Ms. Leader's preference for being around other people may make her unsuited for working on her own. In such a situation, she would probably feel alone and unsupported. Being part of a close-knit team would suit her best.

**Assertiveness and Boldness:** Ms. Leader is unassertive as well as being shy. Thus, she may have some difficulty expressing her point of view or standing up for herself.

**Attitude Toward Authority:** Generally, Ms. Leader may be sensitive to any practice by authorities that she perceives as being unfair or oppressive.

5

Cattell Comprehensive Personality Interpretation Ima Leader

May 07, 2012

She also tends to be discreet and deferential in her dealings with others, and thus, she may keep her grievances to herself. She may sometimes express her grievances indirectly (at worst, she could possibly engage in passive resistance or backbiting). At times, she may decide to risk giving some constructive criticism, but because she fears confrontation, her comments may be overly tactful and too low key to have any real impact.

**Organizational Fit:** Ms. Leader's innovative temperament may not be a good match with a

large, authoritarian institution or one that is set in its ways. She would fit better in a newly emerging industry or an organization where the lines of authority are not highly structured and policies and procedures are flexible. She also is well suited to an environment where teamwork and cooperation are valued rather than independence and individual enterprise.

**Autonomy and Initiative:** Ms. Leader may not have a strong enough imagination to envision theoretical solutions to abstract problems. She also may not be particularly self-reliant or independent. Thus, she is less likely to demonstrate special initiative and enterprise in the workplace or in branching out on her own.

**Self-Discipline and Achievement Orientation:** Ms. Leader appears to have a well-developed work ethic that tends to make her rule-conscious and concerned about doing what is conventionally right. She is likely to feel guilty if she feels she has produced shoddy work or if she receives a poor evaluation.

**Potential Motivators:** Being in a position of authority that requires giving orders, dealing with conflict, and taking unpopular stands may be aversive to Ms. Leader, since she is unassertive. Being in aesthetically pleasing surroundings would enhance her job satisfaction considerably.

#### **PERSONALITY DYNAMICS**

**Tension and Coping:** At the time she took the test, Ms. Leader indicated feeling wound up or physically tense. Her tension may have been just a transitory reaction to current life events, perhaps even the testing situation itself. On the other hand, it is possible that she is a chronically tense, driven person, who finds it difficult to relax.

At the time she took the test, Ms. Leader also gave indications of being easily upset. Unless she was just reacting to current stressful events, it may be that she is unable to regulate her negative affect enough to enable her to cope with difficult life events and frustrations in a balanced, adaptive way. Her high level of tension may also be limiting her ability to cope.

**Social Insecurity and Avoidance:** Ms. Leader's low score on some of the sociability factors suggests that she may be shy and rejection-sensitive. Thus, she may tend to stay in a comfort zone of safe, predictable social interactions, causing her to live a somewhat lonely existence. She probably wants more interpersonal contact than she has but she may be too shy or awkward to initiate it. At times, she may see life as passing her by and become envious of more outgoing, adventurous types who appear to live fuller lives than she.

#### **THERAPEUTIC AND COUNSELING ISSUES**

**Orientation to Feelings:** Since Ms. Leader is an emotional person who often feels (rather than

6

Cattell Comprehensive Personality Interpretation Ima Leader

May 07, 2012

thinks) her way through problems, therapy efforts may involve bringing some balance to her on-sided orientation. However, she may not look at things from an objective viewpoint until she has had the opportunity to express her feelings and is reasonably sure that they have been heard and valued. Thus, her progress will be influenced by the extent to which she feels understood at an emotional level. Her progress will also be influenced by the extent to which she believes the therapist cares about her. She may give more weight to this consideration than to the therapist's expertise or credentials.

**Group Versus Individual Modality:** Since Ms. Leader is a somewhat cautious person who may feel intimidated or self-conscious when speaking in front of others, group therapy would not be the modality of choice for her. Relating to just one person at a time is more comfortable for her, and she probably will progress better in an individualized arrangement if the goal is for her to feel at ease in airing her problems. Later, however, a group modality may be effective in helping to desensitize her social fears.

Cattell Comprehensive Personality Interpretation Ima Leader

Score Summary May 07, 2012

16PF Score Summary

Validity Indices

Raw

Score Percentiles

Impression Management 10 35% within expected range

Infrequency 0 55% within expected range

Acquiescence 62.73% within expected range

Global Scores

Sten Left Meaning 1 2 3 4 5 6 7 8 9 10 Right Meaning

8

4.5 Introversion Extraversion

7.6 Low Anxiety Anxiety

4.1 Receptivity Tough-Mindedness

3.7 Accommodation Independence

6.4 Unrestrained Self-Control

average

Cattell Comprehensive Personality Interpretation Ima Leader

Score Summary May 07, 2012

16PF Primary Scores

Factor Sten Left Meaning 1 2 3 4 5 6 7 8 9 10 Right Meaning

9

**A 5 Reserved, Warm,**

aloof, detached friendly, attentive to others

**B 9 Concrete, Abstract,**

less reasoning ability more reasoning ability

**C 3 Reactive, Emotionally Stable,**

less ego strength more ego strength

**E 4 Deferential, Dominant,**

submissive, humble assertive, competitive

**F 5 Serious, Lively,**

inhibited, somber energetic, carefree

**G 7 Expedient, Rule-Conscious,**

unconventional conventional

**H 2 Shy, Socially Bold,**

socially timid venturesome, seeks attention

**I 8 Utilitarian, Sensitive,**

tough, unsentimental refined, sentimental

**L 4 Trusting, Vigilant,**

accepting, easy-going suspicious, skeptical

**M 4 Grounded, Abstracted,**

practical, pragmatic idea-oriented, imaginative

**N 7 Forthright, Private,**

naive, self-disclosing discreet, shrewd

**O 7 Self-Assured, Apprehensive,**

secure, untroubled guilt-prone, worrying

**Q1 7 Traditional, Open-to-Change,**

resists change experimenting

**Q2 4 Group-Oriented, Self-Reliant,**

socially group-dependent solitary, individualistic

**Q3 5 Tolerates Disorder, Perfectionistic,**

careless orderly, compulsive

**Q4 8 Relaxed, Tense,**

placid, patient driven, fast-paced

average

Cattell Comprehensive Personality Interpretation Ima Leader

Item Summary May 07, 2012

**This page of 16PF scores is intended for qualified professionals only.**

**Data on this page should be treated with utmost confidentiality.**

**Item Responses**

31. c 63. a 95. c 127. a 159. c

32. a 64. a 96. a 128. c 160. c

1. a 33. a 65. a 97. a 129. c 161. c

- 2. a 34. a 66. c 98. c 130. a 162. c
- 3. a 35. c 67. a 99. a 131. a 163. a
- 4. c 36. a 68. a 100. a 132. a 164. c
- 5. a 37. a 69. a 101. a 133. c 165. c
- 6. a 38. a 70. a 102. a 134. c 166. c
- 7. a 39. a 71. a 103. c 135. c 167. a
- 8. c 40. a 72. c 104. a 136. a 168. a
- 9. c 41. a 73. c 105. a 137. c 169. a
- 10. a 42. a 74. a 106. c 138. c 170. c
- 11. c 43. a 75. a 107. a 139. a
- 12. c 44. a 76. a 108. a 140. c 171. a
- 13. c 45. a 77. a 109. a 141. a 172. c
- 14. c 46. a 78. a 110. a 142. c 173. a
- 15. c 47. a 79. a 111. c 143. a 174. a
- 16. a 48. c 80. a 112. c 144. c 175. c
- 17. a 49. a 81. a 113. a 145. c 176. a
- 18. c 50. a 82. c 114. a 146. c 177. b
- 19. a 51. c 83. c 115. a 147. c 178. c
- 20. c 52. c 84. a 116. a 148. a 179. c
- 21. c 53. a 85. a 117. a 149. a 180. c
- 22. a 54. a 86. c 118. c 150. a 181. b
- 23. a 55. a 87. a 119. c 151. c 182. b
- 24. c 56. a 88. a 120. a 152. a 183. b
- 25. a 57. a 89. c 121. c 153. c 184. b
- 26. a 58. a 90. a 122. a 154. c 185. b
- 27. c 59. c 91. c 123. a 155. a
- 28. a 60. c 92. a 124. c 156. a
- 29. c 61. a 93. a 125. a 157. a
- 30. a 62. a 94. c 126. a 158. a

**Summary Statistics:**

# a-responses = 106 out of 170 (62%)

# b-responses = 0 out of 170 (0%)

# c-responses = 64 out of 170 (38%)

# missing responses = 0 out of 185 (0%)

Factor A B C E F G H I L M N O Q1 Q2 Q3 Q4 IM IN AC  
 Raw Scores 14 15 8 10 10 20 0 20 6 2 16 16 22 2 12 18 10 0 62  
 Missing Items 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

10

Cattell Comprehensive Personality Interpretation Ima Leader

Item Summary May 07, 2012

This report was processed using 16PF Fifth Edition Questionnaire combined-sex norms. RGV: 5.4

## CHAPTER 8

### **The Eysenck Personality Scales: The Eysenck Personality Questionnaire-Revised (EPQ-R) and the Eysenck Personality Profiler (EPP)**

JEREMY MILES AND SUSANNE HEMPEL

TEST DESCRIPTION 99

**The EPQ-R 99**

**The EPP 100**

THEORETICAL BASIS 100

TEST DEVELOPMENT 101

PSYCHOMETRIC CHARACTERISTICS 102

RANGE OF APPLICABILITY AND LIMITATIONS 103

CROSS-CULTURAL FACTORS 103

ACCOMMODATION FOR POPULATIONS WITH  
DISABILITIES 103

LEGAL AND ETHICAL CONSIDERATIONS 103

COMPUTERIZATION 103

CURRENT RESEARCH STATUS 104

USE IN CLINICAL OR ORGANIZATIONAL PRACTICE 104

**Clinical Practice 104**

**Organizational Practice 104**

FUTURE DEVELOPMENTS 105

APPENDIX: SOURCES OF SCALES AND MANUALS 105

REFERENCES 105

#### **TEST DESCRIPTION**

The Eysenck Personality Questionnaire-Revised (EPQ-R) and the Eysenck Personality Profiler (EPP) are self-assessment personality tests for administration to normal populations. Both these tests assess the three primary Eysenckian traits of extraversion (E), neuroticism (N), and psychoticism (P). In addition, both tests contain a lie scale (L), designed to assess the degree of socially desirable responding. Both tests are available in a short form suitable for use when time (or inclination of respondents) is limited.

Eysenck and Eysenck (1991) describe the high E scorer as “sociable, likes parties, has many friends, needs to have people to talk to, and does not like reading or studying by himself. He craves excitement, takes chances, often sticks his neck out, acts on the spur of the moment, and is generally an impulsive individual” (p. 3).

The high N scorer is described as being “an anxious, worrying individual, moody and frequently depressed. He is likely to sleep badly, and to suffer from various psychosomatic disorders. He is overly emotional, reacting too strongly to all sorts of stimuli, and finds it difficult to get back onto an even keel after each emotionally arousing experience” (p. 3).

A high P scorer will be tough minded and nonconformist, likely to be aggressive, cold, and impersonal. He or she may also be prone to Machiavellianism and antisocial behavior.

#### **The EPQ-R**

The EPQ-R (Eysenck & Eysenck, 1991) contains 100 items to measure the three personality dimensions of extraversion, neuroticism, and psychoticism, as well as the lie scale. A

dichotomous response format is used with respondents ticking “Yes” or “No.”

Examples of items are:

*Extraversion*

“Do you like telling jokes and funny stories to your friends?” (positively scored item).

“Do you prefer reading to meeting people?” (negatively scored item).

**100 The Eysenck Personality Scales**

**TABLE 8.1 Example Items from the EPP**

Major Scale Subscale Item

Extraversion Sociability Do you spontaneously introduce yourself to strangers at social gatherings? (Positive item)

Activity Do you often feel tired and listless? (Negative item)

Assertiveness Do you prefer to stay in the background rather than push yourself forward? (Negative item)

Neuroticism Anxiety Are you inclined to tremble and perspire when faced with a difficult task ahead? (Positive item)

Inferiority Do you sometimes withhold your opinions for fear that people will laugh and criticize you? (Positive item)

Unhappiness Do you feel a sense of inner calm and contentment most of the time? (Negative item)

Psychoticism Risk-Taking Do people who drive carefully annoy you? (Positive item)

Impulsivity Do you get so excited and involved with new ideas that you never think of possible snags? (Positive item)

Irresponsibility Are you normally on time for appointments? (Negative item)

Lie Have you ever been late for an appointment or work?

*Neuroticism*

“Are you a worrier?” (positively scored item).

*Psychoticism*

“Do you enjoy hurting people you love?” (positively scored item).

Would it upset you a lot to see a child or an animal suffer? (negatively scored item).

*Social Desirability (L)*

“Are you always willing to admit when you have made a mistake?” (positively scored item).

“Have you ever cheated at a game?” (negatively scored item).

The short form of the EPQ-R comprises 48 items, 12 for each of the subscales.

**The EPP**

Like the EPQ-R, the EPP measures the three main personality dimensions of E, N, and P and has a lie scale to ensure that socially desirable responding is controlled. However, the EPP is a departure from the previous Eysenck scales in that it uses facet scales to measure each of the main dimensions. The

three main dimensions and their facets are (with three-letter abbreviations in parentheses):

- *Extraversion*: Activity (ACT), Sociability (SOC), Assertiveness (ASS), Expressiveness (EXP), Ambition (AMB), Dogmatism (DOG), Aggressiveness (AGG).
- *Neuroticism*: Inferiority (INF), Unhappiness (UNH), Anxiety (ANX), Dependence (DEP), Hypochondria (HYP), Guilt (GUI), Obsessiveness (OBS).
- *Psychoticism*: Risk-Taking (RIS), Impulsivity (IMP), Irresponsibility (IRR), Manipulativeness (MAN), Sensation-Seeking (SEN), Tough-Mindedness (TOU), Practicality (PRA).

The response format is expanded to include a third possible response of “Can’t decide” along with “Yes” and “No.” Each scale of the EPP contains 20 items. The 22 scales (7 for each of P, E, and N, plus the lie scale) therefore contain 440 items. Some example items from the EPP are shown in Table 8.1.

The EPP (Eysenck, Barrett, Wilson, & Jackson, 1992) short form reduces the number of facets for each of the main dimensions, rather than reducing the number of items. The facets of the short form are:

- *Extraversion*: Activity, Sociability, Assertiveness.
- *Neuroticism*: Inferiority, Unhappiness, Anxiety.
- *Psychoticism*: Risk-taking, Impulsivity, Irresponsibility.

The short form of the EPP contains a total of 200 items.

### **THEORETICAL BASIS**

The beginnings of the dimensions of extraversion and neuroticism can be traced to Hippocrates in ancient Greece, later popularized by Galen (a Roman physician). In these early theories, persons were divided into one of four types—sanguine, melancholic, choleric, or bilious. These descriptions were also used by Kant in his book *Anthropologie* (1798). These descriptions of types were adapted by Wundt (1903) who used the terms *strong* and *weak* emotions to describe what we would know as the neuroticism dimension, and *changeable* and *unchangeable* to represent what we know as the extraversion dimension (see Figure 8.1).

Eysenck always adopted a top-down, theoretically driven approach in the development of his theory and his scale. This **Figure 8.1** How the Eysenckian dimensions map onto previous theories

of personality: The inner circle represents the four humors of Hippocrates and Galen, the second circle the descriptions of persons from Kant, and the outside the dimensions of Eysenck.

approach contrasted with the bottom-up or data driven approaches employed by, for example, Cattell (see Chapter 4 of this volume). Eysenck was always a proponent of the idea that the fundamental dimensions of personality should be associated with anatomical or biological differences between individuals.

Eysenck considered the biological basis for extraversion to be related to the arousability of the central nervous system (Eysenck, 1967). A greater degree of arousability is associated with a greater degree of introversion. Although this relationship at first seems counterintuitive, it is explained by the fact that people seek an optimal level of stimulation. A

person who has a lower degree of arousability will require greater levels of external stimulation to be aroused to the optimum level—thus a low degree of arousability is associated with extravert behavior such as seeking social stimulation. A person with a high degree of arousability will find this stimulation to be overarousing and tending to increase arousal above the optimum level, and he or she will be sufficiently stimulated by quieter activities. The hypothesis that extraversion, as measured by the Eysenckian personality scales, is associated with nervous system changes has been tested in a range of different ways.

The most well-known demonstration of the link between extraversion and physiology is probably the experiments carried out using lemon juice on the tongue (e.g. Eysenck & Eysenck, 1967). When lemon juice is placed on the tongue of participants, introverts are found to salivate more than extraverts—it is theorized that this is because of the greater arousability of the introvert.

Two of the most compelling findings are direct studies of the electrical activity in the brain using electroencephalography (EEG) and vigilance performance. Gale (1981) reviewed studies that examined the relationship between evoked potentials in EEG measures and extraversion. He found some evidence for a relationship, although it was difficult to draw firm conclusions because of the methodological quality of many of the studies. The arousability hypothesis suggests that extraverts should perform less well than introverts at tasks that require sustained concentration. One such task is a vigilance task, where participants must look for rarely occurring signals and respond appropriately to them. Koelega (1992) carried out a meta-analytic review of studies that examined vigilance and extraversion-introversion and found evidence that introverts did indeed perform better than extraverts on a vigilance task, although several variables were also found to moderate the relationship between extraversion and vigilance performance.

Neuroticism is considered to have its basis in autonomic nervous system lability. The autonomic nervous system of people who are high in neuroticism is more active, and therefore responds more strongly to stressful and anxiety-inducing events. Although the hypothesis that there is a relationship between the autonomic nervous system and neuroticism would make intuitive sense, those studies that have tried to find such a link have generally been unsuccessful.

Psychoticism was the last trait to be added to the Eysenckian personality system and has the weakest theory linking it to any physiological underpinnings. Psychoticism is closely related to such traits as sensation seeking and impulsivity (e.g. Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1994). The nature of any theoretical relationship between psychoticism and physiology has been explored to a much lesser extent by Eysenck and his coworkers than either extraversion or neuroticism.

The theoretical foundations for the Eysenck personality scales were laid down over 30 years ago, and they have stood the test of time with varying degrees of success.

Melancholic Choleric Phlegmatic Sanguine

Unstable

Stable

Introverted Extraverted

Sociable

Outgoing

Talkative

Easygoing

Responsive

Lively

Carefree

Leadership

Passive

Careful

Thoughtful

Controlled

Peaceful

Reliable

Even-

Calm

Tempered

Active

Optimistic

Impulsive

Excitable

Changeable

Aggressive

Restless

Touchy

Quiet

Unsociable

Reserved

Sober

Pessimistic

Rigid

Anxious

Moody

### **TEST DEVELOPMENT**

The Eysenckian scales were developed by H.J. Eysenck and his coworkers (principally his wife, S.B.G. Eysenck) over a long period of time. The first incarnation of the scales was the Maudsley Medical Questionnaire (MMQ; Eysenck, 1952), which measured neuroticism. This scale was modified and became the Maudsley Personality Inventory (MPI; Eysenck 1959), which contained a measure of extraversion, and later

the Eysenck Personality Inventory (EPI; Eysenck & Eysenck, 1964), which added a lie scale. The next change was the addition of the Psychoticism (P) scale in the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975).

The addition of the Psychoticism scale caused some changes to be made to the Extraversion scale. Previously, it was possible to distinguish two subdimensions of the Extraversion scale—sociability and impulsivity—but some of the impulsivity-related items in the extraversion dimension were found to load on the Psychoticism scale, and thus minor modification

of the Extraversion scale took place. The subsequent scale was not found to contain any distinguishable subscales (Roger & Morris, 1990).

The EPQ was revised in 1985, when improvements were made to the Psychoticism scale to adjust some flaws in its psychometric properties—notably its high degree of positive skew and low reliability (Eysenck, Eysenck, & Barrett, 1985).

The EPP (Eysenck et al., 1992) represents the next stage of the development of the Eysenckian scales, and it, with the inclusion of facet scales, represents a move in a slightly different direction.

### **PSYCHOMETRIC CHARACTERISTICS**

Caruso, Witkiewitz, Belcourt-Dittloff, and Gottlieb (2001) conducted a reliability generalization study on data from 69 samples found in 44 studies using the EPQ or EPQ-R. Summary statistics of the reliability of the EPQ scales are shown in Table 8.2. As can be seen in the table, the reliability of the E and N scores is usually adequate, with a minimum reliability being only slightly below 0.7, which is commonly used as a cutoff. The reliability of the P scale often falls below what would normally be considered acceptable.

The reliability of the scores varied considerably between scales, with a tendency for the P scale to show the lowest reliability—both the mean and median are below the level that would be considered acceptable. The revised version of the P scale (Eysenck, S.B.G., et al., 1985) was designed, in part, to address this issue, although Caruso et al. found no consistent differences in the reliabilities of different versions of the P scale. The larger standard deviation of the reliabilities of the P scale is also matched by the relative size of the estimates of the standard deviation of the sampling distribution of the reliability of the P scale, found by Miles, Shevlin, and McGhee (1999), when they used a bootstrapping approach.

### **TABLE 8.3 Weighted Mean Reliabilities Found by Costa & McCrae (1995); Eysenck, H.J., et al. (1992); and Jackson et al. (2000)**

Major Dimension Subdimension Mean Reliability

ACT 0.76

SOC 0.81

ASS 0.75

Extraversion EXP 0.58

AMB 0.79

DOG 0.58

AGG 0.68

INF 0.84

UNH 0.85

ANX 0.83

Neuroticism DEP 0.73

HYP 0.73

GUI 0.79

OBS 0.71

RIS 0.68

IMP 0.75

IRR 0.71

Psychoticism MAN 0.63

SEN 0.75  
TOU 0.57  
PRA 0.74

Caruso et al. note that in the vast majority of papers that used the EPQ or EPQ-R, the authors failed to report the reliability of the scales. They note that, because the results of different studies vary so much in reliability, it is important to report for each sample just how reliable the results are. Without reliability statistics, it is not possible to interpret the results of a study appropriately.

Fewer studies have been carried out using the EPP and so fewer data are available regarding the reliability estimates of the different scales. Table 8.3 shows the mean reliability found in three published studies (Costa & McCrae, 1995; Eysenck et al., 1992; Jackson, Furnham, Forde, & Cotter, 2000; the reliabilities have been weighted by the relative sample size of each of the studies).

With regard to test-retest reliability, we are not aware of any independent published studies that have examined the test-retest reliability of the Eysenck scales. The manual of the

#### **TABLE 8.2 Summary Statistics for Reliability of EPQ/EPQ-R**

EPQ Scale Psychoticism Extraversion Neuroticism Lie

Min 0.36 0.68 0.69 0.59

Max 0.91 0.93 0.97 0.88

Median 0.68 0.82 0.83 0.78

Mean 0.66 0.82 0.83 0.77

SD 0.13 0.05 0.04 0.05

Caruso, J.C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J.D. *Educational and Psychological Measurement*, 61, pp. 675–689, copyright

– 2001 by Sage Publications. Reprinted by permission.

EPQ-R (Eysenck & Eysenck, 1991) gives the following values for test-retest reliabilities over 1 month with a sample of 160 mixed sex subjects: for P,  $r = 0.71$ ; E,  $r = 0.92$ ; N,  $r = 0.89$ ; L,  $r = 0.83$ .

#### **RANGE OF APPLICABILITY AND LIMITATIONS**

The EPQ-R is designed for assessing differences among adults; the existence of the Junior Eysenck Personality Questionnaire (JEPQ) allows high comparability with children's data. The EPPs are long questionnaires with 440 items in the full version and 200 in the short form. They take some time for respondents to complete. If respondents are not highly motivated and determined they may have problems completing the full questionnaire.

#### **CROSS-CULTURAL FACTORS**

The personality tests developed by Eysenck have been translated into various languages, for example, Dutch (Scholte & DeBruyn, 2001), Hebrew (Katz & Francis, 2000), Hindi (Thakur & Thakur, 1986), Persian (Shapurian & Hojat, 1985), and Japanese (Hosokawa & Ohyama, 1993).

Much work has been dedicated to checking the factorial similarity of the scales proposed by Eysenck in different cultures and languages. Probably the definitive study in this area was carried out by Eysenck, Barrett, and Eysenck (1985; although

see also Barrett, Petrides, Eysenck, & Eysenck, 1998). Eysenck et al. collected datasets from the application of the EPQ in 24 different countries—often obtaining multiple datasets from each country (this was expanded to 34 countries by Barrett et al.). The datasets were split according to gender, and the factor congruence was examined by rotating the factors to a target matrix of the UK factor-loading matrix for males and females. The factor congruence coefficients demonstrated a high degree of similarity between the matrices from different countries.

One recent study (Twenge, 2001) has investigated the changes in levels of extraversion over time in American students as measured using the EPI and the EPQ. The study examined the data from studies carried out between 1966 and 1993. It found that there was a strong and significant correlation ( $r = 0.65$ ) between the extraversion scores and the year of data collection, for both males and females. Bulheller and Ha'cker (1997) translated the EPP into German and revised the structure. They reduced the number of scales to 14 and the number of items to 176, measuring the dimensions of extraversion, emotionality, adventure/caution, and openness.

#### **ACCOMMODATION FOR POPULATIONS WITH DISABILITIES**

Reid (2000) examined the accessibility of the EPQ-R to visually impaired adults. She found that no problems arose with the use of any adaptive technology with the short form of the EPQ-R.

#### **LEGAL AND ETHICAL CONSIDERATIONS**

The EPP and the EPQ-R are only available to suitably qualified users—the publishers of the tests can provide this information (see Appendix for contact details).

When providing feedback to users, the tester may consider that terms such as “neuroticism” and “psychoticism” may have pejorative overtones. Testers may prefer to use a more neutral term, such as “proneness to anxiety” rather than neuroticism, and “tough-mindedness” rather than psychoticism.

When interpreting scores, testers should be aware that sex differences occur in the scales. Males tend to score more highly on extraversion and psychoticism, and females score more highly on neuroticism.

#### **COMPUTERIZATION**

The presentation of the EPQ-R via a computer monitor seems to be as suitable as the conventional paper-and-pencil administration.

A study using the German version of the EPQ-R compared both forms and found no systematic differences in group means and standard deviations for the scales (Merten & Ruch, 1996). The authors also found that there was good overall acceptance of the computerized version, which did not depend on previous experience with computer applications. French and Beaumont (1989) used a computerized version of the EPQ with clinical patients. As well as assessing the reliability of each of the scales using pencil-and-paper and computer presentation, they administered an alternative form of the EPQ to the same participants, allowing the authors to assess the equivalence of the two media. They found that the two methods of presentation were comparable in most of the

groups they studied. (The noncomparable group were hospital patients.)

The EPP is available in computerized form from Psi-Press—we are not aware of any published studies that have specifically set out to evaluate this mode of presentation with the EPP. **CURRENT RESEARCH STATUS**

An enduring area of research involving the Eysenckian personality scales is the debate about whether personality theorists should employ three, five, or some other number of fundamental dimensions of personality. Eysenck argued strongly that three dimensions were more appropriate than five (e.g. Eysenck, 1992). The bulk of this research on personality measures involves either analyzing the structure of one of the Eysenck questionnaires or examining the variance shared by one of the Eysenck scales and a scale that measures the Big Five.

Costa and McCrae (1995) and Jackson et al. (2000) both provide examples of research involving the analysis of an Eysenck questionnaire. In both studies, the EPP (full version) was administered to a sample of individuals, and exploratory and confirmatory factor analysis was used to examine the structure of the scales. Costa and McCrae concluded that a five-factor model provided a better interpretation of the EPP than the three-factor model. Jackson et al. criticized the study by Costa and McCrae, arguing that the sample was inappropriate. Jackson et al. replicated the analysis of Costa and McCrae and concluded that the three-factor solution was the more appropriate.

Draycott and Kline (1995) argue that factor analysis of an individual questionnaire is inappropriate—the questionnaire was probably designed using factor analysis, and thus factor analysis will result in a solution intended by the authors of the questionnaire. Instead, participants in their study completed the EPQ-R and the NEO Personality Inventory (NEOPI, a scale designed to measure the Big Five; Costa & McCrae, 1985), and they carried out a canonical correlation analysis to examine the variance shared by the two factors. Draycott and Kline concluded that the NEO-PI did contain variance that was not accounted for by the EPQ-R, but that it was not possible to determine whether this additional variance was useful or meaningful. As they conclude, “a five factor model may not be the best account of personality variance, although it is only through validation work with external criteria that this issue may be laid to rest” (p. 804).

However, the Big Five is not the only model of personality that has been compared with the Eysenckian model. Other authors have proposed alternative biological models of personality that have also been compared with the Eysenckian model. Gray (1982, 1987) has proposed a model of personality that is similar to Eysenck’s theory—it is biologically based and comprises three main dimensions: Impulsivity, Anxiety, and Fight/Flight. Gray identifies a physiological basis and rationale for each of these dimensions. Jackson (2002) used the Gray-Wilson Personality Questionnaire (designed to measure Gray’s dimensions of personality), a learning styles questionnaire, and either the EPQ-R (Study 1) or the EPP

(Study 2), and examined the relationships among them. Results showed that Gray's model of personality managed to explain very little of the variance in the learning styles questionnaire and the Eysenck scales and was not able to describe the whole domain of personality.

## **USE IN CLINICAL OR ORGANIZATIONAL PRACTICE**

In this section we examine use in clinical practice and organizational practice separately.

### **Clinical Practice**

The Eysenck scales are frequently used in studies of health behavior (e.g. Brayne, Do, Green, & Green, 1998) and studies with clinically relevant aspects, such as depressive symptoms and sleep disturbance (e.g. Fischer, 1994).

The neuroticism trait correlates with the expression of several neurotic symptoms, such as free-floating anxiety, phobic anxiety, somatic concomitants of anxiety, obsession, depression, and hysterical personality traits (Manickam, 1996), and may therefore be included in clinical studies.

The EPQ is commonly used as a measure of personality to examine adjustment to different events. For example, Alexander et al. (1996) carried out a study to examine the relative efficacy of two treatments for dysfunctional uterine bleeding. In this study, they used the EPQ as a measure of personality to predict adjustment.

The relationship between psychoticism and schizophrenic symptoms, or schizotypy, is more complex—possibly because of the multidimensional nature of schizotypy found by Bentall, Claridge, and Slade (1989). Studies that have attempted to find a correlation between schizotypal symptoms and the Psychoticism scale tend to find low correlations. For example, Rust, Moncada, and Lepage (1988) found the correlation between the Psychoticism scale of the EPQ and scores on the Inventory of Schizophrenic Cognitions to be 0.12. Chapman, Chapman, and Kwapil (1994) examined the ability of the P scale to predict psychoticism symptoms over a 10-year period. They found that “high scorers on the P-Scale seem to be psychoticlike but are not at heightened risk for psychosis” (p. 363).

### **Organizational Practice**

A number of studies have been published that compared different groups to normative samples. Wilson and Jackson (1994) examined the personality of physicists and found that physicists tended to be introverted and cautious when compared with controls. Miles and Gale (1998) found that cytology screeners scored much lower than norm samples on measures of extraversion and psychoticism (particularly risk taking) from the EPP.

As well as occupational selection, the EPQ and EPP are used to examine aspects of workplace performance. Furnham, Petrides, Jackson, and Cotter (in press) examined the relationship between the EPP and job satisfaction and found that the personality factors could explain only a small amount of variance in job satisfaction. Jackson and Corr (1998) found that extraversion, as measured by the EPP, correlated 0.17 with performance of sales staff. Jackson (1999) again found only low correlations with the EPP and performance in the

workplace but also found that the EPP measures interacted with other measures (attributional style and beliefs about money) to predict significantly more variance than the EPP measures alone.

Although the studies cited above show how the Eysenckian scales might be used to aid in occupational selection, research has also investigated how personality may affect aspects of job performance, with the object of providing appropriate interventions or counseling to solve problems.

Center and Callaway (1999) studied teachers of students with emotional or behavioral disorders. They found positive correlations between stress and both the P and N scales. In addition, teachers who had been injured by a student in the 12 months preceding data collection scored significantly higher on the P scale than those teachers who had not been injured. In a similar study, Fontana and Abouserie (1994) examined the relationship between stress and measures on the EPQ in teachers. This research also found a positive correlation between stress and both N and P, but also a negative correlation between stress and E.

#### **FUTURE DEVELOPMENTS**

With the death of H.J. Eysenck in 1997, it might be thought that the future of the Eysenck scales and model of personality is in doubt. However, work that can lead toward the development and refinement of the scales is continuing. Jackson et al. (2000) carried out a psychometric evaluation of the EPP and identified a number of issues with the scale—including the number of response categories and the low reliabilities of some of the scales—and they suggested that there may be too many scales tapping the neuroticism construct.

The EPQ has been modified by researchers in the past in a number of ways. For example, Francis (1996) developed an abbreviated form of the EPQ-R, which included six items per scale. Corulla (1990) revised the Junior P scale to correct the low reliability found in previous studies. In addition, this scale has been translated into Dutch (De Bruyn, Delsing, & Welten, 1995), although there were still problems with the reliability of the P scale. Combining these two approaches, Maltby and Talley (1998) developed and evaluated an abbreviated form of the Junior EPQ-R in a U.S. sample.

#### **APPENDIX: SOURCES OF SCALES AND MANUALS**

The EPQ-R is available from:

EdITS, PO Box 7234, San Diego, CA 92167, USA

<http://www.edits.net>

E-mail: [edits@k-online.com](mailto:edits@k-online.com)

Phone: (US) 800-416-1666

The EPP is available from:

Psi-Press, 1 Hailsham Road, Worthing, West Sussex BN11

5PA, UK

<http://www.psi-press.co.uk>

E-mail: [admin@psi-press.co.uk](mailto:admin@psi-press.co.uk)

tel: \_44 (0)1903 500 736; fax: \_44 (0)1903 500 737.

#### **REFERENCES**

Alexander, D.A., Naji, A.A., Pinion, S.B., Mollison, J., Kitchener, H.C., Parkin, D.E., et al. (1996). Randomised trial comparing hysterectomy with endometrial ablation for dysfunctional uterine

bleeding: Psychiatric and psychosocial aspects. *British Medical Journal*, 312, 280–284.

Barrett, P.T., Petrides, K.V., Eysenck, S.B.G., & Eysenck, H.J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences*, 25, 805–819.

Bentall, R.P., Claridge, G.S., & Slade, P.D. (1989). The multidimensional nature of schizotypal traits: A factor analytic study with normal subjects. *British Journal of Clinical Psychology*, 28, 363–375.

Brayne, C., Do, K.A., Green, L., & Green, A.C. (1998). Is health protective behaviour in adolescents related to personality? A study of sun protective behaviour and the Eysenck Personality Questionnaire (junior version) in Queensland. *Personality and Individual Differences*, 25, 889–895.

Bulheller, S., & Hacker, H. (1997). *Eysenck Personality Profiler (EPP-D)—Manual*. Frankfurt/Main: Swets Test Services.

Caruso, J.C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J.D. (2001). Reliability of scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement*, 61, 675–689.

Center, D.B., & Callaway, J.M. (1999). Self-reported job stress and personality in teachers of students with emotional or behavioral disorders. *Behavioral Disorders*, 25, 41–51.

Chapman, J.P., Chapman, L.J., and Kwapil, T.R. (1994). Does the Eysenck Psychoticism scale predict psychosis? A ten year longitudinal study. *Personality and Individual Differences*, 17, 369–375.

Corulla, W.J. (1990). A revised version of the psychoticism scale for children. *Personality and Individual Differences*, 11, 65–76.

Costa, P.T., & McCrae, R.R. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P.T., & McCrae, R.R. (1995). Primary traits of Eysenck's P-E-N system: Three- and five-factor solutions. *Journal of Personality and Social Psychology*, 69, 308–317.

De Bruyn, E.E.J., Delsing, M.J.M.H., & Welten, M. (1995). The EPQ-R (Junior): A Dutch replication study. *Personality and Individual Differences*, 18, 405–411.

Draycott, S.G., & Kline, P. (1995). The Big Three or the Big Five—the EPQ-R vs. the NEO-PI: A research note, replication and elaboration. *Personality and Individual Differences*, 18, 801–804.

Eysenck, H.J. (1952). *The scientific study of personality*. London: Routledge and Kegan Paul.

Eysenck, H.J. (1959). *The manual of the Maudsley Personality Inventory*. London: University of London Press.

Eysenck, H.J. (1967). *The biological basis of personality*. Springfield, IL: C.C. Thomas.

Eysenck, H.J. (1992). Four ways why five factors are not basic. *Personality and Individual Differences*, 13, 667–673.

Eysenck, H.J., Barrett, P.T., & Eysenck, S.B.G. (1985). Indices of factor comparison for homologous and non-homologous personality scales in 24 countries. *Personality and Individual Differences*, 6, 400–403.

Eysenck, H.J., Barrett, P., Wilson, G., & Jackson, C. (1992). Primary trait measurement of the 21 components of the P-E-N system. *European Journal of Psychological Assessment*, 8(2), 109–117.

- Eysenck, H.J., & Eysenck, S.B.G. (1964). *The manual of the Eysenck Personality Inventory*. London: University of London Press.
- Eysenck, H.J., & Eysenck, S.B.G. (1975). *Manual of the EPQ (Eysenck Personality Questionnaire)*. San Diego, CA: Educational and Industrial Testing Service.
- Eysenck, H.J., & Eysenck, S.B.G. (1991). *Manual for the EPQ-R*. San Diego, CA: EdITS.
- Eysenck, S.B.G., & Eysenck, H.J. (1967). Salivary response to lemon juice as a measure of introversion. *Perceptual and Motor Skills*, 24, 1047–1053.
- Eysenck, S.B.G., Eysenck, H.J., & Barrett, P. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences*, 6, 21–29.
- Fischer, B.E. (1994). Junior Eysenck Personality Questionnaire: Neuroticism, depressive symptoms and sleep disturbance in elementary school age children. *Personality and Individual Differences*, 15, 233–235.
- Fontana, D., & Abouserie, R. (1994). Stress levels, gender and personality factors in teachers. *British Journal of Educational Psychology*, 63(2), 261–270.
- Francis, L.J. (1996). The development of an abbreviated form of the revised Junior Eysenck Personality Questionnaire (JEPQR-A) among 13–15 year olds. *Personality and Individual Differences*, 21, 835–844.
- French, C.C., & Beaumont, J.G. (1989). A computerized form of the Eysenck Personality Questionnaire: A clinical study. *Personality and Individual Differences*, 10, 1027–1032.
- Furnham, A., Petrides, K.V., Jackson, C.J., & Cotter, T. (in press). Do personality factors predict job satisfaction? *Personality and Individual Differences*.
- Gale, A. (1981). EEG studies of extraversion-introversion. In R. Lynn (Ed.), *Dimensions of personality: Papers in honour of H.J. Eysenck* (pp. 181–208). London: Pergamon.
- Gray, J.A. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. Oxford: Oxford University Press.
- Gray, J.A. (1987). *The psychology of fear and stress*. Cambridge: Cambridge University Press.
- Hosokawa, T., & Ohyama, M. (1993). Reliability and validity of a Japanese version of the short-form Eysenck Personality Questionnaire-Revised. *Psychological Reports*, 72, 823–832.
- Jackson, C.J. (1999, October). Beliefs about money in interaction with personality and attributional style as predictors of sales success. *Selection and Development Review*, 9–13.
- Jackson, C.J. (2002). Comparison between Eysenck's and Gray's models of personality in the prediction of motivational work criteria. *Personality and Individual Differences*, 31, 129–144.
- Jackson, C.J., and Corr, P. (1998). Personality-performance correlations at work: Individual and aggregate levels of analysis. *Personality and Individual Differences*, 24, 815–820.
- Jackson, C.J., Furnham, A., Forde, L., & Cotter, T. (2000). The structure of the Eysenck Personality Profiler. *British Journal of Psychology*, 91(2), 223–239.
- Katz, Y.J., & Francis, L.J. (2000). Hebrew revised Eysenck Personality Questionnaire: Short Form (EPQR-S) and Abbreviated

Form (EPQR-A). *Social Behavior and Personality*, 28, 555–560.

Koelega, H.S. (1992). Extraversion and vigilance performance. *Psychological Bulletin*, 112, 239–258.

The Sixteen Personality Factor (16PF) Questionnaire is a comprehensive measure of normal personality that can be used in any setting where an in-depth, integrated picture of the whole person is desirable. From its inception, the 16PF structure has been multilevel, with the 16 primary traits forming the five global factors (Big-Five) at the broadest level of personality. The test has a long history of empirical research, is embedded in a well-established theory of individual differences, and has been found to predict a wide variety of important behavioral criteria. For example, the 16PF has been used to predict leadership, creativity, conscientiousness, social skills, empathy, self-esteem, power dynamics, and coping patterns. The test is widely used internationally and has been translated and adapted into over 30 languages.

### **Tests, Administration, and Scoring**

The 16PF Fifth Edition contains 185 multiple-choice items that are written at a fifth-grade reading level. The test takes about 35 to 50 minutes to complete in paper-and-pencil format (25 to 35 minutes by computer administration). It provides scores on 16 primary factor scales (one of which is a short ability scale), five global (Big-Five) scales, and three validity scales (see Table 4.1 for a description of the personality dimensions). Each primary scale contains 10 to 15 items, with each item having a three-choice answer format. Because the instructions are straightforward and the test is untimed, administration requires little supervision, whether administered individually or in a group.

The paper-and-pencil version has easy hand-scoring instructions. Computer scoring and interpretive reports are available via the OnSite software program. Additionally, answer sheets can be mailed or faxed to the publisher for scoring and interpretation. Computerized administration and scoring are also available online. For international users, the test can be administered and scored in over a dozen different languages via [www.16PFworld.com](http://www.16PFworld.com) (see Computerization section). One distinguishing characteristic of the 16PF Questionnaire is that items tend to sample a broad range of normal behavior by asking the test taker about concrete situations or behaviors, rather than asking questions that force the test taker to make self-ratings and assess his or her own personality traits (e.g., "I am an even-tempered person; I am a warm and friendly person; I am not a worrier."). Instead, 16PF questions tend to ask about actual behavioral situations:

- "When I find myself in a boring situation, I usually "tune out" and daydream about other things. a. true, b. ?, c. false."

### **40 The Sixteen Personality Factor (16PF) Questionnaire**

#### **TABLE 4.1 16PF Factor Names and Descriptors**

Descriptors of Low Range Primary Factors	Descriptors of High Range
Reserved, Impersonal, Distant	Warmth (A) Warm, Participating, Attentive to Others
Concrete, Lower Mental Capacity	Reasoning (B) Abstract, Bright, Fast Learner
Reactive, Affected by Feelings	Emotional Stability (C) Emotionally Stable, Adaptive, Mature
Deferential, Cooperative, Avoids Conflict	Dominance (E) Dominant, Forceful, Assertive
Serious, Restrained, Careful	Liveliness (F) Enthusiastic, Animated, Spontaneous
Expedient, Nonconforming	Rule-Consciousness (G) Rule Conscious, Dutiful
Shy, Timid, Threat-Sensitive	Social Boldness (H) Socially Bold, Venturesome, Thick-Skinned
Tough, Objective, Unsentimental	Sensitivity (I) Sensitive, Aesthetic, Tender-Minded

Trusting, Unsuspecting, Accepting Vigilance (L) Vigilant, Suspicious, Skeptical, Wary  
 Practical, Grounded, Down-to-Earth Abstractedness (M) Abstracted, Imaginative, Idea-Oriented  
 Fortright, Genuine, Artless Privatness (N) Private, Discreet, Nondisclosing  
 Self-Assured, Unworried, Complacent Apprehension (O) Apprehensive, Self-Doubting, Worried  
 Traditional, Attached to Familiar Openness to Change (Q1) Open to Change, Experimenting  
 Group-Oriented, Affiliative Self-Reliance (Q2) Self-Reliant, Solitary, Individualistic  
 Tolerates Disorder, Unexacting, Flexible Perfectionism (Q3) Perfectionistic, Organized, Self-Disciplined  
 Relaxed, Placid, Patient Tension (Q4) Tense, High Energy, Driven  
 Global Factors  
 Introverted, Socially Inhibited Extraversion Extraverted, Socially Participating  
 Low Anxiety, Unperturbable Anxiety High Anxiety, Perturbable  
 Receptive, Open-Minded, Intuitive Tough-Mindedness Tough-Minded, Resolute, Unempathic  
 Accommodating, Agreeable, Selfless Independence Independent, Persuasive, Willful  
 Unrestrained, Follows Urges Self-Control Self-Controlled, Inhibits Urges  
 Adapted with permission from Conn, S. R., & Rieke, M. L. (1994). *The 16PF Fifth Edition technical manual*.  
 Champaign, IL: Institute for Personality and  
 Ability Testing, Inc.

- “When a bit of tact and convincing is needed to get people moving, I’m usually the one who does it. a. true, b. ?, c. false.”

The history of the 16PF Questionnaire spans almost the whole history of personality measurement. It was first published in 1949 and has been updated periodically, with the most recent release being the 16PF Fifth Edition (Cattell, Cattell, & Cattell, 1993). This latest edition has simpler, clearer, updated language; is easier to hand score; has improved psychometric characteristics; has been reviewed for Americans with Disabilities Act (ADA) compliance and gender, cultural, and racial bias; and is now standardized against a sample that reflects the 2000 U.S. census on sex, age, and race.

The 16PF model has always included parallel versions for lower age ranges, such as the 16PF Adolescent Personality Questionnaire (Schuerger, 2001) and the Children’s Personality Questionnaire (Porter & Cattell, 1975). A shorter version, called the 16PF Select Questionnaire (Kelly & Mead, 1999), is also available for employee screening and selection. The publisher encourages research and education on all versions by providing free or discounted materials for classroom and research use.

### **Global or Big-Five Traits**

From the beginning, the 16PF Questionnaire has been a multilevel measure of personality, based on Raymond Cattell’s model (Cattell, Mead, & Cattell, in press; Cattell, 1946). Thus, it provides information about the global or Big-Five personality dimensions, as well as the more basic primary traits (see Tables 4.1 and 4.2). The global level of personality gives an overview of the most general level of functioning, but these global factors can then be broken down into their more specific primary factors to provide a more nuanced, in-depth picture of the unique individual and are stronger predictors of actual behavior (Ashton, 1998; Goldberg, 1972; Mershon & Gorsuch, 1988).

The 16PF global (or Big-Five) traits were created by factor analyzing the primary traits in order to define the broader, underlying structure among the primaries. For this reason,

the global and primary traits are fundamentally related; each global trait is defined by the primary traits that make it up (see Table 4.2). In order to understand an individual, it is necessary to know both global scores and the scores on the primary scales that make up each global.

Extraversion, for example, is defined by the five primary traits that describe fundamental motives for moving toward

versus away from people **TABLE 4.2 The 16PF Model: Primary Factors Contributing to the Five Global Factors**

Global Factors

Extraversion/

Introversion

High Anxiety/

Low Anxiety

Tough-Mindedness/

Receptivity

Independence/

Accommodation

Self-Control/

Lack of Restraint

(A) Warm/Reserved (C) Emotionally Stable/

Reactive

(A) Warm/Reserved (E) Dominant/Deferential (F) Lively/Serious

(F) Lively/Serious (L) Vigilant/Trusting (I) Sensitive/Unsentimental (H) Bold/Shy (G) Rule Conscious/

Expedient

(H) Bold/Shy (O) Apprehensive/

Self-Assured

(M) Abstracted/Practical (L) Vigilant/Trusting (M) Abstracted/Practical

(N) Private/Forthright (Q4) Tense/Relaxed (Q1) Open to Change/

Traditional

(Q1) Open to Change/

Traditional

(Q3) Perfectionistic/

Tolerates Disorder

(Q2) Self-Reliant/

Group-Oriented

- *Reserve vs. Warmth (Factor A)* has to do with the basic desire to seek close, caring connections with others.
- *Seriousness vs. Liveliness (Factor F)* involves having enthusiastic, exuberant energy that leads to interacting with others in stimulating, spontaneous ways.
- *Shyness vs. Social Boldness (Factor H)* describes the tendency to fearlessly seek attention and adventure regardless of risks and criticism.
- *Forthright vs. Privatness (Factor N)* has to do with a tendency to be discrete and guarded about revealing personal feelings, needs, and motivations (being low on this factor—open and self-revealing—contributes positively to Extraversion).
- *Group-Oriented vs. Self-Reliance (Factor Q2)* involves a preference to do things with other people (e.g., work, play, make decisions) and to rely on them, versus the preference to do things alone without assistance or interference from others.

Thus, two people who are at the 80th percentile on Extraversion tend to move toward people to the same degree, but

may do so for very different reasons and in very different ways. For example, one person might move toward others because he or she is caring and warm (A<sub>1</sub>), group oriented and seeking companionship and support (Q2<sub>1</sub>), but shy and modest (H<sub>1</sub>). Another person might show the same level of Extraversion but instead be bold and attention-seeking (H<sub>2</sub>), high spirited and animated (F<sub>1</sub>), but impersonal and insensitive (A<sub>2</sub>). These two Extraverts differ greatly in their sensitivity to others, for example, and would be very different to work for, to supervise, or to live with.

In this way, the 16PF Questionnaire allows the professional to see an individual's personality at different levels of organization and to understand deeper motivations. Another example: Success in motivating someone to accomplish a goal depends not only on knowing their overall level of Self-Control (conscientiousness) but also on knowing whether their Self-Control is motivated by strong internal standards of right and wrong (Rule-Consciousness—G<sub>1</sub>), by a stylistic tendency to be organized, precise, and planful (Perfectionism—Q3<sub>1</sub>), by a focused and practical perceptual style (low Abstractedness—M<sub>1</sub>), or by a serious, stoic temperament (low Liveliness—F<sub>2</sub>).

### **Interpretation**

In-depth understanding of an individual is also facilitated by considering 16PF primary trait combinations. For example, the forceful, aggressive qualities of Dominance (E<sub>1</sub>) can result in positive interpersonal interaction patterns when combined with the sensitive, caring qualities of high Warmth (A<sub>1</sub>) or the calm maturity of high Emotional Stability (C<sub>1</sub>). However, a high level of Dominance can result in a more difficult interactional style when combined with other factors, such as the high-energy, impatient qualities of high Tension (Q4<sub>1</sub>) or the questioning, oppositional qualities of high Vigilance (L<sub>1</sub>).

These types of interpretive strategies and insights can be found in numerous 16PF resource books, such as:

- *The 16PF: Personality in Depth* (Cattell, 1989).
- *16PF Interpretation in Clinical Practice* (Karson, Karson, & O'Dell, 1997).
- *Personality in Practice* (Lord, 1997).
- *Occupational Interpretation of the 16 Personality Factor Questionnaire* (Schuerger & Watterson, 1998).
- *Overcoming Obstacles to Interpretation* (Lord, 1999)

Many computer-generated reports are available, which provide interpretive insights based on expert opinion and over 50 years of 16PF research. These include:

- The Basic Interpretive Report.
- The Narrative Score Report.
- The Personal Career Development Profile Report.
- The Human Resource Development Report.
- The Couple's Counseling Report.
- The Cattell Comprehensive Personality Interpretation.
- The Karson Clinical Report.
- The Teamwork Development Report.
- The Leadership Coaching Report.

Each report provides narrative regarding the 16 primary

scales and five global scales, as well as predictive scores in key behavioral areas such as leadership style, social skills, compatible occupations, empathy, self-esteem, power dynamics, and coping patterns.

### **THEORETICAL BASIS AND TEST DEVELOPMENT**

Although there are many multiscale measures of normal-range personality, the 16PF Questionnaire was developed from a unique perspective. Rather than measuring preconceived dimensions that were of interest to a particular author, the test was developed from the perspective of trying first to discover the basic structural elements of personality and then to construct scales to measure these fundamental dimensions. The 16PF Questionnaire was developed and refined over several decades by Raymond B. Cattell and his colleagues, based on his theory that the structure of human personality could be discovered by breaking down personality into its basic elements or “source traits” (Cattell, 1943). Cattell believed that these fundamental dimensions of personality were the building blocks of all personality and behavior in the same way that physical elements such as oxygen and hydrogen are basic building blocks of the physical world.

Thus, Cattell and his colleagues set about trying to identify and measure these underlying dimensions of personality through a systematic program of research. He believed that “all aspects of human personality which are or have been of importance, interest, or utility have already become recorded in the substance of language” (Cattell, 1943, p. 483). This “lexical hypothesis” is now supported by a wide range of influential psychologists (Goldberg, 1993) and is the basis for the recent Big-Five theories.

Cattell and his colleagues began with Allport and Odbert’s (1936) compilation of all known personality traits or descriptors in the English language. They supplemented this list with concepts from psychological theorists. They studied these traits by looking at their patterns in actual peer ratings, self-report questionnaires, and behavioral measures. Repeated cluster analyses were performed, and the newly developing methods of factor analysis were used—a powerful tool for discovering and mapping the important, underlying influences among a vast array of observable variables.

After years of factor analytic study, Cattell and his colleagues arrived at their set of underlying source traits—the traits measured by the 16PF Questionnaire (see Table 4.1). This personality structure has been replicated in many studies based on samples differing in language, culture, and education (see Validity section). Its robustness and predictive utility in many kinds of settings probably result from studying personality traits that could be found across all three sources (peer ratings, self-report, and objective behavior measures). The 16PF dimensions also benefit from being embedded within a broader theoretical model that addresses individual differences at multiple ages, through a life-span developmental perspective, and in relation to abilities and motivation (Cattell, 1979).

### **Global or Big-Five Structure**

From the beginning, Cattell’s theory of personality was hierarchical

in structure (Cattell, 1946). Generally, the focus was on the primary-level traits because they provide a more finegrain definition of the individual's personality and are more powerful in predicting behavior (Ashton, 1998; Goldberg, 1972; Mershon & Gorsuch, 1988). However, Cattell found originally that when he factor analyzed the 16 primary traits themselves to find the underlying organizing influences among them, the "second-order" or global factors emerged. These five global factors, which describe personality at the most general, broad-brush level, have changed very little since Cattell's 1957 book. The same five have been scoreable from the 16PF Questionnaire since the release of the fourth edition in 1967 (Cattell, Eber, & Tatsuoka, 1970).

In fact, research leading to the development of the recent Big-Five factors of personality was based on Cattell's original scales (e.g., Costa & McCrae, 1976, 1985; Norman, 1963; Tupes & Christal, 1961). Comparisons between the five 16PF global factors and other Big-Five measures such as the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) indicate a high level of concordance (Cattell, 1996). However, the five 16PF global factor definitions have an important distinction: While other Big-Five systems arbitrarily forced orthogonal factor locations on the data because of statistical convenience, the 16PF oblique rotation methods allowed the data to determine the factors, enhancing the definitions of the factors.

#### **16PF Fifth Edition**

The test has continued to be refined since 1949, resulting in four revisions, in 1956, 1962, 1968, and the Fifth Edition in 1993. This most recent edition was developed by first finding the best items from all previous 16PF forms, and then rewriting these to improve, simplify, and modernize them. Additionally, new items were written by the authors and other 16PF experts. All items were analyzed for gender, race, and disability bias. Cross-cultural translatability was also evaluated. Final items were selected through a series of factor analyses based on diverse national samples of 1,204, 646, 872, and 3,498 participants. A more detailed description of this process can be found in the *16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994).

#### **PSYCHOMETRIC CHARACTERISTICS**

Because of the scientific nature of the test's origins, it shows strong psychometric characteristics supported by a long history of international research.

##### **Reliability**

The reliabilities of the 16PF Fifth Edition Questionnaire primary and global scales are comparable to other personality measures, despite the fact that 16PF scales are fairly short (10 to 15 items). These reliability coefficients are presented in Table 4.3, and further information can be found in the *16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994). Internal consistency reliability (how highly the items in a scale correlate with each other) for the 16PF primary scales averages .75 (ranging from .66 to .86 over the 16 scales) summarized across two general population samples and one university student sample with a total of 4,660 participants.

Test-retest reliabilities (or estimates of the consistency of scores over time) for a 2-week interval ranged from .69 to .87 with a median of .80. Two-month test-retest reliabilities ranged from .56 to .79 with a median of .69. The 16PF global scales have even higher reliabilities, with 2-week test-retest estimates ranging from .84 to .91 with a mean of .87, and 2-month test-retest estimates ranging from .70 to .82 with a median of .80.

### **Validity**

Because 16PF scales were developed through factor analytic methods, the results of these methods provide evidence about the construct validity of the 16 primary scales and the five global factors. A summary of these analyses is provided in the *16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994). The 16PF personality structure has been replicated in studies based on samples differing in language, culture, and education (Bolton, 1977; Boyle, 1989; Cattell, 1946, 1973; Cattell, et al., 1970; Cattell & Krug, 1986; Chernyshenko, Stark, & Chan, 2001; Gerbing & Tuley, 1991; Gorsuch & Cattell, 1967; Hofer, Horn, & Eber, 1997; Krug & Johns, 1986; Mershon & Gorsuch, 1988; Mogenet & Rolland, 1995; Motegi, 1982; Schneewind & Graf, 1998).

Other studies have provided evidence of a high degree of correspondence between the five 16PF global factors and other Big-Five measures (Cattell, 1996). Not only do the NEO PI-R and 16PF five-factor models line up well in factor analyses, but the five 16PF global scales were found to correlate as highly with the NEO PI-R five factors as the two main Big-Five models (NEO-PI-R and Goldberg's Big-Five) correlate with each other.

Many studies provide evidence for the meaning of the 16PF scales by comparing them to similar scales in various other measures of normal personality. Conn and Rieke (1994), for example, provide correlations with all scales from the California Psychological Inventory, the NEO-PI-R, the Personality Research Form, and the Myers-Briggs Type Indicator.

The comprehensiveness of the test has been supported by findings that all dimensions on other major personality tests overlap substantially with the 16PF scales in regression and factor analytic studies (Conn & Rieke, 1994; Cattell, 1996). A substantial body of criterion validity data has accumulated demonstrating the utility of the 16PF scales in predicting a wide range of behavior. For example, these criteria include self-esteem, leadership style, self-discipline, interpersonal needs, empathy, marital compatibility, coping patterns, decision-making style, frustration tolerance, cognitive processing style, compatible career choices, and job performance in a wide range of occupations (e.g., Cattell, et al., 1970; Cattell & Krug, 1986; Cattell, et al., in press; Conn & Rieke, 1994; Guastello & Rieke, 1993b; Krug & Johns, 1990).

In addition, the 16PF Questionnaire has been found to be more powerful than other popular personality inventories in predicting real-life behavior. In a recent study, Goldberg (in press) compared the ability of several popular personality inventories to predict six clusters of behavioral criteria and

found that the 16PF International Personality Item Pool **The Sixteen Personality Factor (16PF) Questionnaire**

**TABLE 4.3 Reliability Estimates for 16PF Fifth Edition Scales**

Internal Consistency Test-Retest

Sample

Primary Scales

1

N\_820

2

N\_2,500

3

N\_1,340

Average

N\_4,660

2-week

N\_204

2-month

N\_159

Warmth (A) .69 .69 .74 .70 (0.82) .83 .77

Reasoning (B) .76 .77 .68 .74 (0.75) .69 .65

Emotional Stability (C) .78 .78 .77 .78 (0.91) .75 .67

Dominance (E) .71 .66 .70 .68 (0.81) .77 .69

Liveliness (F) .73 .72 .70 .72 (0.68) .82 .69

Rule-Consciousness (G) .74 .75 .77 .75 (0.97) .80 .76

Social Boldness (H) .86 .85 .87 .86 (0.74) .87 .79

Sensitivity (I) .79 .77 .79 .78 (0.98) .82 .76

Vigilance (L) .74 .74 .70 .73 (1.02) .76 .56

Abstractedness (M) .74 .74 .75 .74 (1.01) .84 .67

Privateness (N) .77 .75 .78 .76 (0.87) .77 .70

Apprehension (O) .78 .78 .79 .78 (0.94) .79 .64

Openness to Change (Q1) .71 .64 .68 .66 (1.11) .83 .70

Self-Reliance (Q2) .78 .78 .78 .78 (0.89) .86 .69

Perfectionism (Q3) .73 .71 .76 .73 (0.96) .80 .77

Tension (Q4) .75 .76 .74 .75 (0.93) .78 .68

Global Scales<sup>b</sup>

Extraversion .91 .80

Anxiety .84 .70

Tough-Mindedness .87 .82

Independence .84 .81

Self-Control .87 .79

aAverage internal consistency values were weighted with respect to sample size. Standard error of measurement estimates,

using weighted standard deviations, are presented in parentheses.

bInternal consistency values are not available for the global factor scales because their scores are derived from combinations

of the 16 primary factor scores.

Adapted with permission from Conn, S. R., & Rieke, M. L. (1994). *The 16PF Fifth Edition technical manual*. Champaign,

IL: Institute for Personality and Ability Testing, Inc.

(IPIP) version (Goldberg, 1999) had the highest predictive validity coefficients.

**RANGE OF APPLICABILITY AND LIMITATIONS**

The 16PF Questionnaire is a comprehensive measure of normal adult personality, and thus it has a wide range of useful applications in varied counseling and clinical settings (including career counseling and other personal development settings); a range of industrial/organizational settings (employee

selection, development, training, coaching, outplacement, and team building); as well as school and research settings. Because of its comprehensive nature and long history of research, the 16PF Questionnaire can be used to understand a wide range of behavioral patterns from self-esteem, coping patterns, and interpersonal needs to leadership, social skills, and compatible career choices (see Validity section).

The 16PF test was developed to measure normal behavior and cannot be used to measure pathological dimensions or diagnose mental disorders. For these purposes, a clinical measure should be administered in addition. Nonetheless, normal personality dimensions are important in clinical settings; for example, in quickly gathering a comprehensive picture of the individual's whole personality, facilitating empathy and rapport in the therapeutic relationship, choosing therapeutic approaches, planning developmental goals, identifying relevant adjustment issues, and helping the client develop greater self-awareness (Cattell, 1989; Karson et al., 1997).

As with all tests, 16PF results should be treated as hypotheses to be combined with other sources of collateral information (interview data, history information, other psychological measures such as those of pathological dimensions or cognitive functioning, etc.) to arrive at a prediction of behavior.

#### **CROSS-CULTURAL FACTORS**

Because of its international origins, the 16PF primary factor structure has been evaluated and confirmed across many languages and cultures; for example, France (Mogenet & Rolland, 1995), Spain (Prieto, Gouveia, & Fernandez, 1996), Germany (Schneewind & Graf, 1998), the United Kingdom (Smith, 1994), Japan (Motegi, 1982), Latin America (Krug, 1971), and Italy (Barbaranelli & Caprara, 1996). Currently, the 16PF Questionnaire has been translated into over 30 languages worldwide and can be administered via [www.16PFworld.com](http://www.16PFworld.com) in over a dozen different languages. It is used worldwide, for example, in the selection of managers (Bartram, 1992; Chakrabarti & Kundu, 1984), salespersons (Coyne Didsbury Pty. Ltd., 1998), pilots (Bartram and Baxter, 1996), and police officers (Cooper, Robertson, & Sharman, 1986). It is also used internationally for research into diverse topics such as occupational choice (Arbeo, 1994), self-actualization (Kapoor & Shankhla, 1994), student giftedness (Drabkova & Drabkova, 1998), work team roles (Dulewicz, 1995), and adaptation to renal failure treatment (Carbonell, Hernandez, & Ramos, 1992).

#### **ACCOMMODATION FOR POPULATIONS WITH DISABILITIES**

Braille, American Sign Language, and audiotape formats are available for previous editions of the 16PF Questionnaire, but have not been developed and validated for the Fifth Edition. However, the issue of low literacy has been addressed by lowering the reading grade level of the 16PF Fifth Edition to a fifth-grade level. Using the computerized administration can accommodate some disabilities. Generally, for individuals with greater visual or other physical impairments, an objective person can read the test questions to the test taker and/or record answers. Answers can also be spoken into a

tape recorder and later transcribed onto an answer sheet. Since the 16PF Questionnaire is not timed, many individual accommodations can be made that would not interfere with the test administration. Because the test measures normal personality dimensions, test users do not have to be concerned with accommodating individuals with mental disabilities.

### **LEGAL AND ETHICAL CONSIDERATIONS**

As with all tests, the validity of 16PF results depends on proper test selection, administration, scoring, and interpretation by the professional test user. Therefore, it is incumbent on all test users to read *The 16PF Fifth Edition Administrator's Manual* (Russell & Karol, 1994) and *The 16PF Fifth Edition Technical Manual* (Conn & Rieke, 1994) and to gain proper training in the use of the 16PF test. Professional workshops are given each year by the publisher (the Institute for Personality and Ability Testing [IPAT]: (800) 225-4728; [www.ipat.com/workshop.html](http://www.ipat.com/workshop.html)) and other international publishers.

All test users should read the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and be aware of the meaning of measurement error and the limitations of test scores. The 16PF Questionnaire was developed and documented in accordance with these standards. As with all test scores, 16PF results should be combined with information from other sources in drawing conclusions.

### **COMPUTERIZATION**

Computer administration, scoring, and report generation is available on a PC using the OnSite software system (which also allows for hand entering or scanning of pencil-marked answer sheets). This system also provides a disk that can be used to administer the test at several different locations before returning to the office to score the accumulated data and generate reports. The professional can import, export, and organize test data, and the archive function allows the professional to compress and back up data to a network, hard drive, or floppy disk when space or safekeeping are a concern.

Online administration of the 16PF Questionnaire is also available via NetAssess. After the client completes the test online using a pass code, the test is scored and reports are e-mailed to the professional. Online multilanguage testing is available using [www.16PFworld.com](http://www.16PFworld.com), which can administer the test in over a dozen different languages and score the test with national norms for that language group. The narrative report can be produced in the same language as the test or in one or more of the other available languages. This provides the professional with results based on the same personality dimensions across different language groups.

### **CURRENT RESEARCH STATUS**

The extensive body of research findings, accumulated over half a century, that link 16PF traits to important real-life criteria enhances use of the 16PF Questionnaire. A conservative estimate of 16PF research since 1974 includes upwards of 2,000 publications (Hofer & Eber, 2001). Because the test

has been used in varied settings, research has generated a wide range of profiles and prediction equations for criteria such as creativity, emotional intelligence, self-esteem, marital satisfaction, coping patterns, leadership style, interpersonal skills, cognitive processing style, career preferences, job performance, and academic achievement, as well as dozens of occupational profiles (e.g., Cattell, et al., in press; Cattell, et al., 1970; Conn & Rieke, 1994; Karol & Russell, 1995; Krug & Johns, 1990).

### **USE IN CLINICAL OR ORGANIZATIONAL PRACTICE**

The 16PF Questionnaire has a long history of use in a wide variety of applied settings. The two most general ones are described here.

#### **Clinical and Counseling Settings**

Although the 16PF Questionnaire was not constructed to measure pathology or to make a clinical diagnosis, it has been found to be valuable in many aspects of clinical and counseling practice because of its ability to provide a rich, comprehensive, objective picture of the individual's enduring personality makeup. Expert clinicians (Karson et al., 1997) emphasize the importance of getting a picture of the whole person as a context in which to place information gleaned from other tests about symptoms and pathology. These authors demonstrate how 16PF results provide an understanding of basic interpersonal, experiential, emotional, and motivational dynamics and lead to effective treatment planning, especially in time-limited settings.

The Karsons also stress the importance of 16PF scores in increasing the client's self-awareness of both strengths and weaknesses, as well as facilitating early empathy and rapport in the therapy relationship. Unlike pathological dimensions, 16PF scores can be discussed openly and easily with clients to stimulate interaction and enable an early alliance in the assessment and therapy process. The Karson Clinical Report (Karson & Karson, 1995) discusses a range of topics including self-esteem, patterns of coping, frustration tolerance, ability to form and maintain gratifying relationships, cognitive processing style, decision making, and motivational dynamics. These and other experienced clinicians (Cattell, 1989, 1995) have shown the usefulness of 16PF scores in understanding a range of clinically relevant issues such as an individual's capacity for insight and introspection, bases for self-evaluation, internalization of societal standards, quality of attachments, interpersonal needs, capacity for intimacy, and openness to change. The *16PF Cattell Comprehensive Personality Interpretation Manual* (Cattell & Cattell, 1997) demonstrates the value of 16PF results in planning a range of developmental and therapeutic goals, for instance, strategies for developing a good working alliance, power dynamics, overcoming resistance to change, facilitating termination, and choosing between modalities such as group versus individual therapy or directive versus nondirective therapies.

16PF scores have also proven useful in marital or couple's counseling. Karol and Russell (1995) summarize research in the area of marital satisfaction and discuss how 16PF results

help the counselor to better understand both individuals, how their traits might combine and interact, and which therapeutic goals and approaches might be helpful.

### **Industrial and Organizational Settings**

The 16PF Questionnaire is commonly used to help organizations in a variety of functions, including hiring, promotion, development, training, outplacement, team building, and coaching. Meta-analytic reviews of the literature (e.g., Barrick & Mount, 1991) have shown measures of normal, adult personality to be valid predictors of job performance in a wide range of occupations. Although other sources of information are also important, 16PF scores provide an objective, comprehensive, and efficient source of information for the employer.

An extensive literature has accumulated demonstrating the ability of 16PF dimensions to predict important aspects of job performance and occupational satisfaction (e.g., Cattell, et al., in press; Cattell, et al., 1970; Conn & Rieke, 1994; Krug & Johns, 1990; Schuerger & Watterson, 1998). For example, 16PF scores have been used with such diverse occupations as salesperson (Guastello & Rieke, 1993a; Tucker, 1991), manager or executive (Guastello & Rieke, 1993b; Johns, Schuerger, & Watterson, 1980), accountants (Anonsen, 1985), police and security personnel (Fabricatore, Azen, Shoentgen, & Snibbe, 1978; Scholl, 2000), programmer/systems analyst (Schuerger, Johns, & Stazyk, 1985), entrepreneur (Aldridge, 1997), and factory worker (Schuerger, Ekeberg, & Kustis, 1994).

In addition, the 16PF Select (Kelly & Mead, 1999) was developed specifically for use in personnel selection. It is a shorter form of the test that includes the scales that are most predictive of job performance across a wide variety of jobs.

The 16PF Select is not just a personality measure but a process in which the professional defines the personality characteristics that are most important for effective job performance via regression weights or an optimal-range score approach. It provides concise feedback comparing an applicant's personality and behavioral strengths to desired personality dimensions for the job. The extensive history of 16PF research has also produced equations for a wide range of occupationally relevant dimensions, many of which appear in the interpretive reports listed above. These include such dimensions as different types of leadership styles, leadership effectiveness, cognitive processing style, creativity and imagination, problem-solving style, risk taking, independence and initiative taking, emotional intelligence, social skills, conscientiousness, learning style, academic achievement, potential for on-the-job learning, and Holland Occupational Themes (e.g., Cattell, et al., 1970; Conn & Rieke, 1994; Guastello & Rieke, 1993b; Schuerger & Watterson, 1998; Walter, 2000).

Experts on consulting with the 16PF Questionnaire in industrial/organizational settings (e.g., Cattell, 1989; Edelstein, 2001; Lord, 1997, 1999; Walter, 2000; Watterson, Aldridge, & Seelback, 2002) provide practical advice about using 16PF results in a range of industrial/organizational consulting situations, such as executive selection, employee development or coaching, team building, outplacement, and giving feedback

from test results. They demonstrate how to use 16PF scores to answer questions such as, Will the person function effectively in jobs that require a strong technical orientation? Can he or she be counted on to finish things he or she starts? Will he or she be an effective leader? Is this the kind of person who is likely to handle high-stress situations well? The 16PF test is widely used for career development purposes both inside and outside organizations. There are many resources for identifying the degree of fit between an individual's 16PF scores and those scores characteristic of persons in a range of occupations or occupational types (Cattell, et al., 1970; Conn & Rieke, 1994; Guastello & Rieke, 1993b; Lord, 1997, 1999; Schuerger & Watterson, 1998). For example, the Personal Career Development Profile interpretive report (Walter, 2000) compares an individual's 16PF score patterns to those of about 90 different occupational interest groups. Additionally, 16PF feedback increases self-awareness and allows the individual to understand his or her strengths and weaknesses relative to a particular career path and to plan self-development goals.

#### **FUTURE DEVELOPMENTS**

There will soon be a new tool for 16PF interpretation—a volume in Wiley's series of Essentials in Psychological Assessment called *Essentials of the 16PF Questionnaire*. This book will provide step-by-step guidance in the knowledge and practical skills necessary for test administration, scoring, and interpretation, useful to professionals or graduate students. Additionally, a Spanish-American version of the 16PF Fifth Edition Questionnaire will soon be released. A thorough translation process, involving translators from many countries (e.g., Mexico, Cuba, Puerto Rico, Nicaragua, Peru, Argentina, Colombia, the Dominican Republic) and item response theory analyses, make this a quality assessment for the growing U.S. Spanish-speaking population. Online, multilanguage testing via [www.16PFworld.com](http://www.16PFworld.com) is continually expanding the number of languages available for reliable administration, scoring, and reports for multinational employers.

A revision of the Children's Personality Questionnaire, a 16PF version for 8- to 12-year-olds, is currently under development.

An in-depth 16PF leadership development report for use with executives, managers, and supervisors has just been released (Watterson, Aldridge, & Seelback, 2002).

#### **REFERENCES**

- Aldridge, J.H. (1997). *An occupational personality profile of the male entrepreneur as assessed by the 16PF Questionnaire*. Unpublished doctoral thesis, Cleveland State University, Cleveland, OH.
- Allport, G.W., & Odbert, H.S. (1936). Traitnames. A psycho-lexical study. *Psychological Monographs*, 47(211), 171.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anonsen, M.K. (1985). *Personality and interests of artists and accountants*. Unpublished master's thesis, Cleveland State University, Cleveland, OH.

- Arbeo, B.J.G. (1994). Analisis correlacional entre variables de personalidad y la indecision vocacional compleja. *Informacion Psicologica*, 55, 11–17.
- Ashton, M.C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior*, 19, 289–303.
- Barbaranelli, C., & Caprara, G.B. (1996). How many dimensions to describe personality: A comparison of Cattell, Comrey, and the Big Five taxonomies of personality traits. *European Review of Applied Psychology*, 46, 15–24.
- Barrick, M.R., & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26.
- Bartram, D. (1992). The personality of UK managers: 16PF norms for short-listed applicants. *Journal of Occupational & Organizational Psychology*, 65, 159–172.
- Bartram, D., & Baxter, P. (1996). Validation of the Cathay Pacific Airways pilot selection. *International Journal of Aviation*, 6, 149–169.
- Bolton, B. (1977). Evidence for the 16PF primary and secondary factors. *Multivariate Experimental Clinical Research*, 3, 1–15.
- Boyle, B.J. (1989). Re-examination of the major personality type factors in the Cattell, Comrey, and Eysenck scales: Were the factor solutions by Noller et al. optimal? *Personality and Individual Differences*, 10, 1289–1299.
- Carbonell, C., Hernandez, L., & Ramos, J. (1992). Variables asociadas a la adaptacion al tratamiento de los enfermos renales cronicos. *Psicopatologia*, 12(4), 153–156.
- Cattell, H.B. (1989). *The 16PF: Personality in depth*. Champaign, IL: Institute for Personality and Ability Testing.
- Cattell, H.B., (1995). *The six-step method for interpreting the 16PF Questionnaire*. Unpublished manuscript.
- Cattell, H.B., and Cattell, H.E.P. (1997). *16PF Cattell Comprehensive Personality Interpretation manual*. Champaign, IL: Institute for Personality and Ability Testing.
- Cattell, H.E.P. (1996). The original Big-Five: A historical perspective. *European Review of Applied Psychology*, 46, 5–14.
- Cattell, H.E.P., Mead, A.D., & Cattell, R.B. (in press). The Sixteen Personality Factor Questionnaire. In S.R. Briggs, J.M. Cheek, & F.M. Donohue (Eds.), *Handbook of adult personality inventories*. New York: Kluwer-Plenum.
- Cattell, R.B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, 38, 476–506.
- Cattell, R.B. (1946). *The description and measurement of personality*. New York: World Book.
- Cattell, R.B. (1957). *Personality and motivation structure and measurement*. New York: World Book.
- Cattell, R.B. (1973). *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.
- Cattell, R.B. (1979). *Personality and learning theory: Vol. 1. The structure of personality in its environment*. New York: Springer.
- Cattell, R.B., Cattell, A.K., & Cattell, H.E.P. (1993). *Sixteen Personality Factor Questionnaire, Fifth Edition*. Champaign, IL: Institute for Personality and Ability Testing.
- Cattell, R.B., Eber, H.W., & Tatsuoka, M.M. (1970). *Handbook for*

- the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing.
- Cattell, R.B., & Krug, S.E. (1986). The number of factors in the 16PF: A review of the evidence with special emphasis on the methodological problems. *Educational and Psychological Measurement*, *46*, 509–522.
- Chakrabarti, P.K., & Kundu, R. (1984). Personality profiles of management personnel. *Personnel Studies*, *29*(2), 143–146.
- Chernyshenko, E.S., Stark, S., & Chan, K.Y. (2001). Investigating the hierarchical factor structure of the Fifth Edition of the 16PF: An application of the Schmid-Leiman orthogonalization procedure. *Educational and Psychological Measurement*, *61*, 290–302.
- Conn, S.R., & Rieke, M.L. (1994). *The 16PF Fifth Edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.
- Cooper, C.L., Robertson, I.T., & Sharman, P. (1986). A psychometric profile of British police officers authorized to carry firearms: A pilot study. *International Review of Applied Psychology*, *35*, 539–547.
- Costa, P.T., & McCrae, R.R. (1976). Age differences in personality structure: A cluster analytic approach. *Journal of Gerontology*, *31*, 564–570.
- Costa, P.T., & McCrae, R.R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P.T., & McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO-PI-R Five-Factor Inventory (NEO-PI-R-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Coyne Didsbury Pty. Ltd. (1998, November). *The relationship between bottom-line sales performance and two predictive measures of behaviour within the Acme International: A validation Study*. (Technical Series: Number 5.) Melbourne, Australia: Author.
- Drabkova, H., Sr., & Drabkova, H., Jr. (1998). On the problem of the relationship of personality traits to intelligence in intellectually above-average individuals. *Ceskoslavenska Psychologie*, *42*, 462–465.
- Dulewicz, V. (1995). A validation of Belbin's team roles from 16PF and OPQ using bosses' ratings of competence. *Journal of Occupational and Organizational Psychology*, *68*, 81–99.
- Edelstein, B.C. (2001, April). *Executive coaching and development: Does it work? What makes it work?* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.
- Fabricatore, J., Azen, S., Shoentgen, S., & Snibbe, H. (1978). Predicting performance of police officers using the Sixteen Personality Factor Questionnaire. *American Journal of Community Psychology*, *6*, 63–70.
- Gerbing, D.W., & Tuley, M.R. (1991). The 16PF related to the fivefactor model of personality: Multiple-indicator measurement versus the a priori scales. *Multivariate Behavioral Research*, *26*(2), 271–289.
- Goldberg, L.R. (1972). Parameters of personality inventory construction and utilization: A comparison of predictive strategies and tactics. *Multivariate Behavioral Research Monographs*, *72*(2), 1–59.

Goldberg, L.R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.

Goldberg, L.R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several fivefactor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe: Volume 7* (pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L.R. (in press). The comparative validity of adult personality inventories. Applications of a consumer-testing framework. In S.R. Briggs, J.M. Cheek, & E.M. Donohue (Eds.), *Handbook of adult personality inventories*. New York: Kluwer-Plenum.

Gorsuch, R. & Cattell, R.B. (1967). Second stratum personality factors defined in the questionnaire realm by the 16PF. *Multivariate Behavioral Research*, 2, 211–224.

Guastello, S.J., & Rieke, M.L. (1993a). *Technical report #1: Selecting successful salespersons with the 16PF*. Champaign, IL: Institute for Personality and Ability Testing.

Guastello, S.J., & Rieke, M.L. (1993b). *Technical report #2: The 16PF & leadership: Summary of research findings 1954–1992*. Champaign, IL: Institute for Personality and Ability Testing.

Hofer, S.M., & Eber, H.W. (2001). Second-order factor structure of the Cattell Sixteen Personality Factor Inventory (16PF). In B. De Raad & M. Perugini (Eds.), *Big-Five assessment*. Goettingen, Germany: Hogrefe & Huber, 397–404.

Hofer, S.M., Horn, J.L., & Eber, H.W. (1997). A robust five-factor structure of the 16PF: Evidence from independent rotation and confirmatory factorial invariance procedures. *Personality and Individual Differences*, 23, 247–269.

Johns, E.F., Schuerger, J.M., & Watterson, D.G. (1980, May). *Personality measures as predictors of managerial performance and salaries*. Paper presented at the meeting of the Midwest Society for Multivariate Experimental Psychology, St. Louis, MO.

Kapoor, S., & Shankhla, S. (1994). A study of secondary motives in relation to Cattell's personality dimensions. *Indian Journal of Clinical Psychology*, 21(2), 1–6.

Karol, D.L., & Russell, M.T. (1995). Appendix A: Summary of recent research: 16PF fifth edition questionnaire and relationship adjustment. In M.T. Russell (Ed.), *The 16PF fifth edition couple's counseling report user's guide*. Champaign, IL: Institute for Personality & Ability Testing.

Karson, S., & Karson, M. (1995). *The 16PF fifth edition Karson clinical report manual*. Champaign, IL: Institute for Personality and Ability Testing.

Karson, S., Karson, M., & O'Dell, J.W. (1997). *16PF interpretation in clinical practice: A guide to the fifth edition*. Champaign, IL: Institute for Personality and Ability Testing.

Kelly, M.L., & Mead, A.D. (1999). *16PF Select manual*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Krug, S.E. (1971). *The 16PF in Latin America*. Champaign, IL: Institute for Personality and Ability Testing, 83–98.

Krug, S.E., & Johns, E.F. (1986). A large scale cross-validation of second-order personality structure defined by the 16PF. *Psychological Reports*, 59, 683–693.

Krug, S.E., & Johns, E.F. (1990). The 16PF. In C.E. Watkins, Jr. & V.L. Campbell (Eds.), *Testing in counseling practice*. Hillsdale, NJ: Erlbaum.

Lord, W. (1997). *Personality in practice*. Windsor, Berkshire (U.K.):

- NFER-Nelson Publishing Co.
- Lord, W. (1999). *16PF5: Overcoming obstacles to interpretation*. Windsor, Berkshire (U.K.): NFER-Nelson Publishing Co.
- Mershon, B., & Gorsuch, R.L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology*, 5, 675–680.
- Mogenet, J.L., & Rolland, J.P. (1995). *16PF5 de R.B. Cattell*. Paris, France: Les Editions du Centre de Psychologie Appliquée.
- Motegi, M. (1982). *Japanese translation and adaptation of the 16PF*. Tokyo: Nihon Bunka Kagakusha.
- Norman, W.T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574–583.
- Porter, R.B., & Cattell, R.B. (1975). *Children's Personality Questionnaire handbook*. Champaign, IL: Institute for Personality and Ability Testing.
- Prieto, J.M., Gouveia, V.V., & Fernandez, M.A. (1996). Evidence on the primary source-trait structure in the Spanish 16PF, 5th edition. *European Review of Applied Psychology*, 46, 33–43.
- Russell, M.T., & Karol, D. (1994). *The 16PF fifth edition administrator's manual*. Champaign, IL: Institute for Personality and Ability Testing.
- Schneewind, K.A., & Graf, J. (1998). *Der 16-Personlichkeits-Factoren-Test Revidierte Fassung test-manual*. Bern, Switzerland: Verlag Hans Huber.
- Scholl, D. (2000, April). *Practicality and efficacy in assessing suitability for employment*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Schuerger, J.M. (2001). *16PF Adolescent Personality Questionnaire manual*. Champaign, IL: Institute for Personality and Ability Testing.
- Schuerger, J.M., Ekeberg, S.E., & Kustis, G.A. (1994). 16PF scorers and machine operators' performance. *Performance and Motor Skills*, 79, 1426.
- Schuerger, J.M., Johns, E.F., & Stazyk, E. (1985). *Systems analysts: Personality and performance ratings*. Unpublished manuscript.
- Schuerger, J.M., & Watterson, D.G. (1998). *Occupational interpretation of the 16 Personality Factor Questionnaire*. Cleveland, OH: Watterson & Associates.
- Smith, P. (1994). *16PF5: The UK standardization of the 16PF5: A supplement of norms and technical data*. Windsor, Berkshire (U.K.): NFER-Nelson Publishing Co.
- Tucker, T.L. (1991). *Investigating sales effectiveness in the automobile industry in relation to personality variables as measured by the 16PF Questionnaire*. Unpublished doctoral dissertation. Pasadena, CA: Fuller Theological Seminary.
- Tupes, E.C., & Christal, R.E. (1961). *Recurrent personality factors based on trait ratings* (Tech. Rep. Nos. 61–67). Lackland, TX: U.S. Air Force Aeronautical Systems Division.
- Walter, V. (2000). *16PF personal career development profile: Technical and interpretive manual*. Champaign, IL: Institute for Personality and Ability Testing.
- Watterson, D.G., Aldridge, J.H., & Seelback, M. (2002). *The 16PF*

*Leadership Coaching report.* Champaign, IL: Institute for Personality and Ability Testing.

## Chapter 7

# **MINNESOTA MULTIPHASIC PERSONALITY INVENTORY— ADOLESCENT**

The Minnesota Multiphasic Personality Inventory—Adolescent (MMPI-A; Butcher et al., 1992) is a downward extension of the MMPI-2 (see Chapter 6) for adolescents between the ages of 14 and 18. Like the MMPI-2, the MMPI-A is a broadband measure of the major dimensions of psychopathology found in Axis I disorders and some Axis II disorders of the *DSM-IV-TR* (American Psychiatric Association, 2000). The MMPI-A consists of 8 validity and 10 clinical scales in the basic profile, along with 15 content scales, and 11 supplementary scales (see Table 7.1). There also are subscales for the clinical scales and content scales with more than 100 scales that can be scored and interpreted on the MMPI-A. Table 7.2 provides general information on the MMPI-A.

### **HISTORY**

The original standardization sample of the MMPI consisted of white Minnesota adults, primarily between the ages of 16 and 55 (Hathaway & McKinley, 1940). Hathaway and Monachesi (1963), in an extensive study of the MMPI performance of adolescents, administered the MMPI to 3,971 ninth graders (mean age about 15) in the Minneapolis public school system during the 1947 to 1948 school year. Two and four years later, they determined how many of these students had records with the local juvenile division of the police department or probation office. During the spring of 1954, these researchers tested 11,329 additional ninth graders who represented a sample of the entire state of Minnesota. Three years later, they determined how many of these students had records by examining local community police and court files.

In 1957, when most of this second set of ninth graders were now twelfth graders (mean age about 18), Hathaway and Monachesi readministered the MMPI to 3,976 students. Hathaway and Monachesi obviously have a wealth of data, only part of which is relevant to the issue of how adolescents differ from adults in terms of their MMPI performance. (The reader who is interested in the use of the original MMPI to predict delinquency in adolescents, which was the main thrust of Hathaway and Monachesi's research project, should consult their book.)

Relevant to the current topic of MMPI norms, Hathaway and Monachesi found that ninth graders had mean scores with *K*-corrections on Scales 4 (*Pd*: Psychopathic Deviate), 8 (*Sc*: Schizophrenia), and 9 (*Ma*: Hypomania), which were approximately 10 T points

205

### **Table 7.1 MMPI-A scales**

#### **Validity Scales**

? Cannot Say  
*VRIN* Variable Response Consistency  
*TRIN* True Response Consistency  
*F1* Infrequency (First Half)  
*F2* Infrequency (Second Half)  
*F* Infrequency  
*L* Lie  
*K* Correction

#### **Clinical Scales**

1 (*Hy*) Hypochondriasis  
2 (*D*) Depression  
3 (*Hy*) Hysteria  
4 (*Pd*) Psychopathic Deviate  
5 (*Mf*) Masculinity-Femininity  
6 (*Pa*) Paranoia  
7 (*Pt*) Psychasthenia  
8 (*Sc*) Schizophrenia  
9 (*Ma*) Hypomania

0 (Si) Social Introversion

**Content Scales**

*A-anx* Anxiety

*A-obs* Obsessions

*A-dep* Depression

*A-hea* Health Concerns

*A-aln* Alienation

*A-biz* Bizarre Mentation

*A-ang* Anger

*A-cyn* Cynicism

*A-con* Conduct Problems

*A-lse* Low Self-Esteem

*A-las* Low Aspirations

*A-sod* Social Discomfort

*A-fam* Family Problems

*A-sch* School Problems

*A-trt* Negative Treatment Indicators

**Supplementary Scales**

*MAC-R* MacAndrew Alcoholism—Revised

*ACK* Alcohol/Drug Acknowledgment

*PRO* Alcohol/Drug Problems Proneness

*IMM* Immaturity

*A* Anxiety

*R* Repression

**PSY-5 Scales**

*AGGR* Aggression

*PSYC* Psychoticism

*DISC* Disconstraint

*NEGE* Negative Emotionality

*INTR* Introversion/Low Positive Emotionality

**206**

**Minnesota Multiphasic Personality Inventory—Adolescent 207**

**Table 7.2 Minnesota Multiphasic Personality Inventory-A (MMPI-A)**

Authors: Butcher, Williams, Graham, Archer, Tellegen, Ben-Porath, & Kaemmer

Published: 1992

Edition: 1st

Publisher: Pearson Assessments

Website: [www.PearsonAssessments.com](http://www.PearsonAssessments.com)

Age range: 14–18

Reading level: 6th grade

Administration formats: Paper/pencil, computer, CD, cassette

Additional languages: Spanish

Number of items: 478

Response format: True/False

Administration time: 45–60 minutes

Primary scales: 8 Validity, 10 Clinical, 15 Content

Additional scales: 11 Supplementary

Hand scoring: Templates

General texts: Archer (2005), Archer & Krishnamurthy (2002), Fowler et al. (2000)

Computer interpretation: Caldwell Report (Caldwell); Pearson Assessments (Butcher);

Psychological Assessment Resources (Archer)

higher than the original Minnesota normative sample of adults. On the rest of the validity and clinical scales, the ninth graders scored similarly to the adult sample. Thus, these adolescents were more likely to have elevations on Scales 4 (*Pd*: Psychopathic Deviate), 8 (*Sc*: Schizophrenia), and 9 (*Ma*: Hypomania), if adult MMPI norms were used. They also were less likely to have profiles in which all clinical scales were below a T score of 70 than were the normal adults.

Hathaway and Monachesi did not directly address whether these MMPI scale elevations in normal adolescents reflected some form of psychological distress and maladjustment that was characteristic of the turmoil of adolescence or whether these elevations reflected mere

differences in the frequency of item endorsement that may not have psychopathological implications. Hathaway and Monachesi seemed to favor the former interpretation because they did not advocate special adolescent norms with the MMPI. Instead, they suggested that both the standard adult-normed profile and an adolescent-normed profile should be plotted so that the clinician can understand the contrast between adolescents and adults. The clinician, however, when provided with the potentially divergent and contradictory information from two profiles on the same adolescent, needs to know which source of information is more accurate, and Hathaway and Monachesi (1963) did not answer this question.

The students who were retested with the MMPI in the 12th grade provided some interesting information on profile stability. Test-retest reliability coefficients were highest for the *K* (Correction) scale—.52 for males and .56 for females—and Scale 0 (*Si*: Social Introversion)—.54 for males and .61 for females. By contrast, Scales 4 (*Pd*: Psychopathic Deviate)—.36 for males and .38 for females—and 6 (*Pa*: Paranoia)—.32 for males and .36 for females—had the lowest reliability coefficients.

#### **208 Self-Report Inventories**

As might be expected, profile stability as defined by the single high-point scale was highest when that scale was a T score of 70 or higher. More than half of the adolescents, both male and female, with a T score of 70 or higher on Scale 4 (*Pd*: Psychopathic Deviate) on initial testing, had Scale 4 (*Pd*: Psychopathic Deviate) as one of the two highest scales when retested. This relationship, however, did not hold for all scales. Scale 5 (*Mf*: Masculinity-Femininity), for example, was likely to shift from a high-point scale to one of the three lowest clinical scales across this 3-year interval.

Marks, Seeman, and Haller (1974) described how the adolescent norms for the MMPI were derived, primarily crediting Peter F. Briggs. Briggs selected 100 boys and 100 girls aged 14, 15, and 16 years, and 80 boys and 40 girls aged 17 years from the students studied by Hathaway and Monachesi (1963). To obtain a larger and more nationally representative sample, Marks et al. collected 1,046 additional MMPI profiles in 1964 and 1965 from both rural and urban, public and private school students residing in Alabama, California, Kansas, Missouri, North Carolina, and Ohio. All these students were presumed to be white and at the time of testing were neither institutionalized nor being treated for emotional disturbance (Marks et al., 1974). By combining these two groups of students, the adolescent norms for the MMPI validity and clinical scales for boys and girls in four age groups (14 and below, 15, 16, and 17 and 18) were constructed (see appendix H, tables 6 to 9, Dahlstrom, Welsh, & Dahlstrom, 1972). Although Marks et al. (1974) advocated that emotional disturbance in adolescents needs to be evaluated against adolescent norms, they concurred with Hathaway and Monachesi (1963) that adolescent scores on the MMPI also should be compared with adult norms. Thus, Marks et al. recommended that both an adult- and an adolescent-normed profile should be constructed for adolescents.

Several issues arise when the MMPI is used with adolescents. The first issue is whether adult norms, adolescent norms, or both sets of norms should be used with adolescents. As noted, Hathaway and Monachesi (1963) and Marks et al. (1974) recommended that adult and adolescent-normed profiles should be constructed for adolescents. However, Archer (1984, 1987) advocated that adolescent norms should be used exclusively with adolescents, and he presented a cogent rationale for their use that clinicians should read. Williams (1986) also indicated that adolescent norms were the most appropriate for adolescents, but she suggested that both profiles should be plotted for adolescents.

A final issue is whether the adolescent or the adult correlates provide a more accurate interpretation once adolescent norms serve as the base to derive a codetype. Using adolescent and adult norms, Lachar, Klinge, and Grisell (1976) obtained valid profiles on 100 adolescents, most of whom were hospitalized. Interpretations of the two profiles for each adolescent were generated using Lachar's (1974) automated interpretive system for adults. Clinicians then were asked to rate these interpretations for accuracy. The interpretations of profiles generated from adolescent norms, compared with interpretations of profiles generated from adult norms, were rated more accurate 61 times, as accurate 13 times, and less accurate 26 times. Only 10% of the adolescent norm interpretations were judged to

be inaccurate, and 20% of the adult norm interpretations were judged to be inaccurate. Wimbish (1984) reported similar findings in a sample of inpatient adolescent substance abusers.

Hence, it appears that even when an interpretive system (Lachar's) based on adults is used, profiles based on adolescent norms provide a more accurate description of adolescents **Minnesota Multiphasic Personality Inventory—Adolescent 209**

than do profiles based on adult norms. It remains to be seen, however, whether other adult interpretive systems and individual clinicians will demonstrate the same improvement in accuracy with the use of adolescent norms.

Once the decision had been made to use adolescent norms for the MMPI, the clinician had three different sets of norms from which to choose. Marks et al.'s (1974) norms, described previously, have served as the standard for adolescent MMPIs since they were first published. Gottesman, Hanson, Kroeker, and Briggs (1987) developed their norms by using the MMPI responses of 12,953 15-year-olds and 3,492 18-year-olds, who composed the entire sample of adolescents studied by Hathaway and Monachesi (1963). Colligan and Offord (1989) developed their norms by collecting a sample of adolescents in the Mayo Clinic catchment area. This issue of which set of norms to use with the MMPI has been resolved with the development of the MMPI-A (Butcher et al., 1992) and the decision of the University of Minnesota Press to discontinue publishing the original MMPI in the fall of 1997.

The MMPI-A (Butcher et al., 1992) represents the restandardization of the MMPI to create a version that is specifically adapted for adolescents. Restandardization of the MMPI was needed to provide item content that was appropriate for adolescents, develop scales to assess symptoms and behaviors more common to adolescents, and create current norms for the inventory. Continuity between the MMPI and the MMPI-A was maintained as much as possible, although it was recognized that some changes would be necessary to adapt the MMPI for adolescents. Thus, the items on the validity and clinical scales of the MMPI are basically unchanged on the MMPI-A except for the *F* (Infrequency) scale and Scale 5 (*Mf*: Masculinity-Femininity), which lost 27 and 16 items, respectively. The MMPI-A still retains 12 of the 27 items that were dropped from the *F* scale.

In the development of the MMPI-A, the Restandardization Committee (Butcher et al., 1992) started with the 550 items on the original MMPI (they first deleted the 16 repeated items). They reworded 82 of these 550 items to eliminate outdated and sexist language and to make these items more easily understood. Williams, Ben-Porath, and Hevern (1994) determined that the psychometric properties of 20 of these reworded items were the same or better than the original items. They also included 58 items from Form AX that was used to collect the normative data for the MMPI-2. The Restandardization Committee then added 96 provisional items "to assess problems, behaviors, and attitudes of adolescents in areas related to identity formation, negative peer-group influence, school and teachers, relationships with parents and families, and sexuality" (Butcher et al., 1992, p. 5). Thus, Form TX had a total of 704 items and was used to collect the normative data for the MMPI-A. When finalizing the items to be included on the MMPI-A, the Restandardization Committee deleted 121 items from the original MMPI in addition to the 54 items deleted from the standard validity and clinical scales and the 16 repeated items. The rationale for including and dropping items from Form TX that resulted in the 478 items on the MMPI-A has not been made explicit to date.

The MMPI-A was standardized on a sample of 1,620 adolescents who resided in eight different states (California, Minnesota, New York, North Carolina, Ohio, Pennsylvania, Virginia, and Washington). These adolescents "were derived from samples of generally limited representativeness in terms of ethnicity, geographical region, and rural-urban residence" (Butcher et al., 1992, p. 10). Similar to the MMPI-2, the MMPI-A normative sample

#### **210 Self-Report Inventories**

is predominantly white and the adolescents' parents tended to be well educated and to be in professional and managerial occupations.

The MMPI-A is intended to be used with adolescents in the age range of 14 to 18 years. The Restandardization Committee thought that the validity of the data for 13-year-olds

was too questionable to warrant inclusion in the normative sample. Bright, mature 12- and 13-year-olds might be able to take the MMPI-A appropriately, but research is needed to document this position. The Restandardization Committee also suggested that the MMPI-A be used for 18-year-olds who are in high school and the MMPI-2 be used for those who are in college and living an independent lifestyle. Shavel and Archer (1996) have suggested that a conservative approach would be to plot both the MMPI-A profile and the MMPI-2 profile for 18-year-olds when the clinician cannot decide which form of the test is more appropriate. The data on age effects on the MMPI-2 (see Chapter 6, pp. 198–199) would suggest that all 18-year-olds should be given the MMPI-A, but research is needed to answer this question definitively.

### **ADMINISTRATION**

Special efforts must be made to provide adequate instructions and monitoring for all adolescents, because of their age and developmental stage (cf. Archer, 2005, pp. 50–61), and are extremely important for younger adolescents (<15). The adolescent must be able to understand why the MMPI-A is being administered and how the results will be used in order to assent to the assessment process. The clinician should work diligently to make the assessment process a collaborative activity with the adolescent to obtain the desired information. This issue of therapeutic assessment was covered in more depth in Chapter 2 (pp. 43–44).

Completing the 478 items may seem like a nearly impossible task for younger adolescents and any adolescent with language or attention deficits. Before they start, these adolescents need to be aware that there will be frequent breaks any time they become necessary. A number of shorter sessions that are planned beforehand are more likely to produce valid results than relying on these adolescents to recognize when breaks are necessary. Pena, Megargee, and Brody (1996) reported by splitting the MMPI-A administration into two sessions 94% of their sample of delinquent boys produced a valid profile.

Supervision and monitoring of adolescents taking the MMPI-A is important in a group setting. Some adolescents may respond out loud to the MMPI-A items and provoke responses from others that, if not dealt with quickly, have the potential to not only disrupt the session, but render the entire assessment process questionable. The initial instructions should make it explicit that they are to raise their hand if they have a question and the proctor will come to them. Under no circumstances are they to ask out loud about any of the items. The proctors also need to know what types of assistance are allowed and never indicate what the response to any item should be. Butcher et al. (1992, p. 29) suggest that it usually is sufficient to say, “Just answer as you think best.”

Although the proctors need to monitor actively the adolescents’ performance in a group setting, the proctors should not linger near any of the adolescents to prevent them from thinking that their specific responses are of interest. Any type of conversation among the proctors also should be minimized so that the adolescents do not get the impression that their **Minnesota Multiphasic Personality Inventory—Adolescent 211** responses are being discussed. The proctors need to make sure that adolescents’ responses on the answer sheet correspond to their place in the test booklet as nonobtrusively as possible.

Reading level is a crucial factor in determining whether the adolescent can complete the MMPI-A, and if possible, reading level should be checked before the MMPI-A is administered rather than relying on the inconsistency scales to identify these problems. The adolescent should have at least a sixth-grade reading level to take the MMPI-A (Butcher et al., 1992). Krakauer, Archer, and Gordon (1993) have developed the *Items–Easy* and *Items–Difficult* scales for the MMPI-A to identify adolescents who are having reading difficulties. The clinician must check measures of consistency (Baer, Ballenger, Berry, & Wetter, 1997) and accuracy of item endorsement carefully with adolescents because of the higher probability that they will be noncooperative or noncompliant with the assessment process, in addition to the potential problems of reading comprehension.

### **SCORING**

Scoring the MMPI-A is identical to scoring the MMPI-2 (see pp. 141–143), with one major exception, and will not be repeated here. The primary major difference in scoring

the MMPI-A compared with the MMPI-2 is that the five clinical scales (1 [*HS*: Hypochondriasis], 4 [*PD*: Psychopathic Deviate], 7 [*PT*: Psychasthenia], 8 [*SC*: Schizophrenia], and 9 [*MA*: Hypomania]), that are *K*-corrected on the MMPI-2, are *not K*-corrected on the MMPI-A. Consequently, a similar procedure is followed to score and plot all the validity, clinical, content, and supplementary scales on the MMPI-A.

### **ASSESSING VALIDITY**

Figure 7.1 provides the flowchart for assessing the validity of this specific administration of the MMPI-A, and the criteria for using this flowchart are provided in Table 7.3. The clinician is reminded that the criteria provided in Table 7.3 are continuous, yet ultimately the decisions that must be made in implementation of the flowchart in Figure 7.1 are dichotomous. General guidelines are provided for translating these continuous data into dichotomous decisions on the MMPI-A, but these guidelines need to be considered within the constraints of this specific adolescent and the circumstances for the evaluation.

#### **Item Omissions**

The Cannot Say scale (?) consists of the total number of items that the adolescent fails to answer or answers both “true” and “false.” The raw score on the Cannot Say scale (?) is located at the bottom left-hand corner of the profile form and can be easily overlooked, particularly if the clinician is not the person who actually scored the MMPI-A. These omitted item(s) should be reviewed to see if there is any theme to them. The individual should be queried carefully about any omitted items among the dangerousness-to-self items (177, 283) and dangerousness-to-others items (445, 453, 458, 465), and the explanation documented.

#### **212 Self-Report Inventories**

Therapeutic  
Assessment  
Omissions  
Cannot Say (?)  
Consistency of  
Item Endorsement  
VRIN  
Favorable  
Self-Description  
Accuracy of  
Item Endorsement  
Unfavorable  
Self-Description  
L K Fp-A F  
Codetype  
Interpretation

#### **Figure 7.1 Flowchart for assessing validity of the MMPI-A.**

Butcher et al. (1992) and Archer (2005) state that if 30 or more items are omitted and they are evenly distributed throughout the MMPI-A, it should not be scored. Clinicians may want to ask the adolescent to attempt to answer any omitted items if there are more than 10 omitted items, after determining that reading difficulties are not the cause of the omissions. If most of the omitted items occur after item 350, the basic validity (*F1*, *L*, and *K*) and the 10 clinical scales can be scored.

Archer (2005, p. 83) reported the items most frequently omitted by the MMPI-A normative sample: 203 (2.9%); 199 (1.5%); 93 (1.2%); and 16 (1.0%). All the other items on the MMPI-A were omitted by less than 1% of the normative group. These very low rates of item omissions suggest that the restandardization process was successful at rendering the items appropriate for adolescents.

Few items are generally omitted on the MMPI-A with nearly 70% of normal adolescents and 50% of clinical samples omitting no items (Table 7.3). Omitting 10 items is around the 98th percentile for the MMPI-A normative sample (Butcher et al., 1992) and the 93rd percentile in a clinical sample (Archer, Handel, & Mason, 2006). The setting in which the MMPI-A is administered does not appreciably affect the number of items that are omitted.

#### **Consistency of Item Endorsement**

The clinician should carefully note the time that the adolescent started and completed the

MMPI-A. Unusually short administration times, such as less than 40 minutes to complete all 478 items, will alert the clinician to potential problems in consistency of item endorsement.

**Minnesota Multiphasic Personality Inventory—Adolescent 213**

**Table 7.3 Criteria for assessing MMPI-A validity by type of setting**

Percentile

Setting 1 2 7 16 30 50 70 84 93 98 99

**Item Omissions (Cannot Say [?])**

Normal<sup>a</sup> — — — — — — — 0 1 3 7 11

Clinical<sup>b</sup> — — — — — — 0 1 4 9 27 83

**Consistency of Item Endorsement (Variable Response Inconsistency Scale [VRIN])**

Normal<sup>a</sup> — — — 0 1 2 3 5 7 9 12 13

Clinical<sup>b</sup> — 0 1 2 3 4 6 9 12 15 17

**Accuracy of Item Endorsement**

**Self-Unfavorable (Infrequency Psychopathology Scale [Fp-A])**

Normal<sup>a</sup> — — — — 0 1 2 5 9 13 16 17

Clinical<sup>b</sup> — — — 0 1 2 4 7 12 17 21 22

**Accuracy of Item Endorsement**

**Self-Unfavorable (Infrequency Scale [F])**

Normal<sup>a</sup> — — — — 0 1 2 5 11 17 23 28 30

Clinical<sup>b</sup> 0 1 3 4 6 10 16 24 30 34 40

**Self-Favorable Impression Management (Lie Scale [L])**

Normal<sup>a</sup> — — — — 0 1 2 3 4 6 8 9

Clinical<sup>b</sup> — 0 1 2 3 4 6 7 9 11 12

**Self-Deception (Correction Scale [K])**

Normal<sup>a</sup> 2 3 5 7 9 11 14 16 19 22 23

Clinical<sup>b</sup> 4 5 7 9 11 14 16 19 22 25 27

<sup>a</sup>Butcher et al. (1992).

<sup>b</sup>Archer et al. (2006).

This information is easily overlooked, however, unless the clinician routinely notes the administration time.

Consistency of item endorsement on the MMPI-A is assessed by the Variable Response Inconsistency (*VRIN*) and True Response Inconsistency (*TRIN*) scales in an identical manner as on the MMPI-2 and the Infrequency scales *F1* and *F2*. The Variable Response Inconsistency (*VRIN*) scale consists of 50 pairs of items that have similar or opposite item content. These pairs of items are scored if the adolescent is inconsistent in his or her responses. The *VRIN* scale actually consists of 42 pairs of unique items, since two separate response patterns are scored for 8 of these 50 item pairs as with items 6 and 86.

The True Response Inconsistency (*TRIN*) scale consists of 24 pairs of items. The *TRIN* scale is very similar to the *VRIN* scale except that the scored response on the *TRIN* scale is either “true” or “false” to both items in each pair. The *TRIN* scale has 15 pairs of items to which the inconsistent response is “true” and 9 item pairs to which the inconsistent response is “false.” Scoring the *TRIN* scale is somewhat complicated. One point is added to the client’s score for each of the 15 item pairs that are scored if endorsed “true,” whereas one point is subtracted for each of the 9 item pairs that are scored if endorsed “false.” Then

**214 Self-Report Inventories**

9 points are added to this score. (Nine points are added to the score so that it is not possible to obtain a negative score on the *TRIN* scale. If a client endorsed none of the 15 “true” item pairs and all 9 of the “false” item pairs, a score of –9 would be obtained. Adding 9 points avoids this problem.) The *TRIN* scale is intended to identify adolescents who are endorsing the items inconsistently by essentially responding to most of the items as “true” or “false.” The *F* (Infrequency) scale on the MMPI-A has been divided into two 33-item subscales.

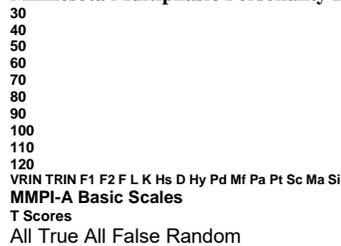
The items on the *F1* scale are found in the first half of the test and the items on the *F2* scale are found in the second half of the test. Butcher et al. (1992) suggested that the *T* scores on the *F1* and *F2* scales can be compared to assess whether the adolescent has endorsed more or less of these infrequently endorsed items on the second half of the test. An adolescent who has an elevated *T* score on the *F2* scale compared with the *F1* scale may have lost interest in the assessment process and began to respond randomly to the items somewhere in the latter half of the MMPI-A. However, Archer, Handel, Lynch, and Elkins (2002) found that the difference between the *T* scores on the *F1* and *F2* scales was ineffective at identifying

profiles with varying levels of random responding. Consequently, clinicians should be cautious about putting too much emphasis on interpreting the relationship between the T scores on the *F1* and *F2* scales as a possible indicator of inconsistent item endorsement in the latter half of the MMPI-A until additional research validates its use. Archer et al. also noted that none of the measures of consistency of item endorsement was very effective when partial (<50%) random responding occurred in the last half of the test. Pincus (2005) reported similar difficulties in the identification of partially random MMPI-A profiles. He did suggest a number of subscales of VRIN (Variable Response Inconsistency) and algorithms based on the relationships among the family of *F* (Infrequency) scales to identify these partially random profiles.

Scores on the *VRIN* scale are very similar in normal and clinical samples (see Table 7.3), which substantiates the point that psychopathology per se does not produce inconsistent item endorsement. Also, a fair amount of inconsistency (3 to 4 inconsistent pairs of items) is found in both the clinical and normal samples. Again, a specific score on *VRIN* has not been selected to indicate that the items were endorsed inconsistently. Raw scores in the range of 12 to 15 have been suggested as a cutting score by the authors of MMPI-A interpretive manuals (Archer, 2005 [girls >12; boys >14]; Butcher et al., 1992 [girls >11; boys >13]). The clinician will have to decide which cutting score is most appropriate given the specific adolescent and the setting in which the MMPI-A is administered.

Figure 7.2 shows the profiles for random, “all-true,” and “all-false” responses on the MMPI-A. The validity scales easily identify the “all-true,” and “all-false” profiles as being invalid with several scales at T scores greater than 100 or below 40. Scales 1 (*Hs*: Hypochondriasis), 2 (*D*: Depression), and 3 (*Hy*: Hysteria) when compared with Scales 6 (*Pa*: Paranoia), 7 (*Pt*: Psychasthenia), 8 (*Sc*: Schizophrenia), and 9 (*Ma*: Hypomania) also are elevated in very divergent manners in these two profiles. The validity scales do not identify a random profile very well with T scores around 80 on *VRIN* (Variable Response Inconsistency) and the family of *F* (Infrequency) scales. Validity scores in this T score range can be seen in adolescents who are accurately reporting relatively severe psychopathology. The clinical scales for the random profile also look interpretable and do not assist the clinician in the identification of a random profile. In short, it is very difficult to identify a randomly generated MMPI-A profile, which underscores the importance of both obtaining the cooperation of adolescents before taking the MMPI-A and monitoring their

**Minnesota Multiphasic Personality Inventory—Adolescent 215**



**Figure 7.2 Random, all true, and all false MMPI-A profiles.**

performance during the test. Figure 7.2 can be compared with Figure 6.3, which provides similar data on the MMPI-2, to note some of the differences between the MMPI-A and MMPI-2.

**Accuracy of Item Endorsement**

*Self-Unfavorable Descriptions*

Self-unfavorable descriptions of psychopathology on the MMPI-A can be assessed by infrequency scales [*F*, *F1*, *F2*, and *Fp - A*] and the *F - K* Dissimulation index (Gough, 1950). Only the *F* (Infrequency) and Infrequency-Psychopathology (*Fp - A*) scale (McGrath et al., 2000) will be described here because they are the two best scales to identify self-unfavorable descriptions. The *Fp - A* scale consists of 40 items that no more than 20% of a sample of 475 inpatient adolescents, a second sample of 356 inpatients adolescents from the same facility, 140 high school students faking the MMPI-A (Stein, Graham, & Williams, 1995), and the MMPI-A normative sample endorsed in the deviant direction. Nearly three-quarters (31) of the items on the *Fp - A* scale also are found on the *F* scale. Given the redundancy

between these two scales, the incremental validity for the *Fp - A* scale is relatively small. McGrath et al.'s data also suggested that the MMPI-A *F* scale may be a better indicator of a self-unfavorable description than the MMPI-2 *F* scale.

Table 7.3 shows that endorsing a number of the *F* and *Fp - A* items is typical in both normal and clinical samples. The adolescent must endorse 13 to 17 of the *Fp - A* items (32% to 42%) and 23 to 30 of the *F* items (35% to 45%) to reach the 93rd percentile in both samples. Adolescents are much more likely to endorse these infrequency items on the MMPI-A than adults endorse the infrequency items on the MMPI-2 (see Table 6.3).

#### *Self-Favorable Description*

Self-favorable descriptions of psychopathology will be organized within Paulhus's (1984, 1986) model of social desirability responding that distinguishes between self-deception and impression management: *self-deception* refers to a motivated unawareness of one of two conflicting cognitions, whereas *impression management* can be conceptualized as a **216 Self-Report Inventories**

strategic simulation, a motive, or a skill. In self-deception, individuals are thought to believe their positive self-evaluation is an accurate description of themselves, whereas in impression management, individuals consciously dissemble to create a favorable impression in others. The *L* (Lie) scale on the MMPI-A is a measure of impression management, whereas the *K* (Correction) scale is a measure of self-deception.

The *L* (Lie) scale includes 15 items that were selected on a rational basis to identify persons who are deliberately trying to avoid answering the MMPI frankly and honestly (Dahlstrom et al., 1972). The scale assesses attitudes and practices that are culturally laudable but actually found only in the most conscientious persons. The content areas within the *L* scale include denial of minor, personal dishonesties and denial of aggression, bad thoughts, and weakness of character. The *K* (Correction) scale consists of 30 items that were empirically selected to assist in identifying persons who displayed significant psychopathology yet had profiles within the normal range.

Table 7.3 shows that endorsing more than 6 to 9 of the items on the *L* (Lie) scale and 19 to 22 items on the *K* (Correction) scale is at or above the 93rd percentile in normal and clinical samples. These ranges on the MMPI-A *L* (Lie) and *K* (Correction) scales are similar to the ranges found on the MMPI-2 (see Table 6.3). The similarity in the scores on these self-favorable scales between the MMPI-A and MMPI-2 is in sharp contrast to the differences that were found on the self-unfavorable scales.

The person can be described as using impression management to create a favorable self-description when the *L* (Lie) scale is elevated ( $T > 59$ ) and the *L* (Lie) scale is at least 5 T points higher than the *K* (Correction) scale. The person can be described as using self-deception to create a favorable self-description when the *K* (Correction) is elevated ( $T > 59$ ) and at least 5 T points higher than the *L* (Lie) scale. When both of these scales (*L* [Lie], *K* [Correction]) are elevated in the same range ( $T > 55$ ), the person can be described as making a generic self-favorable description.

When an adolescent's responses have been identified as being endorsed inaccurately because of an extremely self-favorable description of psychopathology, the MMPI-A profile is no longer interpretable. The clinician will have little reason to try to interpret such a profile, however, because extremely self-favorable descriptions result in no clinical scales being elevated over a T score of 60. The clinician should describe the adolescent's style of self-favorable description of psychopathology (impression management, self-deception, or a combination of both), determine the potential causes for this self-favorable description, and assess the implications for treatment and intervention.

Adolescents who provide extremely self-favorable descriptions of psychopathology see their problems as less troubling to themselves and, hence, are less motivated to change. Their problems also may be more chronic in nature, and consequently, they may be more difficult to treat if they remain in treatment. None of these potential causes of an extremely self-favorable description of psychopathology is a good prognostic sign for any type of psychological intervention.

## **INTERPRETATION**

The clinician whose primary focus has been on the MMPI-2 must be particularly careful

when interpreting the MMPI-A for the first time because of the numerous subtle and not Minnesota Multiphasic Personality Inventory—Adolescent 217

30  
40  
50  
60  
70  
80  
90  
100  
110  
120

#### MMPI-A Basic Scales

##### T Scores

VRIN TRIN F1 F2 F L K Hs D Hy Pd Mf Pa Pt Sc Ma Si

#### Figure 7.3 MMPI-A basic validity and clinical scales for a 1-3/3-1 codetype.

so subtle differences between the two tests. The interpretive guidelines for the MMPI-2 simply *cannot* be applied directly to the MMPI-A. A prime example of this difference in interpretive strategy on the MMPI-A is the use of *marginally* elevated T score values between 60 and 64 in addition to the more usual *clinically* elevated T scores of 65 and higher (see Figure 7.3). The T scores in this marginally elevated range are indicated by the shaded area on the MMPI-A profile sheets. Probably the major difference between the MMPI-A and the MMPI-2 is that the MMPI-A is being interpreted for an adolescent who is still growing and maturing rather than for an adult. Generally, MMPI-A profiles will be characterized by less chronicity and severity of psychopathology than the same MMPI-2 profile simply because of the relative youth of adolescents. Any inferences drawn from the MMPI-A must reflect that adolescents are in a state of developmental transition to adulthood.

#### Codetypes

Once it has been determined that the adolescent has omitted few, if any, items, and endorsed the items consistently and accurately, the MMPI-A can be interpreted. Interpretation of the MMPI-A begins with, and is based on codetypes; that is, the two highest clinical scales elevated to a T score of 65 or higher, or the single clinical scale (spike codetype) when only one scale reaches a T score of 65. In contrast to the MMPI-2, marginally elevated T scores in the range of 60 to 64 on the MMPI-A are considered to be in a transition range between normal and elevated profiles. (This range is indicated by the gray or shaded background on the profile form for the basic validity and clinical scales). Scores in this marginally elevated range are expected to reflect some, but not all, of the behaviors reflected in elevated scores. The interpretation of the MMPI-A codetype then is supplemented by the examination of specific scales, such as the clinical, content, and supplementary scales, as well as individual critical items. Archer (2005), Archer and Krishnamurthy (2002), and Fowler, Butcher, and Williams (2000) have provided examples of the general interpretive strategy for the MMPI-A.

The following pages provide the interpretation of 15 commonly occurring two-point MMPI-A codetypes, which have been adapted from the MMPI-2 Adult Interpretive System 218 Self-Report Inventories

(Greene & Brown, 2006) based on the work of Marks et al. (1974) and Archer (2005).

The 60 new items on the MMPI-A that are not found on the MMPI-2 were not available to be included in these descriptions so the clinician will need to add these items to the description where appropriate. Most of these 60 items are found on the MMPI-A content and supplementary scales that have been included in these codetype descriptions. Figure 7.3 illustrates the basic profile for the MMPI-A validity and clinical scales for a 1-3/3-1 codetype. A codetype description is written about a *group* of adolescents who have elevated these two specific scales, not about this specific adolescent who has produced the codetype. Each statement within the codetype description must be considered a hypothesis to be verified with the clinical interview and history of this specific adolescent.

The clinician must expect to be making some modification to every codetype description on the MMPI-A. This entire process has been made much more straightforward in these descriptions because every statement is referenced either to a scale(s) and/or item(s). When

a scale name is in **bold** font, the individual is expected to have a low score ( $T < 50$ ) on that scale. In the description of Interpersonal Relations for a *1-3/3-1* codetype (p. 219), the ***Pdi*** and ***A-fam*** scales are in bold font because these two scales assess problems and conflicts in family relations that this group of adolescents does *not* report. When an item is listed in **bold** font, the item is endorsed in the **false** direction. Numerous examples can be seen in the Other Problem Areas for a *1-3/3-1* codetype (p. 219).

The clinician should modify any statement for which there are explicit data to support the change. If the codetype description indicates that the adolescent has endorsed an item as being “true” and it actually was endorsed “false,” the clinician has explicit data to change/modify/delete that item. Similarly, if the interpretive text indicates that a scale is/is not elevated, the adolescent’s actual T score on the scale will support/negate that statement explicitly. The clinician, however, should *not* change any statement simply because it does not seem to fit this individual or the clinician does not think it is appropriate. It is vital to remember Meehl’s (1957) admonition of how infrequently we are to use our clinical judgment when we have empirical data available to us.

*1-3/3-1<sub>1</sub>*

**Moods** She reports that she is experiencing very mild emotional distress (*A, NEGE*) characterized by tension and anxiety (*7[Pt], A-anx*). She frequently worries about something (185T). She also experiences a very mild level of dysphoria (*2[D], A-dep*). Her daily life is full of things that keep her interested (9T). Something exciting will almost always make her feel better when she is feeling low (228T).

**Cognitions** She reports that she cannot keep her mind on one thing (279T). She is selfconfident (**70F**). She likes to let people know where she stands on things (341T), and she finds it necessary to stand up for what she thinks is right (115T). She tends to think in a concrete manner and to focus on her physical ailments (*1[Hs]*).

She believes that most people are honest and can be trusted (*3[Hy], A-cyn*). She does not analyze the motives for her own or others’ behavior (*3[Hy]*).

<sup>1</sup>The gender used for the interpretive text for each codetype reflects whether it is more common in boys or girls.

**Minnesota Multiphasic Personality Inventory—Adolescent 219**

**Interpersonal Relations** She reports a balance between extroverted and introverted behaviors (*0[Si], A-sod*). She is very sociable and makes friends quickly (46T, 262T, 336T).

She believes, if given the chance, she would make a good leader of people (329T). She seems to make friends about as quickly as others do (262T). She is very conventional and conforming to social standards (*3[Hy]*). Her relations with her family are good (***Pdi, A-fam***).

**Other Problem Areas** She reports a variety of physical (*1[Hs], A-hea, Hy4*) and neurological symptoms (*Sc6, A-hea2*), headaches (**168F**), and pain in a number of areas of her body (**54F, 210F**). Her physical health is not as good as that of most of her friends (**42F**), and she worries about her health (**4F**). She tires quickly (**146F**). She has difficulty going to sleep because she is bothered by ideas or thoughts (**134F**), her sleep is fitful and disturbed (36T), and she does not wake up fresh and rested most mornings (**3F**). She is less able to work now than she was in the past (**10F**). Her history and background should be reviewed to determine whether a medical or neurological evaluation is warranted. She is unlikely to have behavioral problems (***A-con, A-con2***), have been in trouble with the law (249T), or use alcohol or drugs (***ACK, PRO***).

**Treatment** She often is referred for treatment because of problems or concerns in academic settings (*1[Hs], 3[Hy]*) that reflect attention-seeking behavior or somatic symptoms.

She is able to talk at least superficially about her psychological problems and is not evasive in treatment (*1[Hs], 3[Hy]*). Conservative medical treatment is recommended because her physical ailments are difficult to document (*1[Hs], 3[Hy]*). Short-term interventions that are focused on providing symptomatic relief from her physical ailments may be beneficial and can provide the foundation for more traditional psychotherapy (*1[Hs], 3[Hy]*).

She is not naturally introspective, which will complicate the implementation of any therapeutic intervention (*3[Hy]*).

*2-3/3-2*

**Moods** She reports that she is experiencing a mild to moderate level of emotional distress

(A, *NEGE*) characterized by dysphoria (2[D], *A-dep*), worrying (7[Pt], *A-anx*), and anhedonia (*INTR*, *A-dep1*, *A-trt1*). Most of the time she feels blue (62T, **91F**, **360F**), and her daily life is not full of things that keep her interested (**9F**). She frequently worries about something or someone (185T, 281T, 285T, 383T). She is easily hurt by criticism or scolding (121T) and has difficulty expressing her feelings, describing herself as feeling bottled up (2[D], 3[Hy]). She has had long periods of time when she could not take care of things because she could not get going (35T). She is overcontrolled and fearful of losing control (2[D], 3[Hy]). She is likely to experience increases in depression, fatigue, and physical symptoms in response to stress (3[Hy]). She is unlikely to express her anger overtly (**AGGR**, *A-ang*).

**Cognitions** She reports that she has problems with attention, concentration (28T, 279T, 305T), and memory (**158F**). She likes to let people know where she stands on things (341T), and she finds it necessary to stand up for what she thinks is right (115T). She has low self-esteem, lacks self-confidence (70T, **223F**), and doubts her own abilities (*A-lse*).

#### **220 Self-Report Inventories**

She feels inadequate, helpless, and insecure (*A-lse*). She certainly thinks she is useless at times (124T).

She sometimes thinks that she is about to go to pieces (404T). At times, her mind seems to work more slowly than usual (320T). Her judgment is not as good as it was in the past (**40F**).

She does not analyze the motives for her own or others' behavior (3[Hy], **A-cyn**). She wishes she could be as happy as others seem to be (53T).

**Interpersonal Relations** She reports that she is somewhat introverted (0[Si], *A-sod*, *A-sod1*) and does not like loud parties or social events (**82F**). Her activities are usually engaged in alone (2[D], 3[Hy]). She feels socially alienated (*Pd4*, *Sc1*, *Si2*) and presents herself as helpless, immature, and dependent (*A-lse*). She reports good relations with her family (*Pd1*, **A-fam**).

**Other Problem Areas** She reports a wide variety of physical symptoms (*I[Hs]*, *A-hea*, *Hy4*) including neurological symptoms (*Sc6*, *A-hea2*), headaches (**168F**), and generalized pain (**210F**). Her physical health is not as good as that of most of her friends (**42F**) and she worries about her health (**4F**). During the past few years, she has not been well most of the time (**135F**). She has difficulty going to sleep because thoughts or ideas are bothering her (**134F**), her sleep is fitful and disturbed (36T), and she does not wake up fresh and rested most mornings (**3F**). She is easily awakened by noise (5T). She tires quickly (**146F**) and feels tired, weak, and without energy a good deal of the time (167T, **289F**). She is not as able to work as she once was (**10F**). Her history and background should be reviewed to determine whether a medical or neurological evaluation is warranted. She is a very conventional individual who is unlikely to have behavioral problems (**A-con**). She is not likely to abuse substances (**ACK**, **PRO**).

**Treatment** Her prognosis is generally poor because she sees little chance for significant change in her life (2[D], 3[Hy]). She is referred for treatment because of poor relationships with peers; at school, she is liked by few students and outside school, she has few friends of any kind (2[D], 3[Hy]). Short-term, behavioral therapy that focuses on her reasons for entering treatment may be beneficial and may allow for the development of a therapeutic alliance that would be necessary for long-term therapy (2[D], 3[Hy]). She will prefer to discuss her physical symptoms rather than focus on her psychological processes (*I[Hs]*). She is not naturally introspective, which will complicate the therapeutic process (3[Hy]).

2-4/4-2

**Moods** He reports that he is experiencing a mild to moderate level of emotional distress (A, *NEGE*) characterized by dysphoria (2[D], *A-dep2*, 62T, **91F**, **360F**), resentment (6[Pa]), agitation (7[Pt], *A-anx*), and anhedonia (*INTR*, *A-dep1*, *A-trt1*). His daily life has few things that keep him interested (**9F**). He frequently worries about something or someone (185T, 281T, 285T, 383T). These moods often arise in response to some external problem or difficulty (4[*Pd*]). While he may express guilt and remorse and may promise to change his behavior, his expressions typically are not sincere (4[*Pd*]). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (**60F**, 121T), and

#### Minnesota Multiphasic Personality Inventory—Adolescent 221

he is inclined to take things hard (311T). He is grouchy, irritable, and stubborn (*A-ang2*, 111T, 388T, 416T).

**Cognitions** He reports that his memory seems to be all right (158T). His judgment is not as good as it was in the past (**40F**). He certainly is lacking in self-confidence and believes that he is not as good as other people (*A-lse1*, 70T, **223F**, 379T). He certainly thinks he is useless at times (124T). He has sometimes thought that difficulties were piling up so high that he could not overcome them (318T). He is apt to take disappointments so keenly that he cannot put them out of his mind (377T).

He sometimes thinks that he is about to go to pieces (404T).

He doubts the motives of other people and believes that they are interested only in their own welfare (*A-cyn*, 77T, 107T). He knows who is responsible for most of his troubles (109T). He regrets many things that he has done (78T) and thinks that he has not lived the right kind of life (49T).

**Interpersonal Relations** He reports that he is introverted (*0[Si]*, *A-sod1*). His interpersonal relations tend to be generally problematic and conflicted (*4[Pd]*, *A-fam*), and he is alienated from himself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*). His home life is not as pleasant as that of most people he knows (**119F**). He quarrels with members of his family (*A-fam1*, **79F**) and feels alienated from them (*A-fam2*). He has no one in his family with whom to discuss his personal problems (*2[D]*, *4[Pd]*).

**Other Problem Areas** He reports that he worries about his health (**4F**). His sleep is bothered by his thoughts and ideas (**134F**), and he does not wake up fresh and rested most mornings (**3F**). He tires quickly (**146F**). He does not feel weak all over much of the time (**167F**). He reported a number of conduct problems (*A-con2*). He is likely to abuse alcohol or drugs with the exception of narcotics (*ACK*, *PRO*, 247T); a careful review should be made of the consequences of his alcohol and drug use.

**Treatment** His prognosis is generally poor for traditional methods of individual psychotherapy (*4[Pd]*) unless a mood disorder is the primary diagnosis (*2[D]*, *A-dep*). In that instance, cognitive-behavioral psychotherapy focused on his depressive cognitions will be beneficial (*2[D]*, *A-dep*). Evaluation for antidepressant medication may be indicated in cases of more severe depression (*2[D]*, *A-dep*). He frequently is referred for treatment because he has difficulty concentrating (*2[D]*, *4[Pd]*). Group therapy with individuals with similar behaviors may be effective when a mood disorder is not the primary diagnosis (*4[Pd]*). The consequences of his use of alcohol and drugs cannot be overlooked when determining the interventions that need to be made (*2[D]*, *4[Pd]*).

Several specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: He has difficulty starting to do things (218T); he gives up quickly when things go wrong (340T) or because he thinks too little of his ability (306T); he is hard to get to know (408T); and he is very stubborn (416T).

2-7/7-2

**Moods** He reports that he is experiencing a moderate level of emotional distress (*A, NEGE*) characterized by dysphoric mood (*2[D]*, *A-dep2*, 62T, **91F**, **360F**), guilt, and anxiety (*7[Pt]*, **222 Self-Report Inventories**

*A-anx*). He views himself as being irritable and grouchy (*A-ang2*, 111T, 388T), although others are more aware of his dysphoric mood (*2[D]*, *A-dep2*). He experiences little pleasure from life (*INTR*, *A-dep1*, *A-trt1*) and may be anhedonic. He has had long periods when he could not take care of things because he could not get going (35T). He is a chronic worrier (185T) who broods and ruminates about himself and his problems (*7[Pt]*, *A-anx*, *A-obs*, 281T, 285T, 383T). He is likely to overreact to minor stress with agitation, guilt, and self-punishment (*7[Pt]*). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (**60F**, 121T) and he is inclined to take things hard (311T). He easily becomes impatient with people (282T). He is unlikely to express anger overtly or to be aggressive toward others (**AGGR**).

**Cognitions** He reports that he has problems with attention, concentration (28T, 279T, 305T), and memory (**165F**, 288T). He certainly is lacking in self-confidence and believes that he is not as good as other people (*A-lse1*, 70T, **223F**, 379T). He certainly thinks he

is useless at times (130T). He usually has to stop and think before he acts even in small matters (310T). He has several times given up doing a thing because he thought too little of his ability (306T). He is obsessed with his perceived personal deficiencies (7[Pt], A-obs). He believes that his judgment is not as good as it once was (40F). He has difficulty starting to do things (218T), and he gives up quickly when things go wrong (340T). His inertia and lack of drive reflect his depressive cognitions and negative expectations (2[D], A-dep). He is pessimistic and hopeless about making any substantial changes in his circumstances (2[D], A-dep). He feels guilty when his high standards and expectations are not met (7[Pt]). He wishes he could be as happy as others seem to be (53T).

**Interpersonal Relations** He reports that he is shy and introverted (0[Si], A-sod1). He is easily embarrassed (270T), wishes he was not so shy (178T), and frequently has to fight against showing that he is bashful (158T). He finds it hard to talk when he meets new people (160T) and he has trouble thinking of the right things to talk about when in a group of people (227T). He will avoid people if given the opportunity (Si2). He is likely not to speak to people until they speak to him (248T). Even when he is with people, he feels lonely much of the time (259T). He is passive and dependent in his relationships and is unlikely to be assertive (A-lse2). He reports mild conflict with members of his family (Pd1, A-fam1). He is alienated from himself (Pds, Sis) and others (Pd4, Sci, Sis).

**Other Problem Areas** He reports that he has a number of gastrointestinal (A-hea1) and neurological symptoms (Sc6, A-hea2). He worries about his health (4F), and his health is not as good as that of most of his friends (42F). He has difficulty going to sleep because thoughts or ideas are bothering him (134F), his sleep is fitful and disturbed (36T), and he does not wake up fresh and rested most mornings (3F). He tires quickly (146F) and feels tired and without energy a good deal of the time (289F). His history and background should be reviewed to determine whether a medical or neurological evaluation is warranted. Suicidal thoughts are quite common and should be monitored carefully (2[D], 7[Pt]).

**Treatment** His prognosis is generally quite good because he sees himself as being responsible for his difficulties and is willing to examine his behavior, even at great length (2[D], 7[Pt]). Cognitive-behavioral psychotherapy focusing on his depressive cognitions **Minnesota Multiphasic Personality Inventory—Adolescent 223** will be very beneficial to him (2[D], 7[Pt]). Evaluation for antidepressant medication may be indicated in cases of more severe depression (2[D], A-dep). The primary reasons for his referral for treatment are tearfulness, restlessness, nervousness, anxiety, and worry, as well as excessive fantasy and daydreaming.

A number of specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: He shrinks from facing a crisis or difficulty (27T); he has difficulty starting to do things (218T); he is passive and nonassertive (280T); he gives up quickly when things go wrong (340T) or because he thinks too little of his ability (306T); he is hard to get to know (408T); and he is bothered greatly by the thought of making changes in his life (444T).

2-0/0-2

**Moods** She reports that she is experiencing mild emotional distress (A, NEGE) characterized by chronic brooding (203T), dysphoria (2[D], A-dep2), and anhedonia (INTR, A-dep1, A-trt1). She has had long periods when she could not take care of things because she could not get going (35T). She frequently worries about something or someone (185T, 281T, 285T, 383T). She is more sensitive (266T) and feels more intensely than most people (253T). Her feelings are easily hurt (60F, 121T) and she is inclined to take things hard (311T). She is often irritable and grouchy (A-ang2, 116T, 388T), and she easily becomes impatient with people (130T, 282T). Despite these angry feelings, she is very unlikely to be overtly aggressive (A-ang1, AGGR).

**Cognitions** She reports that she has problems with attention and concentration (28T, 305T). She certainly is lacking in self-confidence and believes that she is not as good as other people (A-lse1, 70T, 223F, 379T). She finds it hard to be assertive because she is so reserved (A-lse2). At times, she thinks she is useless (124T). She shrinks from facing a crisis or difficulty (27T) and is easily downed in an argument (67T). She has often lost out on things because she could not make up her mind quickly enough (129T). She is apt to

feel disappointments so keenly that she cannot put them out of her mind (377T). She gives up quickly when things go wrong (340T), when others criticize her, or when she thinks that she is unable to do something (306T). She is apt to pass up something she wants to do when others feel that it is not worth doing (280T). She thinks that most people will use unfair means and stretch the truth to get ahead (*A-cyn*, 55T, 77T, 107T) and it is safer to trust nobody (225T). She has often been misunderstood when she was trying to be helpful (403T). She wishes she could be as happy as others seem to be (56T).

**Interpersonal Relations** She reports that she is very socially introverted (*O[Si]*, *A-sod1*). She is easily embarrassed (270T), wishes she was not so shy (178T), and frequently has to fight against showing that she is bashful (154T). She avoids interactions with others (*Si2*), and she spends most of her spare time by herself (410T). Whenever possible, she avoids being in a crowd (304T), and she does not enjoy the excitement of a crowd (**335F**). At parties she is more likely to sit by herself or with just one other person than to join in with the crowd (316T). She finds it hard to talk when she meets new people (160T) or is in a group of people (227T), and she is likely not to speak to people until they speak to her (248T). She does not seem to make friends as quickly as others seem to do (**262F**). She sees herself as socially inept and awkward (*A-lse1*); others are more likely to describe her

#### 224 Self-Report Inventories

as shy and reserved (*A-sod2*). She feels lonely even when she is with people (259T). She is alienated from herself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*).

**Other Problem Areas** She reports general concerns about her physical health, but few specific symptoms (*A-hea3*). There is an increased frequency of eating disorders, including bulimia and anorexia (*2[D]*, *O[Si]*) so eating behavior should be reviewed carefully. She is unlikely to abuse substances (**ACK, PROS**) or to get into trouble because of her behavior (**Pd2, A-con**, 249T).

**Treatment** Her prognosis is only fair because she is very accustomed to her characterological problems and she is reluctant to think of making changes in her life (*2[D]*, *O[Si]*).

She often is referred for treatment because she is nervous and anxious, listless, apathetic, shy, and fearful (*2[D]*, *O[Si]*). She is essentially friendless at school (*2[D]*, *O[Si]*). Social skills and assertiveness training will frequently be beneficial (*O[Si]*). She responds well to structured treatment approaches that prescribe what she is to do (*2[D]*, *O[Si]*). Cognitivebehavioral approaches that focus on her depressive cognitions also will be beneficial (*2[D]*, *A-dep*). Group psychotherapy will be helpful in providing a social perspective for her problems and in dealing directly with her avoidant behaviors (*O[Si]*).

Several specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: She shrinks from facing a crisis or difficulty (27T); she has difficulty starting to do things (218T); she believes it is safer to trust nobody (225T); she is passive and nonassertive (280T); she gives up quickly when things go wrong (340T) or because she thinks too little of her ability (306T); and she feels unable to tell anyone all about herself (369T).

4-6/6-4

**Moods** He reports that he is experiencing mild to moderate emotional distress (*A, NEGE*) that is characterized by brooding (*6[Pa]*, *D5*), dysphoric mood (*2[D]*, *A-dep2*), and anhedonia (*INTR*, *A-dep1*, *A-trt1*). He broods and worries constantly over what is happening to him (*6[Pa]*, 285T). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (**60F**, 121T), and he is inclined to take things hard (311T). He is generally stubborn, argumentative, and angry (*6[Pa]*). He is usually able to control the expression of his anger, but he does exhibit episodic angry outbursts, particularly in response to stress (*A-ang1*, *AGGR*, *DISC*).

**Cognitions** He reports that his attention, concentration (**28F**, **305F**), and memory seem to be all right (165T). He remembers very well, and for a long time, anything that people say or do to him (*6[Pa]*). He likes to let people know where he stands on things (341T), and he finds it necessary to stand up for what he thinks is right (115T). He has very peculiar and strange experiences (29T). At times, his thoughts have raced ahead faster than he could speak them (116T). He has sometimes thought that difficulties were piling up so high he could not overcome them (318T). He is very sensitive to and resentful of any demands

being placed on him (6[Pa]). He is sure that he is being talked about (286T). He knows who is responsible for most of his troubles (109T). He has not lived the right kind of life (49T) and wishes that he could be as happy as others seem to be (53T). His way of doing things is apt to be misunderstood by others (211T).

**Minnesota Multiphasic Personality Inventory—Adolescent 225**

*[He reports symptoms that may reflect a psychotic process or a long-term, characterological condition. His presenting problems, background, and history should be reviewed with this possibility in mind.]*

**Interpersonal Relations** He reports a balance between extroverted and introverted behaviors (0[Si]) and that he is comfortable in social situations (**A-sod**). He is very sociable and makes friends quickly (46T, 262T, 336T). He believes, if given the chance, he would make a good leader of people (329T). He enjoys social gatherings and parties (331T, 292T) and the excitement of a crowd (335T, **304F**). However, he is likely to have a long history of poor interpersonal relations (4[*Pd*], 6[*Pa*]) often characterized by resentment, anger, and suspiciousness (6[*Pa*]). He sees his family as extremely uncaring and critical and his home life as very unpleasant (*Pd*, *Pd1*, *A-fam*). He is constantly in trouble with his parents because of his misbehavior (4[*Pd*], 6[*Pa*]). He is alienated and detached from himself (*Pd5*) and others (*Pd4*, *Sc1*, *A-fam2*).

**Other Problem Areas** He reports only a few general concerns about his physical health (1[*Hs*], *Hy4*, *A-hea3*). He has few or no pains (54T, 210T) and very few headaches (168T). He is in just as good physical health as most of his friends (42T). His sleep is not fitful and disturbed (**36F**), but he does not wake up fresh and rested most mornings (**3F**). He does not tire quickly (146T) and is not tired, weak, or without energy a good deal of the time (**167F**, 289T). He is about as able to work as he ever was (10T). He is likely to abuse alcohol or drugs (**ACK PRO**, 247T) so a careful review should be made of the consequences of his alcohol and drug use. He reported a number of conduct problems (*A-con2*) and that he has been in trouble with the law (**249F**). He has done dangerous things for the thrill of it (**96F**).

**Treatment** His prognosis is generally quite poor due to his lack of awareness of his role in his problems, which are chronic and characterological (4[*Pd*], 6[*Pa*]). He is frequently referred for treatment by the legal system because he is defiant, disobedient, tense, restless, and negativistic (4[*Pd*], 6[*Pa*]). He will be very demanding and will engage in frequent testing of the therapist (4[*Pd*], 6[*Pa*]). He typically makes excessive demands on others for attention and sympathy but is resentful of even mild demands that may be placed on him (4[*Pd*], 6[*Pa*]). Short-term, behavioral interventions that are presented in a direct and explicit manner will be most effective (4[*Pd*], 6[*Pa*]).

The therapeutic alliance must be developed slowly and cautiously because of his distrust of others (6[*Pa*]).

4-7/7-4

**Moods** He reports that he is experiencing a mild to moderate level of emotional distress (*A*, *NEGE*) characterized by anxiety and agitation (7[*Pt*], *A-anx*). He is unmotivated and feels unable to do much to help himself at this time and may be anhedonic (*INTR*, *A-dep1*, *A-trt1*). He frequently worries about something or someone (185T, 281T, 285T, 383T). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (**60F**, 121T) and he is inclined to take things hard (311T). He is grouchy and irritable (*A-ang2*) and easily becomes impatient with people (282T).

**226 Self-Report Inventories**

**Cognitions** He reports that he has problems with attention and concentration (28T, 279T, 305T), but his memory seems to be all right (165T). He certainly is lacking in self-confidence and believes that he is not as good as other people (*A-lse*, 70T, **223F**, 379T). He exhibits a cyclical pattern of acting-out followed by excessive concern, regret, and remorse over his behavior (4[*Pd*], 7[*Pt*]). His remorse, however, does not inhibit the repetition of this behavior and further episodes of acting-out (4[*Pd*], 7[*Pt*]). He does many things that he later regrets (78T). He believes that he has not lived the right kind of life (49T) and wishes he could be as happy as others seem to be (53T). He is apt to take disappointments so keenly that he cannot put them out of his mind (377T). He does not analyze the motives for his own or others' behavior (*Pa3*, *A-cyn*).

**Interpersonal Relations** He reports that he is slightly introverted (*0[Si]*), but he is very sociable and makes friends quickly (46T, 262T, 336T). He enjoys social gatherings and parties (331T, 292T) and the excitement of a crowd (335T, **304F**). His familial and interpersonal relations are marked by conflict and are often disrupted by episodic acting out (*4[Pd]*, *Pd1*, *Pd2*, *A-fam*). He is not particularly sensitive or responsive to the needs of others, except after something has happened (*4[Pd]*). He is alienated from himself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*, *A-fam2*).

**Other Problem Areas** He reports a few general concerns about his health (*A-hea3*). He has few or no pains (54T, 210T) and very few headaches (168T). He is in just as good physical health as most of his friends (42T). Most nights his thoughts or ideas bother his sleep (**134F**), and he does not wake up fresh and rested most mornings (**3F**). He is about as able to work as he ever was (10T). He is likely to abuse alcohol or drugs (*ACK*, *PRO*, 247T) so a careful review should be made of the consequences of his alcohol and drug use. He reported a number of conduct problems (*A-con2*), and he is likely to have been in trouble with the law (**249F**).

**Treatment** His prognosis is generally poor for short-term psychotherapy and guarded for long-term, intensive psychotherapy because of the characterological nature of his problems (*4[Pd]*, *7[Pt]*). He sees his main problems as picking the right kind of friends and his parents not liking his friends (*4[Pd]*, *7[Pt]*). His remorse and guilt over acting-out may give the impression of more insight and motivation to change than are actually present (*4[Pd]*, *7[Pt]*). Once his remorse and guilt have dissipated, his motivation will disappear quickly (*4[Pd]*, *7[Pt]*). Helping him to recognize his cyclic patterns and then to understand their dynamics is a primary goal in treatment (*4[Pd]*, *7[Pt]*).

Some specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: He has difficulty starting to do things (218T); he gives up quickly when things go wrong (340T) or because he thinks too little of his ability (306T); and he is very stubborn (416T).

4-8/8-4

**Moods** He reports that he is experiencing a moderate to severe level of emotional distress (*A*, *NEGE*) characterized by dysphoric mood (*2[D]*, *A-dep2*), agitation (*7[Pt]*, *9[Ma]*), and anhedonia (*INTR*, *A-dep1*, *A-trt1*). He frequently finds himself worrying about something (185T), finds it very difficult to get things done in his life, and has little hope of being successful even if he could get motivated at something (*INTR*, *A-dep1*, *A-trt1*). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (**60F**, 121T), and he is inclined to take things hard (311T). He easily becomes impatient with people (282T). He often feels resentful, angry (*6[Pa]*, *A-ang*), irritable, and grouchy (111T, 388T). He has difficulty controlling or expressing his anger appropriately (*8[Sc]*, *DISC*). In response to stress, he is likely to either withdraw completely (*8[Sc]*) or to act out his angry impulses (*6[Pa]*, *A-ang*, *DISC*). At times, he has a strong urge to do something harmful or shocking (81T).

**Cognitions** He reports that he has problems with attention and concentration (28T, 279T, 305T). He likes to let people know where he stands on things (341T) and he finds it necessary to stand up for what he thinks is right (115T). He certainly is lacking in self-confidence and believes that he is not as good as other people (*A-lse*, 70T, **223F**, 379T). He exhibits poor judgment and is often unpredictable and impulsive (*4[Pd]*, *8[Sc]*). His judgment is not as good as it once was (**40F**). He feels insecure, isolated, rejected, and unwanted (*8[Sc]*). He is threatened by a world that he views as hostile and dangerous (*8[Sc]*). He has had peculiar and strange experiences (29T) and thoughts (296T). He thinks and dreams of things that are too bad to talk about (15T, 208T). He is very sensitive to the motives of others and hypervigilant (*6[Pa]*). He has often thought that strangers were looking at him critically (235T). He thinks that most people will use unfair means and stretch the truth to get ahead (*A-cyn*, 55T, 77T, 107T). He often wonders what hidden reason another person may have for being nice to him (118T). He believes that it is safer to trust nobody (225T). He has not lived the right kind of life (49T) and has done many things that he later regrets (78T). He has often been misunderstood when he was trying to be helpful (373T), and his way of

doing things is apt to be misunderstood by others (211T).

[He reports symptoms that may reflect a psychotic process or a long-term, characterological condition. His presenting problems, background, and history should be reviewed with this possibility in mind.]

**Interpersonal Relations** He reports that he is introverted (*O[Si]*) and has difficulty with close, emotional relationships (*8[Sc]*). He lacks basic social skills (*4[Pd]*, *8[Sc]*) and tends to be socially withdrawn and isolated (*O[Si]*), yet he enjoys social gatherings and parties (331T, 292T) and the excitement of a crowd (335T, **304F**). He sees himself as being sociable and making friends quickly (46T, 262T, 336T), although he is emotionally distant, feels lonely, and knows that no one understands him and his problems (*8[Sc]*). His tendency to feel rejected by others often leads to hostility and conflict that only exacerbate his feelings of being alienated from others (*4[Pd]*, *8[Sc]*). He sees his family as extremely uncaring and critical and his home life as very unpleasant (*Pd1*, *A-fam*). He is constantly in trouble with his family because of his misbehavior (*4[Pd]*, *8[Sc]*). Once in a while, he feels hate toward members of his family whom he usually loves (240T). The things that some of his family have done have frightened him (302T). Some of his family have habits that bother and annoy him very much (194T). He is alienated from himself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*, *A-fam2*).

**Other Problem Areas** He reports a few gastrointestinal symptoms (*A-heal*) and a number of neurological symptoms (*Sc6*, *A-hea2*). He does not wake up fresh and rested most

#### **228 Self-Report Inventories**

mornings (**3F**). He does not feel weak all over much of the time (**167F**). At times he is all full of energy (289T). His physical and neurological symptoms warrant careful review and possible referral for a medical or neurological evaluation. He may abuse substances (*ACK*, *PRO*, 247T), which will exacerbate all his problems, so a careful review should be made of the consequences of his use of substances. His problems frequently involve inappropriate sexual behavior (*4[Pd]*, *8[Sc]*, **31F**). He reported a number of conduct problems (*A-con2*) and has been in trouble with the law (**249F**). If he commits a crime, it is likely to be poorly planned and executed and may involve bizarre or violent behavior (*4[Pd]*, *8[Sc]*).

**Treatment** His prognosis is generally very poor because of the characterological nature of his problems (*4[Pd]*, *8[Sc]*). Psychopharmacological interventions, other than possibly antidepressants, also are unlikely to be very effective because of the characterological problems involved (*4[Pd]*, *8[Sc]*). He frequently is referred because of immaturity coupled with extreme narcissism (*4[Pd]*, *8[Sc]*). He has very poor academic achievement and poor grades (*4[Pd]*, *8[Sc]*). Interventions focused on specific behavioral objectives may be useful (*4[Pd]*, *8[Sc]*). Any form of insight-oriented therapy is contraindicated (*4[Pd]*, *8[Sc]*). Social skill and/or assertiveness training in a group setting may be beneficial (*8[Sc]*, *O[Si]*). His difficulties in forming an emotional relationship and his reluctance to self-disclose make the establishment of a therapeutic alliance problematic at best (*4[Pd]*, *8[Sc]*). Some specific issues that must be kept in mind when establishing and maintaining the therapeutic alliance: No one seems to understand him (20T); he shrinks from facing a crisis or difficulty (27T); he has difficulty starting to do things (218T); it is safer to trust nobody (225T); he is so touchy on some subjects that he cannot talk about them (256T); he gives up quickly when things go wrong (340T) or because he thinks too little of his ability (306T); and he is very stubborn (416T).

4-9/9-4

**Moods** He reports that he is experiencing minimal emotional distress or concern (*A*, *NEGE*) about his behavior or present circumstances. In fact, he is happy most of the time (**62F**, 91T). He dislikes being bored and inactive and will stir up excitement if he gets bored (9[*Ma*], 162T). Once a week or oftener, he becomes very excited (226T). Something exciting will almost always pull him out of it when he is feeling low (228T). He frequently has intense feelings of anger and hostility that result in episodic outbursts of anger (9[*Ma*], *A-ang*, *AGGR*, *DISC*). He is likely to exhibit increased anger and hostility in response to stress (*4[Pd]*, 9[*Ma*]). He is oppositional when he is not overtly angry (*4[Pd]*, 9[*Ma*]).

**Cognitions** He reports that his attention, concentration (**28F**, **278F**, **305F**), and memory (165T) seem to be all right. He likes to let people know where he stands on things (341T),

and he finds it necessary to stand up for what he thinks is right (115T). He exhibits very poor judgment, often acts without considering the consequences of his actions, and has difficulty learning from experience (4[Pd], 9[Ma]). He is unwilling to accept responsibility for his own behavior and he exhibits a persistent tendency to get into trouble (4[Pd], 9[Ma]). He thinks that most people will use unfair means and stretch the truth to get ahead (A-cyn, 55T, 77T, 107T).

**Minnesota Multiphasic Personality Inventory—Adolescent 229**

**Interpersonal Relations** He reports that he is extroverted and mixes easily with others (0[Si], A-sod). He is very sociable and makes friends quickly (46T, 262T, 336T). He believes, if given the chance, he would make a good leader of people (329T). He enjoys social gatherings and parties (331T, 292T) and the excitement of a crowd (335T, 304F). He is socially facile (160F, 227F), and he makes a good first impression (4[Pd], 0[Si]). More long-term contact will reveal that his interpersonal relations are superficial and marked by impulsivity, distrust, a lack of empathy, and egocentricity (4[Pd], 9[Ma]). He has numerous, long-standing problems with his family and persons in positions of authority (4[Pd], Pd1, Pd2, A-fam). Some of his family have habits that bother and annoy him very much (194T). He is alienated from himself (Pd5, Si3) and others (Pd4, Sc1, Si3).

**Other Problem Areas** He reports that he is in good physical health (1[Hs], A-hea, Hy4). He has few or no pains (210T) and very few headaches (168T). He is in just as good physical health as most of his friends (42T). Sometimes he becomes so excited that he finds it hard to get to sleep (284T), but his sleep is not fitful or disturbed (36F). He does not tire quickly (146T) and is not tired, weak, and without energy a good deal of the time (167F, 289T). He is about as able to work as he ever was (10T). He is impulsive, risk-taking (4[Pd], 9[Ma], MAC-R, DISC), and is very likely to abuse substances (ACK, PRO, 247T) so a careful review should be made of the consequences of his alcohol and drug use. He has done dangerous things for the thrill of it (96F). He has a history of legal and school problems and acting-out, even if he is not in legal trouble at the present time (4[Pd], 9[Ma], A-con2, 249F). He does not like school because of lack of interest in and dislike of the subjects being taught (4[Pd], 9[Ma]).

**Treatment** His prognosis is guarded because this characterological process tends to result in fixed behavior patterns that are difficult to change (4[Pd], 9[Ma]). He often is referred for treatment because he is defiant, disobedient, impulsive, provocative, mischievous, and truant from school (4[Pd], 9[Ma]). He is experiencing minimal emotional distress that might serve as motivation to change his behavior (A, NEGE). Behavioral or short-term interventions that focus on the specific behaviors that led him to treatment will be most effective (4[Pd], 9[Ma]). Group therapies with similar adolescents frequently are helpful in confronting his characterological pattern of relating to others (4[Pd], 9[Ma]). Only one specific issue must be kept in mind when establishing and maintaining the therapeutic alliance: He is very stubborn (416T).

4-0/0-4

**Moods** He reports that he is experiencing a mild level of emotional distress (A, NEGE) characterized by dysphoric mood (2[D], A-dep2) and anhedonia (INTR, A-dep1, A-trt1). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (60F, 121T), and he is inclined to take things hard (311T). He easily becomes impatient with people (130T, 282T). He is often irritable and grouchy (111T, 388T).

**Cognitions** He reports that his attention, concentration (279F, 305F), and memory seem to be all right (165T). He certainly is lacking in self-confidence and believes that he is not as good as other people (A-lse, 70T, 223F, 379T). He thinks that most people will use unfair means and stretch the truth to get ahead (A-cyn, 55T, 77T, 107T). He often wonders what

**230 Self-Report Inventories**

hidden reason a person may have for doing something nice for him (118T). He has often thought that strangers were looking at him critically (235T). He believes that it is safer to trust nobody (225T). He believes that he has not lived the right kind of life (49T) and has done many things that he later regrets (78T). He has often been misunderstood when he was trying to be helpful (373T) and others are apt to misunderstand his way of doing things (211T). He wishes that he could be as happy as others seem to be (53T).

**Interpersonal Relations** He reports that he is a shy, retiring individual (*O[Si]*) who actively avoids social interactions (*A-sod, Si2*). Whenever possible, he avoids being in a crowd (**335F**, 304T). He is easily embarrassed (269T) and has to fight against showing that he is bashful (154T). He is likely not to speak to people unless they speak to him (248T). He does not like to meet strangers (**336F**) and finds it hard to talk when he meets new people (160T) or is in a group of people (227T). He does not seem to make friends as quickly as others do (**262F**). He is troubled by how slow he is in making friends and tends to have only a few rather close friends (*4[Pa], O[Si]*). His relations with his family are unpleasant, with little love or companionship (*4[Pa], Pd1, A-fam*). He is alienated from himself (*Pd5, Si3*) and others (*Pd4, Sc1, Si3*).

**Other Problem Areas** He reports that he is in good physical health (*I[Hs], A-hea, Hy4*). His sleep is not fitful and disturbed (**36F**), but he does not wake up fresh and rested most mornings (**3F**). He does not feel weak all over much of the time (**167F**). He is about as able to work as he ever was (10T). He is likely to abuse substances (*ACK, PRO, 247T*), so a careful review should be made of the consequences of his alcohol and drug use. He reported a number of conduct problems (*A-con2*) and he has been in trouble with the law (**249F**).

**Treatment** His prognosis is guarded because of the characterological nature of his problems (*4[Pa], O[Si]*). Group interventions focused on social skills or assertiveness training frequently will be beneficial (*O[Si], A-sod*). It will be important for the therapist to provide him support in initial group sessions because of his shyness (*O[Si], A-sod2*). He has little motivation for any long-term intervention (*A, NEGE*).

Many specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: No one seems to understand him (20T); he shrinks from facing a crisis or difficulty (27T); he has difficulty starting to do things (218T); he believes it is safer to trust nobody (225T); he is so touchy on some subjects that he cannot talk about them (256T); he is passive and nonassertive (280T); he gives up quickly when things go wrong (340T) or because he thinks too little of his ability (306T); and he is very stubborn (416T). Developing a therapeutic alliance must proceed slowly because of his sensitivity and shy, retiring style (*O[Si]*). He questions the motives of others and is distrustful of people in general (*4[Pa], O[Si]*), which makes the whole therapeutic process more difficult.

6-8/8-6

**Moods** She reports that she is experiencing severe emotional distress (*A, NEGE*) characterized by dysphoric mood (2[D], *A-dep2*), agitation (9[Ma]), worrying (7[Pt], *A-anx*), and anhedonia (*INTR, A-dep1, A-trt1*). Her affect is likely to be blunted or inappropriate (6[Pa], 8[Sc]). She sees little opportunity of improving her circumstances, further dampening her Minnesota Multiphasic Personality Inventory—Adolescent 231 mood (6[Pa], 8[Sc]). She is very fearful, easily frightened, and generally apprehensive (8[Sc]). Several times a week she feels as if something dreadful is about to happen (402T). She frequently worries about something or someone (185T, 281T, 285T, 383T). She is more sensitive (266T) and feels more intensely than most people (253T). Her feelings are easily hurt (**60F**, 121T), and she is inclined to take things hard (311T). She easily becomes impatient with people (130T, 282T). She is often irritable and grouchy (*A-ang*, 111T, 388T) and has a violent temper (6[Pa], 8[Sc]). She gets angry when people hurry her (401T). At times, she has a strong urge to do something harmful or shocking (81T), and she feels like smashing things (34T). At times she has fits of laughing and crying that she cannot control (21T).

**Cognitions** She reports that she has problems with attention, concentration (28T, 279T, 305T), and memory (**158F**, 288T). She likes to let people know where she stands on things (341T) and she finds it necessary to stand up for what she thinks is right (115T). She certainly is lacking in self-confidence, believes that she is not as good as other people (70T, **223F**, 379T), and feels inferior and insecure (*A-lse*). Sometimes some unimportant thought will run through her mind and bother her for days (308T). Her plans have frequently seemed so full of difficulties that she has had to give them up (370T).

She has had very strange and peculiar experiences (29T) and thoughts (296T). She thinks and dreams of things that are too bad to talk about (15T, 208T). She prefers to daydream

rather than do anything else (45T). She thinks that there is something wrong with her mind (163T, 173T) and that she is about to go to pieces (404T). She has had blank spells in which her activities were interrupted and she did not know what was going on around her (214T) and periods in which she carried on activities without knowing later what she had been doing (161T). Bad words, often terrible words, come into her mind and she cannot get rid of them (307T). She often feels as if things are not real (329T). She has often thought that strangers were looking at her critically (235T). She is sure that she is being talked about (286T), she is being plotted against (132T), people say vulgar and insulting things about her (314T), and someone has it in for her (95T). She has enemies who wish to harm her (294F). If people had not had it in for her, she would have been much more successful (39T). Much of the time, she thinks she has done something wrong or evil (90T) and believes she is a condemned person (219T). She is sure she gets a raw deal from life (16T) and that she is often being punished without cause (137T).

She thinks that most people will use unfair means and stretch the truth to get ahead (*A-cyn*, 55T, 77T, 107T). She often wonders what hidden reason another person may have for doing something nice for her (118T). She believes that it is safer to trust nobody (225T). She has often been misunderstood when she was trying to be helpful (373T). She does many things that she later regrets (78T).

*[She reports symptoms that may reflect a psychotic process or a long-term, characterological condition. Her presenting problems, background, and history should be reviewed with this possibility in mind.]*

**Interpersonal Relations** She reports that she is extremely introverted (*0[Si]*) and very uncomfortable around others (*A-sod*). Her behavior is likely to be unpredictable and inappropriate, also making others uneasy around her (*6[Pa]*, *8[Sc]*). She is suspicious and distrustful of others, and she avoids serious emotional relationships (*6[Pa]*, *8[Sc]*). She

#### **232 Self-Report Inventories**

generally feels apathetic, socially isolated, and withdrawn, and she believes that no one understands her (*6[Pa]*, *8[Sc]*). She feels lonely most of the time even when she is with people (259T). She is liked by few students and has few friends of any kind; any friends that she might have are older (*6[Pa]*, *8[Sc]*). She has poor relations with her family (*Pd1*, *A-fam*) and is alienated from herself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*). Once in a while she feels hate toward members of her family whom she usually loves (240T).

**Other Problem Areas** She reports a number of specific gastrointestinal (*A-hea1*) and neurological symptoms (*Sc6*, *A-hea2*) as well as a wide variety of more general physical symptoms (*1[Hs]*, *A-hea*, *Hy4*). Her head seems to hurt all over much of the time, and she has frequent headaches (37T, 97T, 168F). She has difficulty going to sleep because she is excited (284T) or thoughts or ideas are bothering her (134F), her sleep is fitful and disturbed (36T), and she does not wake up fresh and rested most mornings (3F). She tires quickly (146F) and feels weak and tired a good deal of the time (167T). Her history and background should be reviewed to determine whether a medical or neurological evaluation is warranted. She is likely to abuse substances (*ACK*, *PRO*, 247T), so a careful review should be made of the consequences of her alcohol and drug use. She is worried about sex (159T).

**Treatment** Her prognosis is generally poor because her problems are chronic and severe (*6[Pa]*, *8[Sc]*). She is usually referred because of her bizarre behavior or excessive fantasies (*6[Pa]*, *8[Sc]*). Her ability to work may not be severely impaired as long as the job does not involve any appreciable amount of contact with people (*6[Pa]*, *8[Sc]*). Some type of psychopharmacological intervention may be necessary to stabilize her thought processes and mood and to help her sleep (*6[Pa]*, *8[Sc]*). Short-term, behavioral interventions are warranted rather than any form of insight-oriented psychotherapy (*6[Pa]*, *8[Sc]*).

A number of specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: No one seems to understand her (20T); she shrinks from facing a crisis or difficulty (27T); she has difficulty starting to do things (218T); she believes it is safer to trust nobody (225T); she is so touchy on some subjects that she cannot talk about them (256T); she is very passive and nonassertive (280T, 385T); she gives up quickly when things go wrong (340T) or because she thinks too little of her ability (306T); she

feels unable to tell anyone all about herself (369T); the future seem hopeless to her (399T); she is hard to get to know (408T); she is very stubborn (416T); she is bothered greatly by the thought of making changes in her life (444T) and is not sure that she can make these changes (426T); and she hates going to doctors even when she is sick (434T).

6-9/9-6

**Moods** He reports that he is experiencing a mild to moderate level of emotional distress (A, *NEGE*) characterized by agitation, tension, and excitement (9[*Ma*]). His daily life is full of things that keep him interested (9T). Something exciting will almost always pull him out of it when he is feeling low (228T). When he gets bored, he likes to stir up excitement (162T). Once a week or oftener, he becomes very excited (226T). Sometimes without any reason or even when things are going wrong he feels excitedly happy or unusually cheerful (212T, 298T). He has very few fears compared with his friends (244T). He also reports mild dysphoric mood (*A-dep2*), yet he is happy most of the time (**62F**, 91T). He

**Minnesota Multiphasic Personality Inventory—Adolescent 233**

feels more intensely than most people do (253T). He has difficulty expressing his feelings appropriately and may vacillate between overcontrolling and undercontrolling his emotions (6[*Pa*], 9[*Ma*]). It makes him angry when people hurry him (401T). Others will perceive him as being irritable, grouchy, and angry (9[*Ma*], *A-ang*, *AGGR*).

**Cognitions** He reports that his attention, concentration (**28F**), and memory seem to be all right (158T). He likes to let people know where he stands on things (341T), and he finds it necessary to stand up for what he thinks is right (115T). He has had very strange and peculiar experiences (29T). He often exercises poor judgment, although he thinks that his judgment is better than it ever was (6[*Pa*], 9[*Ma*], 40T). At times, he can make up his mind with unusually great ease (195T). He is sure he is being talked about (286T) and being punished without cause (137T). He is grandiose and egocentric (6[*Pa*], 9[*Ma*], 58T). He thinks that most people will use unfair means and stretch the truth to get ahead (*A-cyn*, 55T, 77T, 107T). He has often been misunderstood when he was trying to be helpful (373T).

[*He reports symptoms that may reflect a psychotic process or a long-term, characterological condition. His presenting problems, background, and history should be reviewed with this possibility in mind.*]

**Interpersonal Relations** He reports that he is extroverted (**0[*Si*]**) and comfortable in social situations (*A-sod*). He is very sociable and makes friends quickly (46T, 262T, 336T). He believes, if given the chance, he would make a good leader of people (329T). He enjoys social gatherings, parties (292T, 331T), and the excitement of a crowd (335T, **304F**). His relations with his family (*Pd1*, *A-fam*) and others (*Pd2*) tend to be strained at best. He is alienated from himself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*).

**Other Problem Areas** He reports gastrointestinal (*A-hea1*) and neurological symptoms (*Sc6*, *A-hea2*), but states that he is in just as good physical health as most of his friends (42T), with few or no pains (210T). Sometimes he becomes so excited, he finds it hard to get to sleep (284T). He wakes up fresh and rested most mornings (3T). He does not tire quickly (146T) and is not tired, weak, and without energy a good deal of the time (**167F**, 289T). He is about as able to work as he ever was (10T). His history and background should be reviewed to determine whether a medical or neurological evaluation is warranted. He is prone to abuse substances (*ACK*, *PRO*, 247T), so a careful review should be made of the consequences of his alcohol and drug use. He reported a number of conduct problems (*Acon2*) and he has been in trouble with the law (**249F**). He is a risk-taker and sensation-seeker (*MAC-R*, **96F**).

**Treatment** His prognosis is generally poor because of his limited concerns about his own behavior (6[*Pa*], 9[*Ma*]). Psychopharmacological interventions are frequently warranted because of the agitation and excitement (6[*Pa*], 9[*Ma*]). Short-term, behavioral interventions that focus on his reasons for entering treatment will be most effective (6[*Pa*], 9[*Ma*]).

7-8/8-7

**Moods** She reports that she is experiencing a moderate to severe level of emotional distress (A, *NEGE*) characterized by dysphoric mood (*A-dep2*, 62T), brooding, and agitation (2[*D*], *A-dep*, 7[*Pt*], *A-anx*). She frequently worries about something or someone (185T, 281T, **234 Self-Report Inventories**

285T, 300T, 383T). She is chronically stressed, and she becomes more agitated or withdrawn as her level of stress increases (7[Pt], 8[Sc]). She is generally apprehensive and fearful of her environment (8[Sc]). Several times a week she feels as if something dreadful is about to happen (402T). She obtains little pleasure from life and is likely to be anhedonic (*INTR*, *A-dep1*, *A-trt1*). She is more sensitive (266T) and feels more intensely than most people (253T). Her feelings are easily hurt (60F, 121T), and she is inclined to take things hard (311T). She easily becomes impatient with people (130T, 282T). She is often irritable and grouchy (*A-ang1*, 111T, 388T). It makes her angry when people hurry her (401T). At times, she has a strong urge to do something shocking or harmful (81T).

**Cognitions** She reports that she has problems with attention, concentration (28T, 279T, 305T), and memory (158F, 288T). She certainly is lacking in self-confidence and believes that she is not as good as other people (*A-lse*, 70T, 223F, 379T). She has often lost out on things because she could not make up her mind quickly enough (129T). Her plans have frequently seemed so full of difficulties that she had to give them up (370T). Sometimes some unimportant thought will run through her mind and bother her for days (308T). She has had strange and peculiar thoughts (296T) and thinks and dreams of things that are too bad to talk about (15T, 208T). She prefers to daydream rather than do anything else (45T). She often engages in sexual fantasies and her sexual adjustment is likely to be poor (7[Pt], 8[Sc]). She often thinks that things are not real (291T), there is something wrong with her mind (163T, 173T), and she is about to go to pieces (404T). Much of the time she feels as if she has done something wrong or evil (90T). She has had periods in which she carried on activities without knowing later what she had been doing (161T, 214T). Her thinking may be quite deviant and she may have experienced either auditory or visual hallucinations (7[Pt], 8[Sc]).

She thinks that most people will use unfair means and stretch the truth to get ahead (*Acyn*, 55T, 77T, 107T). She often wonders what hidden reason another person may have for being nice to her (118T). She believes it is safer to trust nobody (225T). She has often felt that strangers were looking at her critically (235T). She has often been misunderstood when she was trying to be helpful (373T) and her way of doing things is apt to be misunderstood by others (211T). She does many things that she later regrets (78T).

*[She reports symptoms that may reflect a psychotic process or a long-term, characterological condition. Her presenting problems, background, and history should be reviewed with this possibility in mind.]*

**Interpersonal Relations** She reports that she is extremely introverted (0[Si]) and socially uncomfortable (*A-sod*) and has poor social skills and judgment (7[Pt], 8[Sc]). When in a group of people she has trouble thinking of the right things to talk about (227T). She is easily embarrassed (270T) and wishes she was not so shy (178T). She has difficulty forming close, personal relationships (7[Pt], 8[Sc]). Even when she is with people, she feels lonely much of the time (259T). Her relations with her family tend to be poor (*Pd1*, *A-fam*). She is very alienated from herself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*).

**Other Problem Areas** She reports a number of physical symptoms (*1[HS]*, *A-hea*, *Hy4*), general concern about her health (*A-hea3*), gastrointestinal symptoms (*A-hea1*), and neurological symptoms (*Sc6*, *A-hea2*). She has difficulty going to sleep because she is excited

**Minnesota Multiphasic Personality Inventory—Adolescent 235**

(284T) or thoughts or ideas are bothering her (134F), her sleep is fitful and disturbed (36T), and she does not wake up fresh and rested most mornings (3F). She tires quickly (146F) and feels weak and tired a good deal of the time (167T). She is not as able to work as she once was (10F). Her history should be reviewed to determine whether a medical or neurological evaluation is warranted. She is worried about and bothered by thoughts about sex (159T, 251T). She is likely to abuse substances (*AAS*, *PRO*, 247T), so a careful review should be made of the consequences of her alcohol and drug use.

**Treatment** Her prognosis is generally poor given the characterological nature of her problems and her diminished motivation to work (7[Pt], 8[Sc]). Establishing a therapeutic relationship is challenging because of the serious character pathology that is present (7[Pt], 8[Sc]). Psychopharmacological intervention may be necessary to decrease her level of agitation and to help her sleep (7[Pt], 8[Sc]). Cognitive-behavioral interventions focused

on her depressive and anxious cognitive processes will be beneficial (2[D], *A-dep*, 7[Pt], *A-anx*).

Numerous specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: No one seems to understand her (20T); she shrinks from facing a crisis or difficulty (27T); she has difficulty starting to do things (218T); she believes it is safer to trust nobody (225T); she is so touchy on some subjects that she cannot talk about them (256T); she is passive and nonassertive (280T, 385T); she gives up quickly when things go wrong (340T) or because she thinks too little of her ability (306T); she feels unable to tell anyone all about herself (369T); she is hard to get to know (408T); she is very stubborn (416T); she hates going to doctors even when she is sick (434T); and she is bothered greatly by the thought of making changes in her life (444T).

7-0/0-7

**Moods** He reports that he is experiencing a moderate level of emotional distress (*A*, *NEGE*) characterized by dysphoric mood (2[D], *A-dep2*), worrying (7[Pt], *A-anx*), and guilt (7[Pt]). He gets little pleasure from life and may be anhedonic (*INTR*, *A-dep1*, *A-trt1*). He frequently worries about something or someone (185T, 281T, 285T, 383T). He is more sensitive (266T) and feels more intensely than most people (253T). His feelings are easily hurt (60F, 121T), and he is inclined to take things hard (311T). He easily becomes impatient with people (130T, 282T). He is often irritable and grouchy (*A-ang1*, 111T, 388T).

**Cognitions** He reports that he has problems with attention and concentration (28T, 279T, 305T), but his memory seems to be all right (158T). He certainly is lacking in self-confidence and believes that he is not as good as other people (*A-lse*, 70T, 223F, 379T). He has often lost out on things because he could not make up his mind quickly enough (129T). He usually has to stop and think before he acts even in small matters (310T). He lets other people take charge (430T). He shrinks from facing a crisis or difficulty (27T). He is easily downed in an argument (67T). He is apt to pass up something he wants to do because others think it is not worth doing (280T). He takes disappointments so keenly that he cannot put them out of his mind (377T).

He thinks that most people will use unfair means and stretch the truth to get ahead (*A-cyn*, 55T, 77T, 107T). He has often thought that strangers were looking at him critically (236 Self-Report Inventories).

(235T). He has often been misunderstood when he was trying to be helpful (373T). He does many things that he later regrets (78T).

**Interpersonal Relations** He reports that he is introverted (*O[Si]*) and uncomfortable in social situations (*A-sod*). He is shy (178T), bashful (154T), and easily embarrassed (270T). At parties, he is more likely to sit by himself or with one other person than to join in with the crowd (316T). Whenever possible, he avoids being in a crowd (304T, 335F). Even when he is with people, he feels lonely much of the time (259T). He finds it hard to talk when he meets new people (160T) or in a group of people (227T). He is not likely to speak to people until they speak to him (248T). He does not seem to make friends as quickly as others do (262F). He spends most of his spare time by himself (410T). He reports conflicted relations with his family (*Pd1*, *A-fam*). He is alienated from himself (*Pd5*, *Si3*) and others (*Pd4*, *Sc1*, *Si3*).

**Other Problem Areas** He reports a few physical symptoms (*A-hea*, *Hy4*) and general concern about his health (*A-hea3*), but states that he is in just as good physical health as most of his friends (42T). His sleep is not fitful or disturbed (36F), but he does not wake up fresh and rested most mornings (3F). He tires quickly (146F) and feels tired a good deal of the time (167T). He is about as able to work as he ever was (10T). He is not likely to abuse substances (*ACK*, *PRO*). His social isolation and fear of social interaction (7[Pt], *O[Si]*) decrease the possibility of conduct problems (*A-con*).

**Treatment** His prognosis is good for long-term therapy because he is experiencing a significant degree of emotional distress (*A*, *NEGE*), and he is willing to think about his problems (7[Pt], *A-obs*). He usually is referred for his shyness and extreme sensitivity (7[Pt], *O[Si]*). He is generally a serious person who is given to anticipating problems and difficulties (7[Pt], *O[Si]*). Social-skill training or assertiveness training may be beneficial

(0[Si]).

Some specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: He shrinks from facing a crisis or difficulty (27T); he has difficulty starting to do things (218T); he is so touchy on some subjects that he cannot talk about them (256T); he is passive and nonassertive (280T, 385T); he gives up quickly when things go wrong (340T) or because he thinks too little of his ability (306T); he feels unable to tell anyone all about himself (369T); he is hard to get to know (408T); he is very stubborn (416T); and he is bothered greatly by the thought of making changes in his life (444T).

8-9/9-8

**Moods** He reports that he is experiencing moderate to severe emotional distress (A, NEGE) characterized by agitation and excitement (8[Sc], 9[Ma]). He is often first seen in an acute state of hyperactivity, excitement, confusion, and disorientation (8[Sc], 9[Ma]). When he gets bored, he likes to stir up some excitement (162T). Sometimes without any reason or even when things are going wrong, he feels excitedly happy (212T). He has periods in which he feels unusually cheerful without any special reason (298T). Something exciting will almost always pull him out of it when he is feeling low (228T). He easily becomes impatient with people (282T). He is often angry, irritable, and grouchy (A-ang, Minnesota Multiphasic Personality Inventory—Adolescent 237 AGGR, 111T, 388T). At times, he has a strong urge to do something shocking or harmful (81T).

**Cognitions** He reports that he has problems with attention and concentration (28T, 279T, 305T), but his memory seems to be all right (158T). He likes to let people know where he stands on things (341T), and he finds it necessary to stand up for what he thinks is right (115T).

He thinks and dreams of things that are too bad to talk about (15T, 208T) and he has strange and peculiar experiences (29T) and thoughts (296T). He has had periods in which he carried on activities without knowing later what he had been doing (161T, 214T). In response to stress, he is likely to become more disorganized and agitated or to engage in more daydreaming and fantasy (8[Sc], 9[Ma]).

He thinks that most people will use unfair means and stretch the truth to get ahead (A-cyn, 55T, 77T, 100T, 107T). Most people make friends because friends are likely to be useful to them (238T). He often wonders what hidden reason another person may have for doing something nice for him (118T). He believes that it is safer to trust nobody (225T). He has often been misunderstood when he was trying to be helpful (373T), and his way of doing things is apt to be misunderstood by others (211T).

[He reports symptoms that may reflect a psychotic process or a long-term, characterological condition. His presenting problems, background, and history should be reviewed with this possibility in mind.]

**Interpersonal Relations** He reports a balance between extroverted and introverted behaviors (0[Si]) and he is comfortable in social situations (A-sod). He is sociable and makes friends quickly (46T, 262T, 336T). He enjoys social gatherings and parties (82T, 292T, 331T) and the excitement of a crowd (304F, 335T). He likes to be with people who play jokes on one another (217T). He believes, if given the chance, he would make a good leader of people (329T). His behavior may be unpredictable, and he may act out unexpectedly (8[Sc], 9[Ma]). He is fearful of relating to others; consequently, intimate relationships are usually lacking (8[Sc], 9[Ma]). His relations with his family (Pd1, A-fam) and others (Pd2) are marked by conflict. Once in a while, he feels hate toward members of his family whom he usually loves (240T). He is alienated from himself (Pd5, Si3) and others (Pd4, Sc1, Si3).

**Other Problem Areas** He reports a number of general physical symptoms (I[Hs], A-hea, Hy4), gastrointestinal symptoms (A-heal), and neurological symptoms (Sc6, Ahea2). Sometimes he becomes so excited that he finds it hard to get to sleep (282T). His history should be reviewed to determine whether a medical or neurological evaluation is warranted. He is likely to abuse substances, so a careful review should be made of the consequences of his alcohol and drug use (ACK, PRO, 247T). He is a risk-taker and sensation seeker (MAC-R) who is likely to have legal and conduct problems (A-con2).

**Treatment** His prognosis is generally poor because of the serious characterological nature

of his problems (8[Sc], 9[Ma]). Psychopharmacological intervention may be helpful in reducing his agitation (8[Sc], 9[Ma]). The difficulties he experiences in focusing on specific issues and his fear of relating to others often preclude good therapeutic contact and outcome (8[Sc], 9[Ma]). He has little capacity for forming close interpersonal relationships, yet he is

#### **238 Self-Report Inventories**

rarely referred for poor peer or sibling relationships (8[Sc], 9[Ma]). Short-term, behavioral interventions that focus on his reasons for entering treatment will be most effective (8[Sc], 9[Ma]).

The following specific issues must be kept in mind when establishing and maintaining the therapeutic alliance: He has difficulty starting to do things (218T); it is safer to trust nobody (225T); he is so touchy on some subjects that he cannot talk about them (256T); he is hard to get to know (408T); and he is very stubborn (416T).

**Individual Scales** The interpretive text for every MMPI-A codetype provides the broad overview of how the *group* of adolescents within the codetype scored on the various scales. The task for the clinician at this juncture is to review all the individual scales (clinical, content, and supplementary) and critical items for this specific adolescent to determine how the interpretive text needs to be modified. Once the interpretive text for the MMPI-A has been rendered internally consistent, the clinician then must integrate the clinical history and background into the interpretation.

The individual scales on the MMPI-A can be organized within two large categories that reflect how the scales were developed: empirically or rationally. The 10 clinical scales for the MMPI-2 and MMPI-A were developed empirically and the 15 content scales (Williams, Butcher, Ben-Porath, & Graham, 1992), and PSY-5 scales (McNulty, Harkness, Ben-Porath, & Williams, 1997) were developed rationally. These two approaches are conceptually opposed to one another as described in Chapter 5 (pp. 116–117) and provide two methods for interpreting the MMPI-A. When the adolescent endorses the items accurately, these two approaches provide a similar interpretation of the MMPI-A profile, and the clinician's task is relatively straightforward. When the adolescent for any reason is motivated to provide either a self-unfavorable or self-favorable description, the approaches will provide divergent interpretations, and the clinician's task becomes more daunting because it is necessary to determine the basis for these divergent interpretations.

#### *Clinical and Content Scales*

Tables 7.4 and 7.5 provide general interpretive text for the clinical and content scales, respectively, when they are elevated above a T score of 64 or below a T score of 45. Figure 7.3 illustrated the profile for the basic validity and clinical scales, and Figure 7.4 illustrates the profile for the content scales for the same *1-3/3-1* codetype. The major emphasis in the interpretation of the MMPI-A is on elevated scales, yet very important information frequently is found in scales that are not elevated. For example, adolescents in a clinical setting who do not elevate Scales 2 (*D*: Depression) and 7 (*Pt*: Psychasthenia) are reporting that they are not experiencing any distress about the behavior or symptoms that led them to the clinical setting or about being evaluated.

A total of 15 new content scales have been developed for the MMPI-A (Williams et al., 1992): Anxiety, Obsessiveness, Depression, Health Concerns, Alienation, Bizarre Mentation, Anger, Cynicism, Conduct Problems, Low Self-Esteem, Low Aspirations, Social Discomfort, Family Problems, School Problems, and Negative Treatment Indicators (see Table 7.1 and Figure 7.4). These scales were developed in a similar manner to the MMPI-2 content (Butcher, Graham, Williams, & Ben-Porath, 1990) scales. The stages in the development of the MMPI-A content scales are outlined in Butcher et al. (1992, p. 61)

#### **Minnesota Multiphasic Personality Inventory—Adolescent 239**

#### **Table 7.4 Interpretation of the MMPI-A clinical scales**

##### Scale Potential Issues

*1 (Hs) > 64* Adolescents have a preoccupation with vague physical ailments. They are pessimistic about being helped. They are unlikely to be doing well in school, reporting increasing problems, primarily academic.

*1 (Hs) < 45* No specific interpretations can be made.

*2 (D) > 64* Adolescents are reporting generalized distress and likely to be depressed. Their depressive mood should be readily apparent. It is important to determine whether

internal or external factors are producing the negative mood state and to plan treatment accordingly.

2 (*D*) < 45 Adolescents are not reporting any type of emotional distress either as a result of their presence in treatment or the behaviors/symptoms that led them to consider treatment. The possibility of acting out in an impulsive manner should be evaluated. There is little internal motivation for any type of treatment or intervention.

3 (*Hy*) > 64 Adolescents are dependent, suggestible, and lack insight into their own and others' behavior. They deny any type of psychological problems. Under stress, specific physical ailments will be seen. They look for simplistic, concrete solutions to their problems. Treatment should focus on short-term goals because there is limited motivation. They initially may be enthusiastic about treatment, then later resist treatment or fail to cooperate.

3 (*Hy*) < 45 Adolescents are caustic, sarcastic, and socially isolated. They have few defenses for coping with any problems that they encounter. Well-structured, behavioral interventions should be used whenever possible.

4 (*Pd*) > 64 Adolescents are in conflict either with family members or with persons in positions of authority. They are likely to abuse alcohol or drugs. They may make a good initial impression, but more long-term contact will reveal that they are egocentric and have little concern for others. Any treatment should focus on short-term goals with emphasis on behavior change rather than their verbalized intent to change, no matter how sincere they may sound. Low scores on Scales 2 (*D*: Depression) and 7 (*Pt*: Psychasthenia) make elevations on Scale 4 particularly pathognomonic.

4 (*Pd*) < 45 Adolescents are rigid, conventional, and have little psychological insight into themselves or others. Explicit, behavioral directives to change will be most productive if there is sufficient motivation to follow them.

5 (*Mf*) > 64 Adolescents do not identify with their traditional gender role and are concerned about sexual issues. Boys frequently worry and their feelings are easily hurt. Girls are confident and satisfied with themselves.

5 (*Mf*) < 40 Adolescents identify with their traditional gender role. Men are confident and self-assured. Women are trusting of and depend on others and lack self-confidence. Their feelings are easily hurt and they cry easily.

6 (*Pa*) > 64 Adolescents are suspicious, hostile, and overly sensitive, which is readily apparent to everyone. Any treatment is problematic because of the difficulty in developing a therapeutic relationship based on trust. Any intervention must be instituted slowly.

(continued)

#### **240 Self-Report Inventories**

##### **Table 7.4 (Continued)**

###### Scale Potential Issues

6 (*Pa*) < 45 Adolescents have narrow interests, and they tend to be insensitive to and unaware of the motives of others. Explicit, behavioral directives to change will be most productive if there is sufficient motivation to follow them.

7 (*Pt*) > 64 Adolescents are worried, tense, and indecisive, which is readily apparent to everyone. Ineffective ruminative, and obsessive behaviors will be seen. It may be necessary to lower their level of anxiety before implementing treatment of other symptoms.

7 (*Pt*) < 45 Adolescents are secure and comfortable with themselves, which augurs poorly for any type of intervention in a clinical setting.

8 (*Sc*) > 64 Adolescents feel alienated and remote from the environment and others. At higher elevations (>79), difficulties in logic and judgment may become evident.

Interventions should be directive and supportive. Psychotropic medications may be needed.

8 (*Sc*) < 45 Adolescents are conventional, concrete, and unimaginative. Any intervention should be behavioral, directive, and focused on short-term goals.

9 (*Ma*) > 64 Adolescents are overactive, impulsive, emotionally labile, and euphoric with occasional outbursts of anger. Short-term behavioral goals should be pursued.

9 (*Ma*) < 45 Adolescents have low energy and activity levels. They may be manifesting significant signs of psychomotor retardation that should be evaluated carefully.

0 (*Si*) > 64 Adolescents are introverted, shy, and socially insecure. They withdraw from and avoid significant others who exacerbate their distress. Interventions need to

address specifically their tendency to withdraw and avoid others.

*O (Si) < 45* Adolescents are extroverted, gregarious, and socially poised. They may have difficulty in forming intimate relationships with others at very low scores (*T < 35*). They are unlikely to have a thought disorder. The probability of acting out is increased. Group therapies are particularly useful with these adolescents.

30

40

50

60

70

80

90

100

110

120

A-anx A-obs A-dep A-hea A-aln A-biz A-ang A-cyn A-con A-lse A-las A-sod A-fam A-sch A-trt

MMPI-A Content Scales T

Scores

**Figure 7.4 MMPI-A content scales for a 1-3/3-1 codetype.**

**Minnesota Multiphasic Personality Inventory—Adolescent 241**

**Table 7.5 Interpretation of the MMPI-A content scales**

Scale Potential Issues

*A-anx* (Anxiety) > 64 Adolescents report general symptoms of anxiety, nervousness, worries, and sleep and concentration difficulties. Depending on the level of anxiety, psychotropic medications or other anxiety-reducing techniques may be needed before implementing other interventions.

*A-obs* (Obsessiveness) > 64 Adolescents have great difficulty making decisions, ruminate and worry excessively, and have intrusive thoughts. They are good candidates for most insight-oriented therapies.

*A-dep* (Depression) > 64 Adolescents have difficulty getting going and getting things done in their lives (*A-dep1*). They have a dysphoric mood (*A-dep2*) and thoughts and a negative self-concept (*A-dep3*). Their depression has an angry component that involves blaming others, particularly when *A-dep* is higher (+15 T points) than Scale 2 (*D*: Depression).

*A-hea* (Health Concerns) > 64 Adolescents report gastrointestinal symptoms (*A-hea1*) and symptoms associated with neurological functioning (*A-hea2*), as well as general concerns about their health (*A-hea3*). Their physical symptoms may be another manifestation of their emotional distress. They need to be reassured that their symptoms are being taken seriously.

*A-aln* (Alienation) Adolescents report considerable social distance from others. They do not believe that they are liked by others and they do not get along well with others. They believe that others are out to get them and are unkind to them.

*A-biz* (Bizarre Mentation) > 64 Adolescents report overtly psychotic symptoms such as paranoid ideation and hallucinations (*A-biz1*) and various peculiar and strange experiences (*A-biz2*). Psychotropic medications may be indicated, as well as hospitalization.

*A-ang* (Anger) > 64 Adolescents report a number of explosive tendencies such as hitting and smashing things (*A-ang1*), as well as being irritable, grouchy, and impatient (*A-ang2*). Assertiveness training and anger-control techniques should be implemented as part of treatment.

*A-cyn* (Cynicism) > 64 Adolescents expect others are only interested in their own welfare (*A-cyn1*) and are suspicious of others' motives (*A-cyn2*). Establishing a trusting relationship is a difficult process, but imperative, if any progress is to be made in therapy.

*A-con* (Conduct Problems) > 64 Adolescents have attitudes similar to adolescents who break the law (*A-con1*), even if they do not have conduct problems. They report stealing things and other problematic conduct (*A-con2*). Group interventions with similar adolescents will

be most productive.

(continued)

## 242 Self-Report Inventories

**Table 7.5 (Continued)**

### Scale Potential Issues

*A-lse* (Low Self-Esteem) > 64 Adolescents have very low opinions of themselves (*A-lse1*), and they are uncomfortable if people say nice things about them (*A-lse2*). They give in easily to others. Interventions need to be very supportive and allow ample time for change.

*A-las* (Low Aspirations) > 64 Adolescents are not interested in being successful. They report difficulty starting things and quickly give up when things go wrong. They let other people solve problems, and they avoid facing difficulties. They believe that others block their success; others say they are lazy.

*A-sod* (Social Discomfort) > 64 Adolescents are very uneasy around others and are happier by themselves (*A-sod1*). They see themselves as shy and uncomfortable in social situations (*A-sod2*). They need to be supported and encouraged to participate in treatment until they are comfortable interacting with others.

*A-fam* (Family Problems) > 64 Adolescents report considerable familial discord (*A-fam1*). Their families are reported to lack love, support, and companionship. They feel alienated from and unattached to their family (*A-fam2*). Involvement of the family system in treatment may be important unless the adolescent needs to be emancipated from them.

*A-sch* (School Problems) > 64 Adolescents have numerous school problems such as poor grades, suspensions, truancy, negative attitudes toward teachers, and a dislike of school. They do not participate in school activities.

*A-trt* (Negative Treatment

Indicators) > 64

Adolescents are unmotivated and feel unable to help themselves (*A-trt1*). They dislike going to doctors and they believe that they should not discuss their personal problems with others (*A-trt2*). They prefer to take drugs or medicine because talking about problems does not help them. Adolescents with depressive mood disorders will elevate *TRT* because it is primarily a measure of general distress, so clinicians need to be cautious about interpreting *TRT* in a characterological manner.

and explained more fully in Williams et al. (1992). Sherwood, Ben-Porath, and Williams (1997) developed component scales for the MMPI-A content scales again similar to the MMPI-2 content component scales (Ben-Porath & Sherwood, 1993). Arita and Baer (1998) examined the convergent and discriminant validity of the content scales of Anxiety, Depression, Health Concerns, Alienation, Anger, Conduct Problems, and Social Discomfort in adolescent inpatients. Neither the Anxiety nor the Depression scale discriminated among the various other measures and appear to be more general measures of nonspecific distress. The other content scales that they studied demonstrated reasonable convergent and discriminant validity.

### Minnesota Multiphasic Personality Inventory—Adolescent 243

#### Supplementary Scales

There are many supplementary scales to assist the clinician in interpreting the MMPI-A codetype. Three of these supplementary scales will be familiar to the clinician who has used the MMPI-2 and will have generally similar interpretations: MacAndrew Alcoholism-Revised (*MAC-R*) scale and Welsh's Anxiety (*A*) and Repression (*R*) scales. These scales have had a few minor changes made in them at the item level. Three new scales are represented in the supplementary scales on the MMPI-A: Alcohol/Drug Problem Acknowledgment (*ACK*: Weed, Butcher, & Williams, 1994); Alcohol/Drug Proneness (*PRO*: Weed et al., 1994); and Immaturity (*IMM*: Archer, Pancoast, & Gordon, 1994). The Alcohol/Drug Problem Acknowledgment (*ACK*) and Alcohol/Drug Proneness (*PRO*)

scales were developed in the same manner as the Addiction Admission (AAS) and Addiction Potential (APS) scales on the MMPI-2, respectively. McNulty et al. (1997) developed scales to measure the Personality Psychopathology Five (PSY-5) in adolescents from the MMPI-A items; and Bolinsky, Arnau, Archer, and Handel (2004) developed facet subscales for these PSY-5 scales.

#### *MMPI-A Structural Summary*

The MMPI-A Structural Summary (Archer & Krishnamurthy, 1994) has been created to provide a conceptual framework for the eight factors that have been identified in the MMPI-A item pool (Archer, Belevich, & Elkins, 1994; Archer & Krishnamurthy, 1997): general maladjustment; immaturity; disinhibition/excitatory potential; social discomfort; health concerns; naivete; familial alienation; and psychoticism. These factors are listed in relative order of the amount of variance for which they accounted. Krishnamurthy and Archer (1999) showed that simply counting the number of scales within each factor—and if a majority of them are elevated to a T score of 65 or higher to consider a factor as being elevated—produced similar results as computing the average T score for all scales within the factor.

The MMPI-A Structural Summary organizes all the MMPI-A scales and subscales within each of these factors based on their degree of correlation with each factor. Archer, Krishnamurthy, and Jacobson (1994) and Archer (1997) describe the MMPI-A Structural Summary in more detail and also provide examples of its use with clinical cases.

#### **Critical Items**

Archer and Jacobson (1993) investigated the frequency with which the Koss and Butcher (1973) critical items were endorsed by adolescents in the MMPI-A normative group and clinical samples. Both groups of adolescents endorsed these critical items more often than adults. However, the adolescents in clinical settings did not endorse these items more often than the normal adolescents, which should serve as a caution to clinicians who might be tempted to interpret these critical items in adolescents.

Forbey and Ben-Porath (1998) developed a critical item set specifically for adolescents based on the MMPI-A items. They used a multiple-step process that involved comparing endorsement rates of the individual items among normal and various clinical samples and review of potential critical items by judges experienced with adolescents and the MMPI-A. This process resulted in 82 items that Forbey and Ben-Porath rationally placed

**244 Self-Report Inventories** in 15 content groupings: Aggression; Anxiety; Cognitive Problems; Conduct Problems; Depressed/Suicidal Ideation; Eating Problems; Family Problems; Hallucinatory Experiences; Paranoid Ideation; School Problems; Self Denigration; Sexual Concerns; Somatic Complaints; Substance Use/Abuse; and Unusual Thinking. There have been no published studies on these MMPI-A critical items, but they appear to be much more suitable for adolescents than the MMPI-2 critical item sets that were designed for adults.

#### **Summary**

Archer (2005), Archer and Krishnamurthy (2002), Fowler et al. (2000), and Williams et al. (1992) have provided interpretive guidelines for the individual validity, clinical, content, and supplementary scales of the MMPI-A. These references are required reading for any clinician who is interpreting the MMPI-A. Archer (1997) and Butcher, Graham, and Ben-Porath (1995) have provided general guidelines for research with the MMPI-A that should be consulted by anyone who is interested in working in this area.

#### **APPLICATIONS**

The applications of the MMPI-A have been generally limited to the identification of psychopathology within a clinical setting and delinquent behaviors in criminal justice settings.

The former application was the primary purpose of the original MMPI and also is the primary application of the MMPI-A. Consequently, this application has been the focus of the entire chapter and will not be pursued further here.

The identification of delinquent behavior was the focus of Hathaway and Monachesi's (1963) longitudinal investigations. They found that elevations on Scales 4 (*Pd*: Psychopathic Deviate), 8 (*Sc*: Schizophrenia), and 9 (*Ma*: Hypomania) as "excitatory" scales increased the risk of delinquent behavior, while elevations on Scales 2 (*D*: Depression), 5 (*Mf*:

Masculinity-Femininity), and 0 (*Si*: Social Introversion) as “inhibitory” decreased the risk of delinquent behavior. The effects of these excitatory and inhibitory scales were relatively small, however, with approximately a 7% increase or decrease from the base rate for delinquent behavior of 35%, respectively. Several subsequent studies (e.g., Huesmann, Lefkowitz, & Eron, 1978; Pena et al., 1996; Rempel, 1958) supported these general findings of the effects of the excitatory and inhibitory scales on delinquent behaviors.

More recent studies of using the MMPI-A to identify delinquent behavior have found somewhat divergent results. Cashel, Rogers, Sewell, and Holliman (1998) reported the external correlates of the MMPI-A clinical scales. Scales 4 (*Pd*: Psychopathic Deviate), 7 (*Pt*: Psychasthenia), and 8 (*Sc*: Schizophrenia) were related significantly to most of the correlates, while Scales 5 (*Mf*: Masculinity-Femininity) and 0 (*Si*: Social Introversion) were related to none of them. The correlates of Scales 2 (*D*: Depression) and 9 (*Ma*: Hypomania) were as expected, depressive and excitatory variables, respectively. However, Scale 3 (*Hy*: Hysteria) was related to cruelty toward others and suicidal ideation, neither of which would be expected from either the adult or adolescent research. They also reported that White delinquents tended to have higher scores on the clinical scales than African American or Mexican American delinquents. That is, there was no tendency for the MMPI-A to overpathologize African American or Mexican American delinquents.

#### **Minnesota Multiphasic Personality Inventory—Adolescent 245**

Morton, Farris, and Brenowitz (2002) compared a large sample of male delinquents with the MMPI-A normative group. Lower scores on Scale 5 (*Mf*: Masculinity-Femininity) characterized the delinquent boys and was the most effective Scale to differentiate the two groups. Elevations on Scales 4 (*Pd*: Psychopathic Deviate) and 6 (*Pa*: Paranoia) also were common in the delinquent boys.

Thus, it appears that elevations on Scales 4 (*Pd*: Psychopathic Deviate), 8 (*Sc*: Schizophrenia), and 9 (*Ma*: Hypomania) as excitatory scales must be considered in evaluating adolescents for increased risk for delinquent behaviors. In addition, elevations on Scales *F* (Infrequency), 3 (*Hy*: Hysteria), and 6 (*Pa*: Paranoia) and lower scores on Scale 5 (*Mf*: Masculinity-Femininity) must be considered as potential risk factors.

## **PSYCHOMETRIC FOUNDATIONS**

### **Demographic Variables**

#### *Age*

There were substantial differences in performance between adolescents and adults on the original MMPI that resulted in the development of the MMPI-A specifically for adolescents. Within the limited age range of the MMPI-A (14 to 18), there is little reason to expect much difference as the function of age. Younger adolescents (< 15) would be expected to have more trouble reading the MMPI-A items because of their developmental level in reading skills.

#### *Gender*

Gender does not create any general issues in MMPI-A interpretation because separate norms (profile forms) are used for boys and girls. Any gender differences in how adolescents responded to the items on each scale are removed when the raw scores are converted to T scores.

#### *Education*

The limited levels of education represented in ages 14 to 18 also limit any potential effect of education on the MMPI-A.

#### *Ethnicity*

Only very limited research has looked specifically at the effects of ethnicity on either the original MMPI or the MMPI-A in adolescents. This limited research in adolescents is somewhat surprising given the vast number of studies of ethnicity with the MMPI and MMPI-2 in adults. The effects of ethnicity on the MMPI and MMPI-2 in adults were summarized in Chapter 6 (see p. 200) and should be perused to provide a context for looking at this issue in adolescents.

The MMPI-A has been examined in three studies with normal Hispanic adolescents (Gumbiner, 1998; Mendoza-Newman, 1999; Negy, Leal-Puente, Trainor, & Carlson, 1997), all of which found minimal differences when compared with the MMPI-A normative group.

Negy et al. (1997) found that MMPI-A profiles did vary as a function of acculturation and socioeconomic status (SES). Gumbiner (1998) found that boys scored higher than girls

#### **246 Self-Report Inventories**

on several scales and she concluded that the scale elevations might be related to SES. Mendoza-Newman (1999) examined the relationship between acculturation and SES on Scales *L* (Lie) and 5 (*Mf*: Masculinity-Femininity) and found no relationship between acculturation or SES as individual variables. However, there was a significant negative correlation between the combination of acculturation and SES and Scale *L* (Lie), but not for Scale 5 (*Mf*: Masculinity-Femininity). All three studies suggest that normal Hispanic adolescents score very similarly to the MMPI-A normative group and underscore the importance of considering the importance of acculturation and SES in looking at the effects of ethnicity when any differences are found.

Gomez, Johnson, Davis, and Velasquez (2000) found no differences on any of the MMPI-A scales between small samples of African American and Mexican American first-time offenders. They did find that the African American offenders produced more within-normal-limit profiles (50% versus 25%).

Multivariate regressions of age, gender, and ethnicity on the MMPI-A scales (Schinka, Elkins, & Archer, 1998) have shown that the total percentage of variance accounted for by these factors did not exceed 10% for any of the MMPI-A scales. The largest percentage of variance (9.50%) was found on the Bizarre Mentation (*BIZ*) scale. Such small percentages of variance are unlikely to impact the interpretation of most MMPI-A profiles. Ethnicity (White versus nonwhite) accounted for 7.73% of the variance in the *F* (Infrequency) scale and typically was in the range of 1% to 3% on most of the MMPI-A scales. Gender only accounted for 0.79% of the variance on Scale 5 (*Mf*: Masculinity-Femininity) on the MMPI-A, in contrast to the MMPI-2 in which slightly over 50% of the variance was accounted for by gender.

It appears that demographic variables will have minimal impact on the MMPI-A profile in most individuals. It is important to monitor the validity of the MMPI-A profile more closely in younger adolescents because of the potential limitations on their reading level because of their level of education.

#### **Reliability**

The *MMPI-A Manual* (Butcher et al., 1992, Appendix E) reports the reliability data for 45 boys and 109 girls who were retested 1 week later. The test-retest correlations ranged from .65 to .84 across the 10 clinical scales and averaged .74. The test-retest correlations ranged from .62 to .82 across the 15 content scales and averaged .73. The standard error of measurement is 4 to 6 T points for the clinical and content scales; that is, the individual's true score on the clinical and content scales will be within  $\pm 4$  to 6 T points two-thirds of the time.

#### **Codetype Stability**

There are no data on the stability of MMPI-A codetypes. It would be assumed that MMPI-A codetypes, at best, would be no more stable than MMPI-2 codetypes (see pp. 200–201), and probably would be less stable than MMPI-2 codetypes. Consequently, any interpretation should describe the adolescent's *current* status rather than be used for any long-term predictions.

#### **Minnesota Multiphasic Personality Inventory—Adolescent 247**

#### **REFERENCES**

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Archer, R. P. (1984). Use of the MMPI with adolescents: A review of salient issues. *Clinical Psychology Review, 4*, 241–251.
- Archer, R. P. (1987). *Using the MMPI with adolescents*. Hillsdale, NJ: Erlbaum.
- Archer, R. P. (1997). Future directions for the MMPI-A: Research and clinical issues. *Journal of Personality Assessment, 68*, 95–109.
- Archer, R. P. (2005). *MMPI-A: Assessing adolescent psychopathology* (3rd ed.). Mahwah, NJ: Erlbaum.
- Archer, R. P., Belevich, J. K. S., & Elkins, D. E. (1994). Item-level and scale-level factor structures of the MMPI-A. *Journal of Personality Assessment, 62*, 332–345.
- Archer, R. P., Handel, R. W., Lynch, K. D., & Elkins, D. E. (2002). MMPI-A validity scale uses and

- limitations in detecting varying levels of random responding. *Journal of Personality Assessment*, 78, 417–431.
- Archer, R. P., Handel, R. W., & Mason, J. A. (2006). [MMPI-A data from a community outpatient setting]. Unpublished raw data.
- Archer, R. P., & Jacobson, J. M. (1993). Are critical items “critical” for the MMPI-A? *Journal of Personality Assessment*, 61, 547–556.
- Archer, R. P., & Krishnamurthy, R. (1994). A structural summary approach for the MMPI-A: Development and empirical correlates. *Journal of Personality Assessment*, 63, 554–573.
- Archer, R. P., & Krishnamurthy, R. (1997). MMPI-A scale-level factor structure: Replication in a clinical sample. *Assessment*, 4, 337–349.
- Archer, R. P., & Krishnamurthy, R. (2002). *Essentials of MMPI-A assessment*. Hoboken, NJ: Wiley.
- Archer, R. P., Krishnamurthy, R., & Jacobson, J. M. (1994). *MMPI-A casebook*. Odessa, FL: Psychological Assessment Resources.
- Archer, R. P., Panoast, D. L., & Gordon, R. A. (1994). The development of the MMPI-A Immaturity scale: Findings for normal and clinical samples. *Journal of Personality Assessment*, 62, 145–156.
- Arita, A. A., & Baer, R. A. (1998). Validity of selected MMPI-A Content scales. *Psychological Assessment*, 10, 59–63.
- Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment*, 68, 139–151.
- Ben-Porath, Y. S., & Sherwood, N. E. (1993). *The MMPI-2 Content Component scales: Development, psychometric characteristics, and clinical application*. Minneapolis: University of Minnesota Press.
- Bolinsky, P. K., Arnau, R. C., Archer, R. P., & Handel, R. W. (2004). A replication of the MMPI-A PSY-5 scales and development of facet subscales. *Assessment*, 11, 40–48.
- Butcher, J. N., Graham, J. R., & Ben-Porath, Y. S. (1995). Methodological problems and issues in MMPI, MMPI-2, and MMPI-A research. *Psychological Assessment*, 7, 320–329.
- Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 Content scales*. Minneapolis: University of Minnesota Press.
- Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., et al. (1992). *Minnesota Multiphasic Personality Inventory—Adolescent (MMPI-A): Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.
- Cashel, M. L., Rogers, R., Sewell, K. W., & Holliman, N. B. (1998). Preliminary validation of the MMPI-A for a male delinquent sample: An investigation of clinical correlates and discriminant validity. *Journal of Personality Assessment*, 71, 49–69.
- Colligan, R. C., & Offord, K. P. (1989). The aging MMPI: Contemporary norms for contemporary teenagers. *Mayo Clinic Proceedings*, 64, 3–27.
- 248 Self-Report Inventories**
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook: Vol. I. Clinical interpretation (Rev. ed.)*. Minneapolis: University of Minnesota Press.
- Forbey, J. D., & Ben-Porath, Y. S. (1998). *A critical item set for the MMPI-A*. Minneapolis: University of Minnesota Press.
- Fowler, R. A., Butcher, J. N., & Williams, C. L. (2000). *Essentials of MMPI-2 and MMPI-A interpretation* (2nd ed.). Minneapolis: University of Minnesota Press.
- Gomez, F. C., Johnson, R., Davis, Q., & Velasquez, R. J. (2000). MMPI-A performance of African and Mexican American adolescent first-time offenders. *Psychological Reports*, 87, 309–314.
- Gottesman, I. I., Hanson, D. R., Kroeker, T. A., & Briggs, P. F. (1987). New MMPI normative data and power-transformed T-score tables for the Hathaway-Monachesi Minnesota cohort of 14,019 15-year-olds and 3,674 18-year-olds. In R. P. Archer (Ed.), *Using the MMPI with adolescents* (pp. 241–297). Hillsdale, NJ: Erlbaum.
- Gough, H. G. (1950). The *F* minus *K* dissimulation index for the MMPI. *Journal of Consulting Psychology*, 14, 408–413.
- Greene, R. L., & Brown, R. C. (2006). *MMPI-2 adult interpretive system* (3rd ed.). Lutz, FL: Psychological Assessment Resources.
- Gumbiner, J. (1998). MMPI-A profiles of Hispanic adolescents. *Psychological Reports*, 82, 659–672.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): Pt. I. Construction of the schedule. *Journal of Psychology*, 10, 249–254.
- Hathaway, S. R., & Monachesi, E. D. (1963). *Adolescent personality and behavior: MMPI patterns of normal, delinquent, dropout, and other outcomes*. Minneapolis: University of Minnesota Press.
- Huesmann, L. R., Lefkowitz, M. M., & Eron, L. D. (1978). Sum of MMPI scales *F*, 4, and 9 as a measure of aggression. *Journal of Consulting and Clinical Psychology*, 46, 1071–1078.
- Koss, M. P., & Butcher, J. N. (1973). A comparison of psychiatric patients’ self-report with other sources of clinical information. *Journal of Research in Personality*, 7, 225–236.

- Krakauer, S. Y., Archer, R. P., & Gordon, R. A. (1993). The development of Items-easy (*Ie*) and Items-difficult (*Id*) scales for the MMPI-A. *Journal of Personality Assessment*, *60*, 561–571.
- Krishnamurthy, R., & Archer, R. P. (1999). A comparison of two interpretive approaches for the MMPI-A structural summary. *Journal of Personality Assessment*, *73*, 245–259.
- Lachar, D. (1974). *The MMPI: Clinical assessment and automated interpretation*. Los Angeles: Western Psychological Services.
- Lachar, D., Klinge, V., & Grisell, J. L. (1976). Relative accuracy of automated MMPI narratives generated from adult norm and adolescent norm profiles. *Journal of Consulting and Clinical Psychology*, *44*, 20–24.
- Marks, P. A., Seeman, W., & Haller, D. L. (1974). *The actuarial use of the MMPI with adolescents and adults*. Baltimore: Williams & Wilkins.
- McGrath, R. E., Pogge, D. L., Stein, L. A. R., Graham, J. R., Zaccario, M., & Piacentini, T. (2000). Development of an Infrequency-Psychopathology scale for the MMPI-A: The *Fp* - A scale. *Journal of Personality Assessment*, *74*, 282–295.
- McNulty, J. L., Harkness, A. R., Ben-Porath, Y. S., & Williams, C. L. (1997). Assessing the Personality Psychopathology Five (PSY-5) in adolescents: New MMPI-A scales. *Psychological Assessment*, *9*, 250–257.
- Meehl, P. E. (1957). When should we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*, 268–273.
- Mendoza-Newman, M. C. (1999). *Level of acculturation, socioeconomic status, and the MMPI-A performance of a non-clinical Hispanic adolescent sample*. Unpublished doctoral dissertation, Pacific Graduate School of Psychology, Palo Alto, CA.
- Morton, T. L., Farris, K. L., & Brenowitz, L. H. (2002). MMPI-A scores and high points of male juvenile delinquents: Scales, 4, 5, and, 6, as markers of juvenile delinquency. *Psychological Assessment*, *14*, 311–319.
- Minnesota Multiphasic Personality Inventory—Adolescent 249**
- Negy, C., Leal-Puente, L., Trainor, D. J., & Carlson, R. (1997). Mexican American adolescents' performance on the MMPI-A. *Journal of Personality Assessment*, *69*, 205–214.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598–609.
- Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires: Current issues in theory and measurement* (pp. 143–165). Berlin, Germany: Springer-Verlag.
- Pena, L. M., Megargee, E. I., & Brody, P. (1996). MMPI-A patterns of male juvenile delinquents. *Psychological Assessment*, *8*, 388–397.
- Pinsoneault, T. B. (2005). Detecting random, partially random, and nonrandom MMPI-A protocols. *Psychological Assessment*, *17*, 476–480.
- Rempel, P. P. (1958). The use of multivariate statistical analysis of MMPI scores in the classification of delinquent and nondelinquent high school boys. *Journal of Consulting Psychology*, *22*, 17–23.
- Schinka, J. A., Elkins, D. E., & Archer, R. P. (1998). Effects of psychopathology and demographic characteristics on MMPI-A scale scores. *Journal of Personality Assessment*, *71*, 295–305.
- Shaevel, B., & Archer, R. P. (1996). Effects of MMPI-2 and MMPI-A norms on T-score elevations for 18-year-olds. *Journal of Personality Assessment*, *67*, 72–78.
- Sherwood, N. E., Ben-Porath, Y. S., & Williams, C. L. (1997). *The MMPI-A Content Component scales: Development, psychometric characteristics, and clinical applications*. Minneapolis: University of Minnesota Press.
- Stein, L. A. R., Graham, J. R., & Williams, C. L. (1995). Detecting fake-bad MMPI-A profiles. *Journal of Personality Assessment*, *65*, 415–427.
- Weed, N. C., Butcher, J. N., & Williams, C. L. (1994). Development of MMPI-A Alcohol/Drug Problem scales. *Journal of Studies on Alcohol*, *55*, 296–302.
- Williams, C. L. (1986). MMPI profiles for adolescents: Interpretive strategies and treatment considerations. *Journal of Child and Adolescent Psychotherapy*, *3*, 179–193.
- Williams, C. L., Ben-Porath, Y. S., & Hevern, V. W. (1994). Item level improvements for use with the MMPI with adolescents. *Journal of Personality Assessment*, *63*, 284–293.
- Williams, C. L., Butcher, J. N., Ben-Porath, Y. S., & Graham, J. R. (1992). *MMPI-A Content scales: Assessing psychopathology in adolescents*. Minneapolis: University of Minnesota Press.
- Wimbish, L. G. (1984). *The importance of appropriate norms for the computerized interpretations of adolescent MMPI profiles*. Unpublished doctoral dissertation, Ohio State University, Columbus.

## CHAPTER 5

### **The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS)**

DAVID J.A. DOZOIS AND ROGER COVIN  
BECK DEPRESSION INVENTORY-SECOND EDITION  
(BDI-II) 51

#### **Test Description 51**

#### **Theoretical Basis 51**

#### **Test Development 51**

#### **Psychometric Characteristics 52**

#### **Range of Applicability and Limitations 54**

#### **Age and Cross-Cultural Factors 55**

#### **Accommodation for Populations with Disabilities 56**

#### **Legal and Ethical Considerations 56**

#### **Computerization 56**

#### **Current Research Status 57**

BECK HOPELESSNESS SCALE (BHS) 57

#### **Test Description 57**

#### **Theoretical Basis and Test Development 57**

#### **Psychometric Characteristics 58**

#### **Range of Applicability and Limitations 59**

#### **Age and Cross-Cultural Factors 59**

#### **Accommodation for Populations with Disabilities 60**

#### **Legal and Ethical Considerations 60**

#### **Computerization 60**

#### **Current Research Status 60**

BECK SCALE FOR SUICIDE IDEATION (BSS) 61

#### **Test Description 61**

#### **Test Development 61**

#### **Psychometric Characteristics 61**

#### **Range of Applicability and Limitations 62**

#### **Accommodation for Populations with Disabilities 62**

#### **Legal and Ethical Considerations 62**

#### **Current Research Status 62**

THE BECK SCALES IN CLINICAL PRACTICE 63

SUMMARY AND FUTURE DEVELOPMENTS 64

REFERENCES 64

In addition to his substantial contributions to the development and validation of cognitive theory and therapy (see Dobson & Dozois, 2001), Aaron T. Beck has, over the past 40 years, established himself firmly in the area of test construction.

Along with his colleagues, Beck has developed some of the most well known and frequently utilized self-report instruments available for research and practice. These measures cover depressive (Beck, Rush, Shaw, & Emery, 1979; Beck, Steer, & Brown, 1996; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and anxious (Beck & Steer, 1990) symptomatology, hopelessness (Beck & Steer, 1988), suicidal ideation

*Acknowledgments:* During the preparation of this chapter, David J.A. Dozois was supported by a fellowship from the Ontario Mental Health Foundation, and Roger Covin was assisted by a studentship from the Natural Sciences and Engineering Research Council of Canada. The authors gratefully acknowledge this support.

(Beck & Steer, 1991), dysfunctional attitudes (Weissman & Beck, 1978), self-concept (Beck, Steer, Epstein, & Brown,

1990), and personality (Beck, Epstein, Harrison, & Emery, 1983). More recently, the Beck Youth Inventories, which purport to evaluate emotional and social impairment in youth, have been produced (Beck, Beck, & Jolly, 2001).

This chapter describes the Beck Depression Inventory-Second edition (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). Given that the BDI-II is the most widely used of these measures, coupled with the fact that comprehensive reviews of this revised instrument have yet to appear in the literature, the primary focus of this chapter concerns the examination of the BDI-II.

However, the remaining scales that we review are used widely as well, especially in the assessment of depression. Although we do not review the Beck Anxiety Inventory (BAI), which is another of the most commonly used Beck scales, readers are directed to some recent review papers (see Steer & Beck 1997; Wilson, de Beurs, Palmer, & Chambless, 1999).

We begin with a review of the principal features, test development, psychometric characteristics, research status, and applicability of each of these instruments. We also discuss the limitations of these measures, mention age and cross-cultural factors, highlight accommodations made for persons with disabilities, address legal and ethical issues, and summarize each instrument's current research status. Following this examination, we underscore how these measures may be used in clinical practice.

## **BECK DEPRESSION INVENTORY-SECOND EDITION (BDI-II)**

### **Test Description**

The BDI-II is a 21-item self-administered inventory designed to measure the intensity of depressive symptoms in psychiatric and nonpsychiatric populations of both adults and adolescents (Beck et al., 1996). Each item contains a header that is intended to focus the examinee on the general purpose of the response options. Directly below this label are four statements listed in order of increasing severity. Respondents are instructed to choose the alternative that best describes how they felt during the "past two weeks, including today."

A sample item follows:

#### *5. Guilty Feelings*

0 I don't feel particularly guilty.

1 I feel guilty over many things I have done or should have done.

2 I feel guilty most of the time.

3 I feel guilty all of the time.

Items are rated on a 4-point scale (0 to 3) and total scores are obtained by tallying the ratings for all 21 items. Scores range from 0 to 63, with higher scores reflecting increased depressive severity. For instance, scores ranging between 0 and 13 are indicative of "minimal depression"; scores that fall between 14 and 19 are considered to reflect a "mild" level of depression; scores of 20 to 28 are considered "moderate"; and a score ranging from 29 to 63 is labeled "severe." Researchers studying dysphoria or depression in analogue samples should consult Dozois, Dobson, and Ahnberg (1998) for recommended cutoff scores for college populations. The BDI-II requires approximately 5 to 10 minutes to complete

and may be administered to individuals 13 to 80 years of age. Although this instrument is typically self-administered, it can also be administered orally with only slight modification to the instructions.

### **Theoretical Basis**

The BDI-II items were specifically selected to evaluate the symptoms and attitudes characteristic of the phenomenology of depression rather than to adhere to any particular theory (Beck et al., 1996). Additionally, although the BDI-II's items are congruent with the criteria outlined in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*; American Psychiatric Association, 1994), the BDI-II is intended to identify the severity of symptoms and not nosological depression. Thus, the BDI-II should be supplemented with other information for a comprehensive assessment and diagnosis of depression.

### **Test Development**

The original BDI (Beck et al., 1961) was designed to be administered in an interviewer-assisted fashion by trained professionals (Beck et al., 1996; Katz, Katz, & Shaw, 1999). The BDI items were initially drawn from clinical observations and typical descriptions of symptoms provided by depressed patients. These descriptions were refined and assembled to yield a 21-item measure with response choices ranging from 4 to 7 per item. Each statement was given a weight between 0 and 3 points. The original BDI asked respondents to describe "the way you feel today, that is, right now" (Beck & Steer, 1984). The 1978 revision, which was published as the BDI-IA in Beck et al. (1979), permitted simpler administration and scoring (see Beck & Steer, 1984). For example, the items were standardized so that they would each involve only four possible choices, alternative ways of asking the same questions were eliminated, and the language of the items was clarified (e.g., the use of double negatives was avoided). The BDI-IA was designed as a self-report index and the temporal focus was on the "past week, including today."

The BDI-IA demonstrated adequate reliability and validity (see Beck, Steer, & Garbin, 1988, for an elaborate review). However, it became apparent that this instrument did not correspond adequately to current diagnostic symptom criteria, and questions were raised regarding its content validity. For example, the BDI-IA covered only six of the nine symptoms highlighted in *DSM-IV*. In addition, this instrument only permitted the assessment of insomnia and decreases in appetite and weight rather than reversed neurovegetative symptoms (Moran & Lambert, 1983; Vredenburg, Krames, & Flett, 1985). Several changes were made in the BDI-II to increase its correspondence with *DSM-IV*: Four items (i.e., "body image change," "work difficulty," "weight loss," and "somatic preoccupation") were eliminated, 17 response options were reworded, 2 items were relocated, 4 new items were constructed (i.e., "agitation," "worthlessness," "loss of energy," and "concentration difficulty"), item labels were provided to make the intention of each item more explicit, and the time frame was extended 2 weeks (see Beck, Steer, Ball, & Ranieri,

1996; Beck et al., 1996).

### **Psychometric Characteristics**

The BDI-II represents a significant change to earlier editions of this instrument. Therefore, it is important for researchers and clinicians to be familiar with the psychometric properties of this particular instrument and to be acquainted with how this measure corresponds to its previous editions.

#### **Reliability**

A number of studies have now documented that the BDI-II exhibits high internal consistency. Table 5.1 presents the coefficient alphas across 13 studies. The average coefficient alpha was .91 (range  $\_$  .89 to .94). Regardless of the population investigated, the internal reliability of the BDI-II appears to be excellent.

There is a paucity of information on the test-retest reliability of the BDI-II. A 1-week test-retest reliability coefficient of .93 ( $n$   $\_$  26 outpatients) was reported in the test manual (Beck et al., 1996). Aside from this finding, it is necessary to rely on previous research using the BDI and BDI-IA to address the temporal stability of this instrument. Beck, Steer, and Garbin (1988) reported the findings from 10 different studies. The test-retest reliability estimates ranged from .48 to .86 in psychiatric patients and from .60 to .83 in nonpsychiatric samples. The assessment periods varied widely in this review from a few hours to a few months. However, the higher overall correlation in nonclinical, relative to clinical, samples and the fact that higher correlations are found when shorter test-retest periods are used support the reliability of this measure (Beck, Steer, & Garbin, 1988; Richter, Werner, Heerlein, Kraus, & Sauer, 1998). Across 20 studies, Yin and Fan (2000) found an average test-retest reliability coefficient of .72.

The use of test-retest reliability on a measure that is supposed to measure a construct reliably but also be sensitive to treatment change is difficult. On one hand, a short temporal period between assessments may overestimate reliability because participants are better able to remember how they responded previously. On the other hand, reliability may be underestimated when using a longer time frame between assessments, because true changes in depressive symptoms may have occurred (e.g., improvement due to treatment; see Richter et al., 1998). Notwithstanding the fact that the BDIII should be both sensitive to clinical change and reasonably stable over time (Boyle, 1985), a number of researchers have reported that BDI/BDI-IA scores decrease significantly upon reassessment even without the introduction of treatment (Ahava, Iannone, Grebstein, & Schirling, 1998; Hatzenbuehler, Parpal, & Matthews, 1983; Yin & Fan, 2000; Zimmerman, 1986). Ahava et al. (1998) tested the stability of the BDI over multiple assessment periods and found a 40% reduction in scores over 8 weeks. These authors argued that this reduction was due to measurement error rather than authentic changes in depressive severity. It is possible that the BDI-II will show greater temporal stability because of the increased time frame in the instructions, but this remains to be demonstrated empirically.

**TABLE 5.1 Internal Consistency Estimates for the Beck Depression Inventory-II**

Reference Sample

Coefficient

Alpha

*Psychiatric*

Beck, Steer, Ball, &

Ranieri (1996)

140 adult outpatients 0.91

Beck, Steer, & Brown

(1996)

500 adult outpatients 0.92

Buckley et al. (2001) 416 substance-abusing males 0.91

Steer, Beck, & Brown

(1997)

210 adult outpatients 0.92

Steer et al. (1999) 210 adult outpatients 0.90

Steer, Clark, Beck, &

Ranieri (1999)

840 adult outpatients 0.92

Steer, Kumar, Ranieri, &

Beck (1998)

210 adolescent outpatients 0.92

*Nonpsychiatric*

Arnau et al. (2001) 340 primary care patients

**Validity**

The content validity of the BDI-II appears to be excellent.

The BDI-II now covers the major content domains of depression, including sadness, pessimism, beliefs of being a failure, loss of pleasure, feelings of guilt, punishment feelings, self-dislike, self-criticalness, suicidal thoughts or wishes, experiences of crying, agitation, anhedonia, indecisiveness, feelings of worthlessness, lack of energy, altered sleep patterns (i.e., hypersomnia and insomnia), irritability, increases or decreases in appetite, concentration difficulties, fatigue, and loss of interest in sex.

The convergent and divergent validity of the BDI-II also appears to be well supported. The BDI-II correlates significantly with other indices of depression and depression-related constructs, including the BDI-IA ( $r = .93$ ; Beck et al., 1996; Dozois et al., 1998), the Hamilton Rating Scale for Depression ( $r = .71$ ), and the BHS ( $r = .68$ ; Beck et al., 1996).

BDI-II scores also correlate more highly with measures of depression than with measures of anxiety (Beck et al., 1996; Osman et al., 1997; Steer, Ball, Ranieri, & Beck, 1997). For example, Steer et al. (1997) found that the BDI-II was more strongly associated with the Depression subscale of the Symptom Check List-90-Revised ( $r = .89$ ) than with the Anxiety subscale of this same instrument ( $r = .71$ ). The divergent validity of the BDI-II is also upheld by low correlations between this instrument and age, sex, ethnicity, and social desirability (Beck, Steer, Ball, & Ranieri, 1996; Osman et al., 1997; Steer & Clark, 1997). One criticism of the BDI-II that also applies to many depression assessment instruments (see Dozois & Dobson, 2002) is that this measure correlates highly with other measures of anxiety and may not discriminate adequately between depression and other affective

states. Lovibond and Lovibond (1995) argued that this problem may be a result of item overlap. However, given the high rates of comorbidity between depression and other emotional disorders, it is difficult to ascertain whether this inability to differentiate between such groups is a function of the measure itself or the construct being assessed (e.g., the heterogeneous nature of depression and the polythetic criteria in the *DSM-IV*).

The BDI-II does appear to differentiate well between depressed and nondepressed persons (Ambrosini, Metz, Bianchi, Rabinovich, & Undie, 1991; Arnau, Meagher, Norris, & Bramson, 2001; Beck et al., 1996; Martinsen, Friis, Hoffart, 1995). Other research has demonstrated that the BDI-II is also able to distinguish among varying levels of depressive severity (Steer, Brown, Beck, & Sanderson, 2001), and between mood disorders and other forms of psychopathology, including anxiety (Beck et al., 1996). Not surprisingly, this instrument does not differentiate among varying types of mood disorders (e.g., major depressive disorder and dysthymia; Richter et al., 1998). Although the BDI-II coincides with *DSM-IV* symptomatology, it was intended to be used as an index of severity, not necessarily as an indicator of nosological depression. However, further study of the diagnostic specificity of this instrument is warranted (see Dori & Overholser, 2000).

Evidence for the construct validity of the BDI-II also stems from factor analytic studies. Table 5.2 displays the number of factors described in the manual and by nine additional studies. As shown in this table, a stable factor structure exists for the BDI-II (Arnau et al., 2001; Beck et al., 1996; Buckley, Parker, & Heggie, 2001; Dozois et al., 1998; Steer, Ball, Ranieri, & Beck, 1999; Steer & Clark, 1997; Steer, Kumar, Ranieri, & Beck, 1998; Whisman, Perez, & Ramel, 2000). With few exceptions, two main factors appear to emerge consistently in the literature. In clinical samples, the two factors typically represent the somatic-affective and cognitive aspects of depression. A similar factor structure is found in nonclinical samples, but the affective items appear to load more consistently on the cognitive than the somatic factor.

#### ***Comparability of the BDI-II to the BDI***

The BDI-II appears comparable to its earlier versions in terms of reliability, but it is a clearly superior instrument in terms of its validity (Dozois & Dobson, 2002; Dozois et al., 1998). Beck and Steer (1984) found that the original (1961) version and 1978 (BDI-IA) revision yielded coefficient alphas of .88 and .86, respectively (also see Beck, Steer, & Garbin, 1988). The average internal consistency for the BDI-II is somewhat higher ( $\alpha = .91$ ). Earlier factor analytic studies of the BDI revealed that a three-factor solution (Negative Attitudes Toward Self, Performance Impairment, Somatic Disturbance) was most frequently identified in the literature (Beck, Steer, & Garbin, 1988). However, the number of factors extracted ranged anywhere from one to seven and the average number of factors was 3.96 ( $SD = 1.91$ ) (see Beck, Steer, & Garbin, 1988). Conversely, research conducted on the BDI-II indicates

that a stronger and more stable factor structure exists than for the BDI and BDI-IA (Beck et al., 1996; Dozois et al., 1998; Steer et al., 1999; Steer, Kumar, et al., 1998). Beck et al. (1996) noted that “the transition from the usage of the BDI-IA to that of the BDI-II should introduce no meaningful interpretative problems” (p. 596). Although the mean BDI-II score is approximately 2 points higher than the BDI-IA (Beck et al., 1996; Dozois et al., 1998), a similar relationship to other inventories is demonstrated (Beck et al., 1996) and conversions are available in the test manual. Many of the limitations of the BDI appear to have been resolved with the 1996 revision, making the BDI-II an even stronger instrument than its earlier versions. **54 The BDI-II, BHS, and BSS**

**TABLE 5.2 Factor Analytic Studies of the Beck Depression Inventory-II**

Reference Sample Method # Factors Factor Labels

*Psychiatric*

Beck, Steer, & Brown (1996) 500 outpatients (53% with mood disorders) Principal factors 2 Somatic-Affective  
Cognitive

Buckley et al. (2001) 416 male substance abusers CFA 3 Cognitive  
Affective  
Somatic

Steer et al. (1999) 210 depressed outpatients Principal factors 2 Somatic-Affective  
Cognitive

(same sample) CFA 2 Cognitive  
Noncognitive

Steer, Kumar, et al. (1998) 210 outpatient adolescents Principal factors 3 (2)a Cognitive  
Somatic-Affective

*Nonpsychiatric*

Arnau et al. (2001) 340 primary care patients Principal components 2 Somatic-Affective  
Cognitive

Beck, Steer, & Brown (1996) 120 college students Principal factors 2 Cognitive-Affective  
Somatic

Dozois et al. (1998) 511 college students Maximum likelihood 2 Cognitive-Affective  
Somatic-Vegetative

511 college students CFA 2 Cognitive-Affective  
Somatic-Vegetative

Osman, Downs, et al. (1997) 230 college students Maximum likelihood 3 Negative Attitudes  
Performance Difficulty  
Somatic Elements

Steer & Clark (1997) 160 college students Principal factors 2 Cognitive-Affective Somatic

Whisman et al. (2000) 576 college students CFA 2 Cognitive-Affective Somatic

aOnly two factors were generalizable; CFA \_ confirmatory factor analysis.

**Range of Applicability and Limitations**

The BDI-II and its predecessors have been used extensively in research and practice and are among the most frequently used psychological tests to date (Camara, Nathan, & Puente, 2000). A number of populations have been studied over the years using the BDI scales, including different psychiatric groups, nonclinical (analogue, undergraduate, and community) samples, ethnic groups, medical populations, and age groups. This instrument is an extremely useful research tool and is also a clinically sensitive instrument that may be used for determining a baseline level of severity, formulating clinical hypotheses, deriving a case conceptualization, monitoring session-by-session treatment change, and determining treatment outcome (see Dozois & Dobson, 2002).

Nonetheless, there are a number of limitations of this instrument that need to be considered. The liabilities include difficulties at the individual item level, its limited ability to detect deviant response sets and styles, the instability of scores over time, the lack of normative information for different ethnic groups, and the potentially premature development of BDI-II derivatives.

Most of the BDI-II items and response options discriminate among individuals who differ in their severity of depression. In the BDI-II manual, Beck et al. (1996) reported item-response curves for the BDI-II that indicated that the majority of its items map appropriately onto the construct of depression. However, the item weights did not perform as expected for four items (“punishment feelings,” “suicidal thoughts or wishes,” “agitation,” and “loss of interest in sex”). For instance, on some of these items severely depressed individuals are more likely than less depressed persons to endorse statements that represent lower rather than higher a priori item weights (i.e., a score of 1 or 2 rather than 3). Similar findings were reported by Santor, Ramsay, and Zuroff (1994) for the BDI-IA.

Concern has also been expressed that women tend to score higher on the BDI-II than do men (Beck et al., 1996; Steer, Kumar, et al., 1998), which might imply that the items are biased or that different psychometric properties exist across gender. However, item analyses using item-characteristic curves demonstrate that few biases exist. For example, although some BDI-IA items were shown to be somewhat biased by gender (“punishment,” “crying,” and “body image change”), for the most part equally depressed males and females do not respond differently on the BDI-IA (Santor et al., 1994). Moreover, the item concerning body image was dropped in the most recent revision of the BDI. Because the gender difference in total scores was not present with the

BDI-IA, however, research needs to determine whether separate cutoffs and norms are necessary (Steer, Kumar, et al., 1998).

Another limitation of the BDI-II pertains to the lack of established criteria for determining the accuracy of an examinee’s response. Some researchers have pointed out that very low BDI scores may reflect social desirability. Clark, Crewdson, and Purdon (1998), for example, found that participants whose total score was 0 or 1 were characterized by a positive impression management style. Although possibly less problematic in clinical settings, Clark et al. suggested that for research purposes extremely low-scoring participants should be excluded from nondepressed control groups (also see Kendall, Hollon, Beck, Hammen, & Ingram, 1987). In addition to low-end specificity (i.e., the degree to which low scores are truly indicative of the absence of psychopathology), Kendall et al. (1987) also raised the issue of high-end specificity (i.e., the degree to which high scores reflect nosological depression as opposed to other clinical conditions).

The BDI-II may differentiate depression from other conditions (Beck et al., 1996), but high scores do not necessarily imply specificity to depression. Thus, accessing other sources

of information is essential for appropriate diagnosis. There are limited data on the test-retest reliability of the BDI-II, but what is available suggests that this instrument is moderately stable over short periods of time yet also sensitive to clinical change. The BDI is an excellent measure for monitoring session-by-session changes in therapy and for the assessment of treatment outcome (see Dozois & Dobson, 2002).

The issue of repeated assessment and the extent to which decreases in BDI-II scores are artifactual (e.g., due to error variance) or genuine has, however, been contested of late (e.g., Ahava et al., 1998; Yin & Fan, 2000). Because this problem is not unique to the BDI-IA and BDI-II (cf. Sharpe & Gilbert, 1998), we discuss this issue further in a subsequent section of the chapter.

Another limitation of the BDI-II is that there are not adequate norms for diverse ethnic groups. The majority of psychometric studies on the BDI-II have used Caucasians. In the BDI-II studies reviewed in this chapter, the average sample composition was 80% Caucasian (range = 37% to 97%). Although some population-specific norms have been developed for the BDI-II, appropriate norms for different ethnic groups are needed (Buckley et al., 2001; O'Hara, Sprinkle, & Ricci, 1998).

Finally, a recent criticism of the BDI-II involves the development of controversial spin-offs. There are many derivatives of the BDI-II, including the BDI 13-item short form (Leahy, 1992), the Beck Depression Inventory for Primary Care (BDI-PC), and the BDI for youth. The psychometric properties of these alternative forms appear to be good (e.g., Leahy, 1992; Reynolds & Gould, 1981) but there are limited data on some of these measures and questionable utility for others. For example, there are no published data on the Beck youth scales aside from what is reported in the test manual (J.S. Beck, personal communication, November 2001). The BDI-PC, which has been marketed by Psychological Corporation as the BDI-Fast Screen, is similarly problematic.

This instrument consists of seven items (sadness, pessimism, past failure, loss of pleasure, self-dislike, self-criticalness, and suicidal thoughts or wishes) from the BDI-II. Beck and his colleagues developed the BDI-Fast Screen in response to the confound that exists in primary care patients between somatic items and physiological problems: "The average rate of specificity may be lower because the somatic and performance symptoms of depression, which are contained in most self-report measures of depression, overlap with the types of somatic and performance symptoms that occur in medical illnesses" (Beck, Guth, Steer, & Ball, 1997, p. 785).

The rationale for developing the BDI-Fast Screen appears reasonable and the test seems to be reliable and valid in primary care populations of adults (Beck, Guth, et al., 1997; Steer, Cavalieri, Leonard, & Beck, 1999) and adolescents (Winter, Steer, Jones-Hicks, & Beck, 1999). However, it is probable that the developers of this derivative were too hasty in excluding somatic-related items in the development of this instrument. Arnau et al. (2001), for instance, argued that there are good reasons not to exclude somatic items when assessing

depression in medical patients. These items may overlap with medical problems but nonetheless contribute importantly to the overall prediction of depressive severity. In a sample of primary care patients, Arnau et al. found that the receiver operating characteristics of the BDI-II full scale were excellent. Other researchers have also found that the omission of somatic items does not enhance predictive ability in this population (Aikens et al., 1999; Geisser, Roth, & Robinson, 1997; Levin, Llabre, & Weiner, 1988; Lustman, Clouse, Griffith, Carney, & Freedland, 1997). Deleting items in order to increase specificity often results in a scale's decreased sensitivity (Aikens et al., 1999). Given these criticisms, researchers and clinicians should be cautious and ensure that supplementary data are collected if they choose to use the BDI-Fast Screen.

### **Age and Cross-Cultural Factors**

The BDI appears to be appropriate for research with elderly populations and as a clinical screening instrument (Gallagher, 1986; Gallagher, Nies, & Thompson, 1982; Keane & Sells,

1990). Some of the issues that have been identified in using the BDI with older adults include the readability of the items, possible social desirability (e.g., elderly patients tend to underreport subjective distress), and whether somatic complaints are the best indicators of depression in the elderly because of their ubiquitous nature (Gallagher, 1986). Some patients may find multiple-choice questionnaires such as the BDI-II too cumbersome to complete, particularly if they are severely impaired. Clinicians may therefore opt to use questionnaires that are less complicated (cf. Dozois & Dobson, 2002). Because the BDI-II also contains somatic items, clinicians who choose to use this measure in elderly or medical populations would be prudent to follow up with questions about the etiology of such complaints (e.g., whether they pertain more to one's physical problem or affective status; see Gallagher, 1986). Conversely, there are some data that suggest that the inclusion of the somatic items does not result in a biased estimate of depression in the elderly. Laprise and Ve'zina (1998) examined the receiver operating characteristic curves of French translations of both the BDI and the Geriatric Depression Scale and found no differences in their diagnostic performance indices. Moreover, as the research we reviewed earlier demonstrated, the exclusion of the BDI's somatic items may not improve the instrument's performance.

The BDI scales have been translated into several languages, including Spanish (Bonicatto, Dew, & Soria, 1998), French (Byrne, Baron, & Campbell, 1994), Chinese (Chan, 1991; Shek, 1990), Portuguese (Gorenstein, Andrade, Filho, Tung, & Artes, 1999), Dutch (Bosscher, Koning, & Van Meurs, 1986), Persian (Hojat, Shapurian, & Mehryar, 1986), Arabic (Abdel-Khalek, 1998; West & Al-Kaisi, 1995), and Hmong (Mouanoutoua, Brown, Cappelletty, & Levine, 1991). In general, these translated versions show psychometric properties that are comparable to the untranslated version. A Chinese translation of the BDI-IA, for example, was found to be reliable, valid, and quite commensurate with the English version (Chan, 1991). Nontranslated versions of the BDI also

appear to be reliable and valid in English-speaking minority groups (e.g., Gatewood-Colwell, Kaczmarek, & Ames, 1989). Across languages, the BDI has acceptable reliability and validity (Bonizzato et al., 1998). Simply because a measure seems appropriate for some minorities or cultural groups does not, however, necessarily imply that it will be equally valid and reliable in other similar groups. Furthermore, direct translations may not produce similar reliability at the item level. For example, the item “loss of libido” may be correct semantically but has been shown to lack content validity in Chinese cultures (see Zheng & Lin, 1991; Zheng, Wei, Lianggue, Guochen, & Chenggue, 1988). Clinicians and researchers should ensure that the BDI scale used is appropriate for their particular clientele or population, taking into account their level of identification and acculturation with a given culture.

#### **Accommodation for Populations with Disabilities**

The reading difficulty of the BDI-II is quite low, which makes this task easy to understand and use. Assuming that there are no severe language disabilities or thought disorders, the BDI can be reliably administered to educationally and developmentally delayed individuals (Beck, Carlson, Russell, & Brownfield, 1987). The BDI-II may also be administered orally for individuals with reading impediments or severe concentration difficulties (Beck et al., 1996).

#### **Legal and Ethical Considerations**

As mentioned previously, the BDI-II should not be used as the sole determinant of an individual’s diagnosis of depression. First, the BDI-II was not intended to diagnose depression but to serve as an index of symptom severity. Second, the BDI-II is not comprehensive enough to provide conclusive diagnostic information. For example, the *DSM-IV* exclusionary criteria are not included in the BDI-II nor is the requirement that respondents endorse at least one of the sadness and anhedonia questions. As such, the BDI-II should be used to determine a patient’s symptom severity, to monitor the efficacy of treatment over time, or to suggest the need for a more thorough assessment.

Clinicians and researchers should be aware that some of the BDI-II items are related to an increased risk of suicidality. In particular, patients or participants who score higher than 1 point on items 2 (“pessimism”) or 9 (“suicidal thoughts or wishes”) should be evaluated for suicide potential (see Beck et al., 1996). Moreover, because hopelessness and the risk of suicide are part of the nature of depression, clinicians working in this area are encouraged to be familiar with local laws and procedures regarding involuntary hospitalization.

#### **Computerization**

Scannable record forms of the BDI-II are available and a computer-administered version of the BDI-IA has been examined (Steer, Rissmiller, Ranieri, & Beck, 1994). Few empirical studies have evaluated the utility or cost efficiency of this measure relative to the standard administration. Steer et al. (1994) examined computerized versions of the BDI and BHS in a sample of patients with mixed psychiatric diagnoses. Both instruments yielded results that concur with the

published literature on paper-and-pencil administrations. Research is necessary, however, to directly test the relationship

between computer- and questionnaire-administered measures.

Four of the Beck scales may also be scored and interpreted simultaneously through the Beck InterpreTrak\_. According to information provided by Psychological Corporation, this software program permits clinicians to track patient progress on the BDI-II, BHS, BAI, and BSS. Item responses are analyzed by this program and an interpretive summary is provided.

In addition, this software generates session-by-session graphs of symptom change to allow clinicians to monitor treatment outcome. Given that InterpreTrak\_ was released in 2000, it is presently unclear whether this software corrects for repeated administrations or provides additional information (e.g., suggestions for the timing of treatment termination or for the prevention of relapse) that would not be easily available elsewhere.

### **Current Research Status**

The BDI-II has numerous assets that make it an excellent choice for both research and practice (see Dozois & Dobson, 2002). These strengths include the BDI-II's consistency with *DSM-IV* criteria, its excellent psychometric properties, the ease of administration and scoring, its sensitivity to treatment change, and its large empirical database with which to compare results. Some of the limitations of the BDI-II include its inability to provide conclusive diagnostic evidence, the potentially inaccurate weighting of some item statements (as identified via item-characteristic curves), problems with high and low-end specificity, the instability of scores over time even among nonclinical samples, inadequate normative data for different ethnic groups, and the potentially premature development of BDI-II derivatives. Many of these limitations are not uncommon among self-report instruments and have to do more with the *use* of the BDI-II than with the instrument per se. As such, this inventory will likely remain within the top ten most frequently used psychological tests (Camara et al., 2000).

Existing research on the BDI-II suggests that this instrument is comparable and yet superior to its earlier editions.

Nonetheless, there are a number of important research directions that should be pursued, including: (1) determining the utility of the short forms of the BDI-II for screening purposes—for example, how do these instruments compare to other screening instruments in terms of their overall correct classification rates?; (2) assessing whether the exclusion of somatic items actually enhances or impairs the predictive utility of the BDI-Fast Screen; (3) developing norms for different ethnic groups; (4) evaluating whether different norms are required for men and women separately; (5) testing the utility and psychometric properties of the Beck Youth scales; (6) investigating the extent to which the Beck InterpreTrak\_ enhances clinical conceptualization, the provision of feedback to patients, the evaluation of treatment outcome, and clinical decision making; and (7) examining how the BDI-II can be used to test clinical hypotheses regarding treatment choice, treatment outcome, and prevention of relapse. As there are a

host of other important research questions that may be addressed, the recommendations we provide are not intended to be exhaustive, but rather to serve as a springboard for further empirical investigation.

### **BECK HOPELESSNESS SCALE (BHS)**

#### **Test Description**

The BHS is a 20-item true-false questionnaire that is easy to administer and score. Nine of the items are false keyed to control for acquiescent response styles. Each item is scored either 0 or 1, with total scores ranging from 0 to 20. Higher scores reflect more intense levels of hopelessness. Overall, the content of the items represents negative expectations for the future. For example, item 2 states "I might as well give up because I can't make things better for myself," (true-keyed response) and item 13 states "When I look ahead to the future, I expect I will be happier than I am now" (false-keyed response). Generalized pessimism regarding the future is the main focus of the BHS and appears to account for the majority of response variance in factor analytic studies.

#### **Theoretical Basis and Test Development**

Based on previous theory implicating hopelessness with depression and suicide (e.g., Beck, 1963), Beck, Weissman, Lester, and Trexler (1974) developed and tested a measure that would allow the construct of hopelessness to be examined quantitatively. Although hopelessness is somewhat of an abstract concept, Beck et al. (1974) followed the suggestions of Stotland (1969) and defined hopelessness objectively in terms of a person's negative expectancies for the future. Several stages were involved in the development of this scale. First, 9 of the 20 items were adapted from a previous test of attitudes toward the future (Heimberg, 1961, as cited in Beck & Steer, 1988). The remaining items were derived from pessimistic statements about the future made by psychiatric patients who were previously rated by clinicians as exhibiting significant hopelessness. Next, the scale was administered to depressed and nondepressed patients and feedback was obtained regarding the appropriateness of each item to the construct. Finally, several clinicians were asked to rate the scale on its face validity and comprehensibility. Changes in wording were made based on the suggestions of patients and clinicians.

#### **Psychometric Characteristics**

The reliability of the BHS is well supported, primarily with clinical populations. In adult psychiatric patients, internal reliability coefficients (Kuder-Richardson-20) range from .83 to .93 and are often around .90 (Beck, Steer, & Brown, 1993; Beck, Steer, Beck, & Newman, 1993; Durham, 1982; Mendonca, Holden, Mazmanian, & Dolan, 1983; Steer & Beck, 1997; Steer, Beck, Brown, & Beck, 1993; Young, Halper, Clark, & Scheftner, 1992). Reliability estimates are equally strong with adolescent psychiatric patients (Kumar & Steer, 1985; Steer, Kumar, & Beck, 1993a). In contrast to research on psychiatric samples, the internal reliability of the BHS is typically lower and more variable in nonpsychiatric populations (Durham, 1982; Holden & Fekken, 1988; Rew, Taylor-Seehafer, Thomas, & Yockey, 2001; but see Johns &

Holden, 1997). It is also possible that some individual test items may lower the overall reliability of the scale (Holden & Fekken, 1988; Steed, 2001; Steer & Beck, 1997; Young et al., 1992). In a nonpsychiatric sample, Steed (2001) found that items 4, 5, 8, and 13 exhibited especially low item-total correlations and recommended excluding these items when testing this population. Similarly, Young et al. (1992) asserted that while the BHS items seem to tap the construct of hopelessness well, they do so only for persons who exhibit moderate to severe levels of hopelessness. Congruent with the weaker internal reliability coefficients found in nonclinical groups, the BHS appears to be less reliable in individuals who display low levels of hopelessness.

Test-retest reliability of the BHS appears to be high in nonpsychiatric students. Holden and Fekken (1988), for instance, reported a 3-week test-retest reliability coefficient of .85. For clinical groups, these estimates are lower, which may reflect genuine clinical change. In mixed psychiatric samples, Beck and Steer (1988) reported test-retest reliability coefficients of .69 and .66 over 1 and 6 weeks, respectively.

Substantial support exists for the validity of the BHS. This instrument shows robust relationships with related constructs, across all types of populations. Among adult clinical patients, the BHS correlates significantly with depression symptoms (Beck et al., 1993; Dyck, 1991; Strosahl, Chiles, & Linehan, 1992; Wilkinson & Blackburn, 1981; Winefield, 1979), suicidal ideation (Beck et al., 1993; Dyck, 1991, Study 3; Ellis, 1985; Mendonca & Holden, 1998; Wetzel, Margulies, Davis, & Karam, 1980), and suicide intent (Strosahl et al., 1992; Wetzel, 1976; Wetzel et al., 1980). In addition, the BHS correlates with suicide intent even after depressive severity has been partialled out (Beck & Steer, 1988; Wetzel et al., 1980).

The BHS also correlates negatively and significantly ( $r = -.63$ ) with the Reasons for Living Inventory (RFL; Dyck, 1991), a measure developed to assess people's reasons for staying alive when pondering suicide. Further evidence for the validity of the BHS emerges from findings that this measure distinguishes between psychiatric and nonpsychiatric groups and between ideators and nonideators. For example, suicide attempters and psychiatric inpatients display significantly higher BHS scores than do nonclinical controls (Durham, 1982; Simonds, McMahon, & Armstrong, 1991).

Moreover, depressed patients exhibit higher scores than nonpsychiatric controls, recovered depressed patients, and neverdepressed controls (Hamilton & Abramson, 1983; Wilkinson & Blackburn, 1981). After controlling for initial depression severity and length of hospital stay, McCranie and Riley (1992) found that pretreatment hopelessness scores were significant predictors of depressive severity 3 weeks later. Finally, the BHS also differentiates between suicide ideators and nonideators (Beck et al., 1993). Hopelessness, in fact, appears to be a better predictor of suicide intent than is depression (Beck & Steer, 1988; Wetzel, 1976; Wetzel et al., 1980).

The validity of the BHS is also supported in nonclinical groups. In these samples, the BHS correlates significantly

with severity of depressive symptoms (Dixon, Heppner, Burnett, & Lips, 1993; Dyck, 1991; Johns & Holden, 1997; Joiner & Rudd, 1995; Prociuk, Breen, & Lussier, 1976; Rudd, 1990; Thackston-Hawkins, Compton, & Kelly, 1994; Weber, Metha, & Nelsen, 1997; Whatley & Clopton, 1992; Wilkinson & Blackburn, 1981), suicidal ideation (Dyck, 1991; Johns & Holden, 1997; Rudd, 1990; Weber et al., 1997; Whatley & Clopton, 1992), and suicide intent (Dyck, 1991), and negatively with measures of hope (Herth, 1991; Obayuwana et al., 1982). It is important to note, however, that the validity of the BHS is limited in this population because of its questionable reliability.

There exists support for the validity of the BHS in adolescent samples. This instrument correlates significantly and positively ( $r = .68$ ) with the BDI (Johnson & McCutcheon, 1981) and negatively ( $r = -.64$ ) with the RFL-A (an adolescent version of the RFL; Gutierrez, Osman, Kopper, & Barrios, 2000). In fact, Kumar and Steer (1995) found that the BHS was a more powerful predictor of suicide ideation than 12 other variables, including history of sexual abuse, current suicide attempt, past suicide attempt, ethnicity, and diagnosis of a mood disorder. The BHS also appears to predict suicide ideation better than both the BDI and BAI in adolescent inpatients (Steer, Kumar, & Beck, 1993a, 1993b).

Although this evidence is encouraging, some studies raise concern about whether the BHS is as valid for adolescents as it is for adults. To illustrate, the BDI and BHS correlated highly ( $r = .75$ ) in a sample of female, adolescent suicide attempters, but neither instrument was significantly associated with suicide intent (Rotheram-Borus & Trautman, 1988). Furthermore, Gutierrez et al. (2000) found that the BHS did not significantly differentiate those individuals who never seriously considered suicide from those who attempted.

The predictive utility of the BHS has been examined across a number of studies. Beck, Brown, Steer, Dahlsgaard, and Grisham (1999) used receiver operating characteristic (ROC) analyses to identify cutoff scores that provided the best prediction of future suicide. In these analyses high sensitivity was understandably considered more important than high specificity in identifying potential suicide attempters. With a cutoff of 8 and higher representing a high-risk group, the BHS was highly sensitive (90%) but limited in specificity (42%). Individuals scoring at or above this threshold were six times more likely to commit suicide than those scoring below the cutoff. These findings concur with those of Beck, Brown, Berchick, Stewart, and Steer (1990), who found that a cutoff score of 9 yielded 94% sensitivity and 41% specificity. In this study, individuals in the high-risk group were 11 times more likely to commit suicide than those in the low-risk group. Using a more stringent cutoff criterion, Cochrane-Brink, Lofchy, and Sakinofsky (2000) found that scores of 15 or higher yielded a sensitivity rate of 100% and a specificity rate of 71%. Negative predictive power was excellent at 100%, and positive predictive power (45%) was much higher than the 1.3% reported by Beck et al. (1999). Therefore, using a slightly higher cutoff score than the recommended 8 or 9 may increase overall classification rates

without necessarily jeopardizing sensitivity. Factor analytic studies have also supported the construct validity of the BHS. The BHS was originally reported to consist of three factors (Beck et al., 1974), but more recent studies have demonstrated that the variance is best explained by a one-factor solution (see Steed, 2001). This finding is replicable across both adult clinical (Dyce, 1996; Mendonca et al., 1983; Young et al., 1992) and nonclinical (Chang, D'Zurilla, & Maydeu-Olivares, 1994; Steed, 2001; Ward & Thomas, 1985) samples. Thus, the BHS is arguably a unidimensional measure best described as general hopelessness or pessimism for the future.

#### **Range of Applicability and Limitations**

The BHS was initially developed and tested on adult clinical patients, and there is a wealth of evidence implying that hopelessness, as measured by the BHS, is an excellent and reliable predictor of suicidal tendencies, including ideation, intent, and attempt. The BHS was not developed for use with nonpsychiatric individuals, yet its validity seems to be supported in this population as well. Because the internal reliability of the BHS is lower and more variable in nonclinical groups, however, researchers may wish to consider Steed's (2001) recommendation to eliminate items that do not fit well with the rest of the scale. Regardless of the sample composition, researchers or clinicians should be cautious using the BHS when moderate to high levels of hopelessness are not anticipated (see Young et al., 1992).

There has been some debate in the literature over whether the BHS measures hopelessness or high, negative social desirability in patients (see Glanz, Haas, & Sweeney, 1995).

Fogg and Gayton (1976) first reported that the BHS correlated negatively with social desirability (coefficients ranged from  $-.47$  to  $-.64$ ), leading these authors to warn that scores on the BHS could be contaminated by response sets. This finding has since been replicated by several other researchers (e.g., Ivanoff & Jang, 1991; Linehan & Nielsen, 1981; Mendonca et al., 1983). For example, Mendonca et al. (1983) studied 78 patients who presented at a crisis unit in a psychiatric hospital. BHS scores were significantly different in nonsuicidal individuals, suicidal ideators, and attempters, but these differences disappeared when social desirability was partialled out. Although one would expect that hopelessness is inversely related to social desirability, the BHS items related more to responding undesirably than to the magnitude of suicidality. Thus, even though social desirable response sets should be considered when using the BHS, social undesirability is a genuine feature of many aspects of psychopathology.

#### **Age and Cross-Cultural Factors**

Because of the lack of psychometric information on the BHS in adolescents, coupled with the fact that some studies indicate that this measure may not be as predictive of suicidality as it is in adult clinical groups, it is unclear whether this instrument should be used with adolescent clinical populations. Congruent with this notion, Beck and Steer (1991) stated that this measure is most appropriate for adults aged

17 years and older. Regardless of whether this instrument is to be used in younger individuals, BHS users should supplement their assessment of suicidality with additional indices to provide convergent evidence. Additional research is also necessary to more adequately assess the psychometric characteristics of the BHS in adolescents and to ascertain the relationship between the BHS and suicidality in this population.

There is a similar paucity of data on the use of the BHS with various ethnic groups. As with the BDI-II, the majority of BHS studies have used primarily Caucasian samples. Steer, Iguchi, and Platt (1994) found that Black participants scored significantly lower on the BHS than Caucasian and Hispanic individuals. Whether this decreased hopelessness is a veridical finding or may reflect cultural biases in the BHS is not presently known. Additional research is required to test the utility of the BHS in ethnic minority groups (Ivanoff & Jang, 1991), including ethnic minority adolescents (Steer et al., 1993a), and to determine whether separate normative information is warranted.

Both standard and translated versions of the BHS have been examined in various geographic regions of the world, including Brazil (Feijo, Saueressig, Salazar, & Chaves, 1997), China (Chiles et al., 1989), Finland (Suominen, Isometsa, Henriksson, Ostamo, & Lonnqvist, 1997), Japan (Tanaka, Sakamoto, Ono, Fujihara, & Kitamura, 1996; Tanaka, Sakamoto, Ono, Fujihara, & Kitamura, 1998), and Sweden (Nordstroem, Schalling, & Asberg, 1995). These studies reveal that there may indeed be significant cultural differences in the meaning of BHS scores and its psychometric properties. Studies conducted in European populations report findings similar to those using North American populations (Nordstroem et al., 1995; Suominen et al., 1997). Studies of Asian populations have, however, produced disparate results. Tanaka et al. (1998) found a two-factor structure for the BHS as opposed to the one-factor solution typically found in North American studies. Also, there were no significant correlations between depression symptoms and the factor scores, nor any notable differences in BHS scores between individuals with a psychiatric history and those without such a history. Similarly, Chiles et al. (1989) found that the BHS was a significant predictor of suicide intent for American, but not Chinese, attempters. The relationships among depression, hopelessness, and suicidal intent (and the extent and manner in which it is manifested) may be quite different across cultures.

#### **Accommodation for Populations with Disabilities**

Some persons might have difficulty completing the BHS on their own, because of extreme fatigue, severe concentration problems, chronic illness, or visual impairments. In such instances, it is possible to administer the questionnaire orally (for instructions, see Beck & Steer, 1988).

#### **Legal and Ethical Considerations**

The BHS is considered to be an indirect measure of suicidal risk (Beck & Steer, 1988) and should not be used as the sole index of suicidality. Rather, suicide potential should be assessed with a thorough clinical interview and the use of more direct measures. Moreover, when deciding which cutoff to

use as indicators of high suicidal risk, the concern for high sensitivity must override considerations to reduce the number of false negatives.

### **Computerization**

Steer et al. (1994) tested a computerized version of the BHS with 330 inpatients. The reliability of the computerized BHS concurs with the printed version. The computerized version correlates highly with the total BDI-IA ( $r = .70$ ) and the BDI-IA pessimism item ( $r = .67$ ). The relationship between the BHS and the BDI-IA was virtually the same ( $r = .68$ ) when the pessimism item was excluded, thereby ruling out the confound that the relationship between these measures was driven by item overlap. Although the computerized BHS was able to significantly differentiate mood-disordered and non-mood-disordered groups it produced mean scores that were significantly lower than those reported previously using the printed version (Steer et al., 1994). Thus, test users should be cognizant of the fact that the cutoff scores recommended for the printed version may not be appropriate for the computerized version. Further research is needed to assess the comparability of the printed and computerized versions.

### **Current Research Status**

The BHS is an excellent measure and is the most widely used measure of hopelessness available (Glanz et al., 1995). Empirical data indicate that the BHS is a highly reliable and valid measure for use among adult clinical groups. Although the BHS provides only an indirect assessment of suicidality, important information concerning the relative severity of pessimism may also be acquired. As such, the BHS may be an important measure to include in the evaluation of treatment outcome. Although this measure is used most frequently with depressed and suicidal individuals, hopelessness is appropriate to assess in a variety of other populations as well. To illustrate, the BHS has been used to assess levels of hopelessness in identified carriers of the Huntington's disease gene (Tibben, Timman, Bannink, & Duivenvoorden, 1997), HIV-positive persons (Swindells et al., 1999), and alcohol and heroin-dependent women (Beck, Steer, & Shaw, 1984). The BHS may be used to assess baseline levels of hopelessness and to monitor fluctuations and improvements in hopelessness over time.

Until more consistent data are available, ascertaining whether the BHS is as useful for adolescents as it is for adults is difficult. Rotheram-Borus and Trautman (1988) suggested that hopelessness may be a symptom of depression in adolescents rather than a separate factor. Clinicians are cautioned to supplement the use of the BHS with another instrument designed specifically for assessing suicidality in adolescents. Use of the BHS with nonclinical groups has been ubiquitous, despite the rather poor reliability reported for this population. Considering that the BHS may be unreliable among groups with low levels of hopelessness (Young et al., 1992), researchers should ensure that at least moderate levels of hopelessness are anticipated among the population they intend to test. Also, users should be aware of the presence of social desirability among certain populations and how this could

affect BHS scores.

An important direction for research involves examining the psychometrics and appropriateness of the BHS with ethnic and cross-cultural groups. This research should address whether reliable ethnic differences in BHS scores exist and if separate norms are necessary for different cultures. Also, test users need to be aware of cultural differences in expectations for the future and how this might influence BHS scores (Tanaka et al., 1998).

Finally, the incremental validity of the BHS, relative to other measures of suicidality, is another empirical question worth addressing. For example, it is not presently clear whether the BHS should be used instead of, or in addition to, the BSS or the Suicide Intent Scale (SIS; Beck, Schuyler, & Herman, 1974) for predicting or assessing suicide risk. Schnyder, Valach, Bichsel, and Michel (1999) argued that there has been a tendency to overvalue the role of hopelessness in suicide assessment. If a clinician is concerned only with determining suicide risk, rather than assessing general pessimism that may be related to psychopathology or treatment outcome, the BSS appears to be the measure of choice (Cochrane-Brink et al., 2000), and this measure is discussed next.

### **BECK SCALE FOR SUICIDE IDEATION (BSS)**

#### **Test Description**

Beck et al. (1988) developed a self-report version of suicidal ideation that could be administered by either paper and pencil or computer. The BSS is a 21-item measure of which only the first 19 items are scored. The final two are items used to record information concerning previous suicide attempts.

All items consist of three response options, ranging from 0 to 2. Respondents are asked to circle the statement that best describes how they have been feeling over the past week, including the current day. An example item follows:

0 I have a moderate to strong wish to live.

1 I have a weak wish to live.

2 I have no wish to live.

Total scores, which range from 0 to 38, are obtained by adding the item values. The first five items (i.e., “wish to live,” “wish to die,” “reasons for living or dying,” “active suicide attempt,” and “passive suicide attempt”) serve as an initial screen for suicide ideation. If respondents circle zero on both of the items pertaining to suicide attempts, they are instructed to skip to the end of this scale to complete the last two items.

#### **Test Development**

Beck et al. (1972) published recommendations for the design and operation of programs for suicide prevention and amelioration. Among these recommendations was a classification system for suicidal behaviors, which was comprised of three broad categories: suicide completions, attempts, and ideation. Although Beck and colleagues had already developed an intent scale for suicide attempters that was also partly applicable to suicide completers (Beck, Kovacs, & Weissman, 1979), there was a need to develop an instrument that would

assess suicidality for the third category of suicidal behaviors—suicide ideation.

Suicide ideation can be defined as “the presence of current plans and wishes to commit suicide in individuals who have not made any recent overt suicide attempts” (Beck et al., 1988, p. 499). Beck et al. (1979) initially developed a clinician rating scale called the Scale for Suicide Ideation (SSI).

The items for this rating scale were generated on the basis of clinical observations, interviews with suicidal patients, and previous research in the area of suicide. This measure was originally piloted on suicidal patients and its items were refined or eliminated if they were ambiguous, difficult to rate, or consisted of overlapping content. This test construction phase resulted in a 19-item clinician rating scale, designed to be administered as a semistructured interview. The test developers were later interested in adapting the SSI into a self-report index of suicidal ideation that could be used independently or concurrently with this instrument. This new inventory, the BSS, was then tested on inpatient and outpatient samples.

### **Psychometric Characteristics**

The internal consistency of the BSS is excellent in both adult and adolescent clinical samples (Beck & Steer, 1991; Beck et al., 1988; Kumar & Steer, 1995; Steer, Kumar, & Beck, 1993b). For the printed version of the BSS, coefficient alpha

ranges from .87 (Beck & Steer, 1991) to .93 (Beck et al., 1988). The computer-administered version yields high coefficients that range from .90 (Beck & Steer, 1991) to .96 (Beck et al., 1988). Item total correlations are generally acceptable and range from .20 to .73 ( $M = .52$ ,  $SD = .15$ ). The test-retest reliability of the BSS appears to be moderate. Beck and Steer (1991), for example, reported a 1-week test-retest reliability coefficient of .54 in a sample of 60 adult inpatients.

BSS scores decreased significantly during this time frame, which the authors hypothesized was the result of clinical improvement (Beck & Steer, 1991).

The validity of the BSS has also been tested using both administration formats. The paper-and-pencil and computerized versions of the BSS each correlate highly (e.g.,  $r = .90$ ) with clinician ratings (Beck et al., 1988), self-report (Beck & Steer, 1991; Cochrane-Brink et al., 2000), and related indices of suicidality (e.g., previous attempts, BDI-IA scores, the BDI-IA suicide ideation item; Beck & Steer, 1991). To date, only the printed version has been tested with adolescent inpatients (Kumar & Steer, 1995; Steer et al., 1993). In this population, the BSS correlates well with past history of suicide attempts, the BDI-IA (excluding suicide ideation and hopelessness items), the presence of a mood disorder, the BHS, and the BAI (Kumar & Steer, 1995; Steer et al., 1993).

The predictive utility of the BSS appears promising, but additional research is needed in this area. A cutoff score of 24 yields excellent sensitivity (100%) and specificity (90%), as well as impressive positive (100%) and negative (71%) predictive power (Cochrane-Brink et al., 2000). The BSS also significantly predicts the decision to admit a patient because

of risk for suicide (Cochrane-Brink et al., 2000). Given that only two studies have factor analyzed the BSS (Beck & Steer, 1991; Steer, Rissmiller, Ranieri, & Beck, 1994), it is not presently known whether a stable factor structure exists for this measure. Three main factors (active suicidal desire, suicidal ideation, and preparation for suicide) emerged in a fairly clear and consistent manner in these studies (although Beck & Steer, 1991, actually reported a fivefactor solution). Visual inspection of the factor loadings reported in these studies points to discrepancies that appear to limit the generalizability of the factor solutions reported. However, further investigation of this factorial validity of the BSS is needed.

#### **Range of Applicability and Limitations**

The BSS was developed as a self-report version of the SSI and has been appraised using adult clinical patients. Therefore, the BSS is best used as a measure of suicidal ideation in this population. However, it should be noted that the BSS should never be used alone to clinically assess suicidality and should be used in conjunction with a professional clinical assessment (Beck & Steer, 1991).

Limited data are available on the use of the BSS with adolescent patients (see Kumar & Steer, 1995; and Steer et al., 1993, for exceptions). Given that the BSS was designed to detect the severity of suicidal thoughts in adults and adolescents (Beck & Steer, 1991), further research should focus on the psychometric properties and the operating characteristics of the BSS in adolescents. Currently, it is most appropriate to use the BSS in adult psychiatric patients older than 17 years (Beck & Steer, 1991). Similarly, no research has examined the psychometrics of the BSS in nonpsychiatric groups. Such studies are important to increase the generalizability of this scale to community samples. In addition, the development of nonclinical norms would permit normative comparisons and enhance the evaluation of clinical significance.

#### **Accommodation for Populations with Disabilities**

As with the BDI-II and BHS, the BSS may be administered orally if test takers experience difficulties with the standard format (see Beck & Steer, 1991).

#### **Legal and Ethical Considerations**

The BSS was not developed to replace clinical assessments (Beck & Steer, 1991) and should only be used as an adjunct to professional evaluations conducted by trained clinicians. Given the low base rate of actual suicide attempts, and consequently the increased difficulty in predicting suicidal behavior, it might be quite difficult to arrive at a cutoff score for the BSS that produces high specificity, while at the same time maintaining high sensitivity (but see Cochrane-Brink et al., 2000). Therefore, despite the possibility of increased false positives, it is in the best interest of clients to utilize lower cutoff scores to maintain high sensitivity. The BSS was developed to assist in the assessment of suicide risk as opposed to actual suicide prediction (Beck & Steer, 1991), but it is important to remember the gravity of false negatives. Finally, test users should be familiar with the protocol as well as the legal and ethical issues surrounding involuntary hospitalization.

### **Current Research Status**

Given its good psychometric properties and the ease with which the BSS is administered and scored, it is expected that the BSS will become more widely used than it has been in

studies examining suicide ideation in adult clinical groups. Compared to the clinician-scored SSI, the BSS has the potential to be administered to large groups of research participants at a time, making it a convenient measure for investigating suicidality. Clinically, it has been described as the “clinical scale of choice” for the assessment of suicidality (Cochrane-Brink et al., 2000, p. 450). The BSS also has the potential to be an excellent measure of suicidal risk in adolescents, but more research is needed to determine its psychometric properties in younger populations. There is a paucity of information available concerning the use of the BSS with nonclinical groups. Consequently, except for research purposes, the BSS should not be used with nonclinical groups until adequate psychometric data are available. Finally, additional research is needed to evaluate the utility of the BSS in ethnic minority and cross-cultural groups.

### **THE BECK SCALES IN CLINICAL PRACTICE**

The Beck scales have been used extensively in research and practice. In this section, we highlight some of the uses of the BDI-II, BHS, and BSS in clinical practice and raise practical issues for clinicians who use these measures. In general, we highly recommend the use of these Beck scales for the assessment of depressive severity, hopelessness, and suicidality and for the ongoing evaluation of psychotherapy outcome.

As this chapter has demonstrated, these scales are reliable and valid and provide clinically meaningful information. Numerous other depression symptom scales are available (Dozois & Dobson, 2002; Nezu, Ronan, Meadows, & McClure, 2000), but they often focus on slightly different themes than the BDI-II and include items that are not directly pertinent to diagnostic criteria, thereby decreasing their specificity. The BDI-II is most congruent with the *DSM-IV* criteria, exhibits excellent psychometric properties, and emphasizes the cognitive and attitudinal symptoms of depression more than other popular self-report measures. The BDI-II is also the most frequently used self-report measure of depression, which affords practitioners the opportunity to compare their clinical results to the literature.

If clinicians are interested in assessing suicidal risk, we recommend using the BSS over the BHS because of its more direct link with suicidality. This does not mean the BHS should not be used with suicidal patients. Hopelessness, as measured by the BHS, is consistently predictive of suicidal tendencies. The BHS is also an excellent instrument for assessing treatment outcome and for monitoring a patient’s feelings of hopelessness over the course of treatment. The BHS deals primarily with general pessimism about the future and may also be used to assess a patient’s motivation for therapy and expectations for treatment change. Westra, Dozois, and Boardman (2002), for example, found that pretreatment hopelessness was significantly higher in therapy dropouts than in individuals who completed treatment. Hopelessness

about symptom control was also related to fewer reductions in dysfunctional attitudes and to poorer overall treatment response.

The BSS is an excellent tool for assessing suicide risk and for tracking fluctuations in a patient's suicidal thoughts over the course of therapy. The BSS is also a valuable tool for initial assessment as it can be used to assess imminent risk for suicide. Although there exist no published recommendations for BSS cutoff scores, Cochrane-Brink et al. (2000) used a cutoff of 24, which produced excellent predictive values.

The Beck scales may also be helpful for treatment planning. Patients often present to clinics with comorbid conditions and it is often difficult for clinicians to know which condition to target as an initial strategy for intervention.

When a patient exhibits comorbid depression and anxiety, for instance, understanding the severity of his or her depressive symptoms, degree of hopelessness, and risk for suicide may differentially guide one's approach to intervention. If the depression, hopelessness, and suicidality are very severe, it would be important to deal with these issues first so that the patient will have the resources, motivation, and energy to manage exposure-based approaches for treating anxiety. Conversely, if the patient does not present a high risk for suicide and does not show a high level of depression or hopelessness, then the clinician may opt to target the anxiety, with the hypothesis being that once the anxiety has improved, the depression will dissipate as well. Thus, having accurate data from these self-report instruments can facilitate case conceptualization and treatment planning (see Dozois & Dobson, 2002).

There are a number of other uses of the Beck scales in practice. These instruments may be used to (1) ensure that one's approach to treatment is effective; (2) monitor problems (e.g., motivational issues) that may arise during the course of treatment; (3) encourage patients by using data and demonstrating that the practitioner is confident in his or her approach and that he or she respects accountability; (4) illustrate to patients the amount of progress being made (e.g., depressed patients are notorious for disqualifying the positives and believing that they are not making significant progress when they in fact are); (5) examine the stability of the treatment response (e.g., minimizing the chances that a patient's change simply reflects a flight into health); (6) indicate when treatment has been successful and may safely be terminated; (7) determine the clinical significance of treatment change; and (8) prevent relapse (see Dozois & Dobson, 2002).

The clinical significance of symptom change is important in both psychotherapy outcome trials and in clinical practice. One strategy for determining clinical significance is to use normative comparisons. Normative comparisons allow clinicians to determine whether a patient's functioning on a given measure has shifted from being within the dysfunctional range to being within a nomothetically average range.

Nonclinical norms have been developed for both the BDI-II (Kendall & Sheldrick, 2000) and the BHS (Dozois, Covin,

& Brinker, 2003). We recommend the use of these norms for the assessment of clinical significance.

One issue that was raised earlier in this chapter pertains to the use of the Beck scales in repeated assessment. Although we recommend using the Beck scales frequently in practice to, among other things, gauge efficacy, researchers and practitioners need to be aware of the issues that surface when administering a task repeatedly to clients. As previously noted, a number of studies have found that self-report symptom scores drop with repeated assessments. Ahava et al. (1998) administered the BDI-IA over 7 weekly administrations and found that scores decreased substantially over time in nonclinical participants with no intervention. In the BDI-II manual, Beck et al. (1996) mentioned that the effects of memory and response sets need to be examined but that they should be the same with the BDI-II as they were with the earlier versions of the BDI. This is not very encouraging given the findings from Ahava et al. (1998) and others (e.g., Sharpe & Gilbert, 1998; Yin & Fan, 2000). Clinicians who decide to use the Beck scales recurrently in practice should consider the effects that repeated assessment have on obtained scores. One implication is that it is important to ensure that the decrease in depression, hopelessness, or suicidality scores are in fact related to treatment change rather than due exclusively to repeated assessment. Although this would deviate from standardized administration, clinicians may also consider randomizing response options to minimize the effects of earlier exposure to these tests. Researchers using any of these instruments for prescreening may wish to use alternative measures for the initial screen and the indicated measure for determining whether participants meet criteria for their study. Although regression to the mean may account for some of the findings, they are instructive and should serve as a caution when conducting repeated assessments (see also Yin & Fan, 2000).

#### **SUMMARY AND FUTURE DEVELOPMENTS**

This chapter provided a comprehensive review of the BDIII, BHS, and BSS. As our review has documented, these scales exhibit excellent psychometric properties and are extremely useful for research and practice. In terms of research, these Beck scales are used widely. The BDI-II, in particular, has been cited in numerous psychotherapy outcome studies as one of the core dependent variables. This instrument is also used frequently for determining inclusion and exclusion criteria in myriad studies of clinical and analogue depression. Although the BHS and BSS have not been as dominant as the BDI-II, these measures have been utilized to test pessimism and suicide risk in a variety of populations. Clinically, these three scales appear to be excellent measures of the constructs they purport to measure and are useful for case conceptualization, treatment planning, monitoring patient change over time, and evaluating treatment outcome and the clinical significance of therapeutic change. Throughout this chapter, we have highlighted some of the limitations of each of these measures and provided a number of suggestions for further empirical work. Rather than reiterating these recommendations,

we simply conclude that there are many exciting avenues for future research that we hope researchers will investigate. The BDI-II, BHS, and BSS are highly useful instruments and we anticipate that a future review of this research will confirm this generally positive review.

## REFERENCES

- Abdel-Khalek, A.M. (1998). Internal consistency of an Arabic adaptation of the Beck Depression Inventory in four Arab countries. *Psychological Reports, 82*, 264–266.
- Ahava, G.W., Iannone, C., Grebstein, L., & Schirling, J. (1998). Is the Beck Depression Inventory reliable over time? An evaluation of multiple test-retest reliability in a nonclinical college student sample. *Journal of Personality Assessment, 70*, 222–231.
- Aikens, J.E., Reinecke, M.A., Pliskin, N.H., Fischer, J.S., Wiebe, J.S., McCracken, L.M., & Taylor, J.L. (1999). Assessing depressive symptoms in multiple sclerosis: Is it necessary to omit items from the original Beck Depression Inventory? *Journal of Behavioral Medicine, 22*, 127–142.
- Ambrosini, P.J., Metz, C., Bianchi, M.D., Rabinovich, H., & Undie, A. (1991). Concurrent validity and psychometric properties of the Beck Depression Inventory in outpatient adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*, 51–57.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Arnau, R.C., Meagher, M.W., Norris, M.P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychology, 20*, 112–119.
- Beck, A.T. (1963). Thinking and depression. *Archives of General Psychiatry, 9*, 324–333.
- Beck, A.T., Brown, G., Berchick, R.J., Stewart, B.L., & Steer, R.A. (1990). Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients. *American Journal of Psychiatry, 147*, 190–195.
- Beck, A.T., Brown, G.K., Steer, R.A., Dahlsgaard, K.K., & Grisham, J.R. (1999). Suicide ideation at its worst point: A predictor of eventual suicide in psychiatric outpatients. *Suicide and Life-Threatening Behavior, 29*, 1–9.
- Beck, A.T., Davis, J.H., Frederick, C.J., Perlin, S., Pokorny, A.D., Schulman, R.E. et al. (1972). Classification and nomenclature. In H.L.P. Resnik & B.C. Hathorne (Eds.), *Suicide prevention in the 70s* (pp. 7–12; DHEW Publication No. HSM 72–9054). Washington, DC: U.S. Government Printing Office.
- Beck, A.T., Epstein, N., Brown, G., & Steer, R.A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56*, 893–897.
- Beck, A.T., Epstein, N., Harrison, R.P., & Emery, G. (1983). *Development of the Sociotropy-Autonomy Scale: A measure of personality factors in psychopathology*. Unpublished manuscript, Center for Cognitive Therapy, University of Pennsylvania Medical School, Philadelphia.
- Beck, A.T., Guth, D., Steer, R.A., & Ball, R. (1997). Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. *Behaviour Research and Therapy, 35*, 785–791.

Beck, A.T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal ideation: The scale for suicide ideation. *Journal of Consulting and Clinical Psychology*, 47, 343–352.

Beck, A.T., Rush, A.J., Shaw, B.F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford Press.

Beck, A.T., Schuyler, D., & Herman, I. (1974). Development of suicidal intent scales. In A.T. Beck, H. Resnick, & D. Lettieri (Eds.), *The prediction of suicide* (pp. 45–55). Oxford: Charles Press.

STATE-TRAIT ANGER EXPRESSION INVENTORY-2 (STAXI-2) 74

**Test Description 74**

**Theoretical Basis 75**

**Test Development 76**

**Psychometric Characteristics of the STAXI-2 78**

**Range of Applicability, Computerization, and**

**Accommodation for Diverse Populations 79**

**Use in Clinical or Organizational Practice 79**

THE STATE-TRAIT PERSONALITY INVENTORY (STPI) 80

**Test Description 80**

**Theoretical Basis 81**

**Test Development 81**

**Use in Clinical Practice: Assessing Psychological Vital Signs 82**

SUMMARY AND CONCLUSIONS 83

REFERENCES 83

Practical considerations in psychological assessment are guided by theories of personality and psychopathology that identify fundamental emotional states and personality traits, and combinations of these dimensions that define major diagnostic syndromes. The nature of anxiety, anger, depression, and curiosity as emotional states and personality traits, and the assessment procedures employed in measuring these constructs with the State-Trait Anxiety Inventory (STAI), State-Trait Anger Expression Inventory (STAXI), and State-Trait Personality Inventory (STPI) are described in detail in this chapter.

The chapter is divided into three major sections, in which the STAI, STAXI, and STPI are individually discussed. Each section begins with a test description, which includes definitions of the constructs that are measured, information about administration and scoring, and examples of the items that are included in the inventory scales and subscales. The historical background and theoretical concepts that guided test construction and development, and the psychometric characteristics, range of application, and current research with each measure are also described. The chapter concludes with a brief discussion of anxiety, anger, depression, and curiosity as emotional vital signs of psychological distress and wellbeing that should be carefully assessed and continuously monitored in diagnostic evaluations, psychotherapy, and studies of treatment outcome.

**THE STATE-TRAIT ANXIETY INVENTORY (STAI)**

**Test Description**

The State-Trait Anxiety Inventory (STAI, Form Y) is comprised of two 20-item self-report scales for measuring state

anxiety (S-Anxiety) and trait anxiety (T-Anxiety) as distinct, clearly defined psychological constructs. S-Anxiety was conceptualized as a transitory psychobiological emotional state or condition that is characterized by subjective, consciously experienced thoughts and feelings relating to tension, apprehension, nervousness, and worry that vary in intensity and fluctuate over time. T-Anxiety refers to relatively stable individual differences in anxiety proneness as a personality trait (i.e., differences in the strength of the disposition to respond to situations perceived as threatening with elevations in S-Anxiety).

The STAI was designed to be self-administering, has no time limits, and may be given either individually or to groups of respondents. The S-Anxiety and T-Anxiety items are printed on the front and back sides of a single-page test form. The instructions for the S-Anxiety items require respondents to report the *intensity* of their feelings of anxiety, “*right now, at this moment,*” by rating themselves on the following 4-point scale: (1) “Not at All”; (2) “Somewhat”; (3) “Moderately So”; or (4) “Very Much So.” In responding to the T-Anxiety scale, subjects are instructed to indicate how they *generally* feel by reporting how often they experience the anxiety-related feelings and cognitions described by each item on the following 4-point *frequency* scale: (1) “Almost Never”; (2) “Sometimes”; (3) “Often”; or (4) “Almost Always.”

The STAI S-Anxiety and T-Anxiety scales are comprised of approximately equal numbers of anxiety-present and anxiety-absent items. The S-Anxiety scale consists of 10 anxiety-present items (e.g., “I feel frightened”; “I feel upset”) and 10 anxiety-absent items (e.g., “I feel calm”; “I feel relaxed”). The T-Anxiety scale consists of 11 anxiety-present items (e.g., “I feel nervous and restless”; “I have disturbing thoughts”) and 9 anxiety-absent items (e.g., “I feel secure”; “I am content”). Each anxiety-present item is given a direct score of 1 to 4, which is the score recorded on the test form. The anxiety-absent items are reverse scored (i.e., responses to items that are marked 1, 2, 3, or 4 are scored 4, 3, 2, or 1, respectively). Ratings of 4 for the direct and reverse-scored items indicate the presence of a high level of anxiety. The scores for the S-Anxiety and T-Anxiety scales are obtained by simply summing the scores for the items that make up each scale. Thus, the score range for the 20-item S-Anxiety and T-Anxiety scales can vary from a minimum of 20 to a maximum of 80.

When the STAI S-Anxiety and T-Anxiety scales are administered together, it is recommended that the S-Anxiety scale be given first, followed by the T-Anxiety scale. Scores on the S-Anxiety scale, which was designed to measure the intensity of anxiety as an emotional state, are sensitive to the conditions under which the test is administered, and may be influenced by the emotional climate that is created when the T-Anxiety scale is given first. In contrast, because the T-Anxiety scale is relatively stable and impervious to the conditions under which the test is given (e.g., Auerbach, 1973; Lamb, 1969; Spielberger, Auerbach, Wadsworth, Dunn, & Taulbee, 1973; Spielberger, 1983), scores on this scale are not likely to be influenced by having previously responded

to the S-Anxiety scale.

### **Theoretical Basis**

Freud (1924) defined anxiety as “something felt,” a specific unpleasant emotional state or condition that included feelings of apprehension, tension, worry, and physiological arousal, and equated fear with objective anxiety (Freud, 1936), which he considered to be an emotional reaction that was proportional in its intensity to a real danger in the external world.

Objective anxiety was generally beneficial because it served to warn the individual that some form of adjustment was necessary. Consistent with Darwin’s (1872/1965) evolutionary perspective, Freud (1936) emphasized the adaptive utility of anxiety in motivating behavior that helped a person to cope more effectively with threatening and potentially harmful situations.

Freud’s (1936) danger signal theory and Cattell’s concepts of state and trait anxiety (Cattell, 1966; Cattell & Scheier, 1958, 1963), as refined and elaborated by Spielberger (1966, 1972a, 1972b, 1977, 1979a), provided the conceptual framework that guided the construction of the STAI. As previously noted, S-Anxiety was defined as a psychobiological state or condition consisting of subjective feelings of tension, apprehension, nervousness, and worry, with associated activation (arousal) of the autonomic nervous system. T-Anxiety was defined in terms of relatively stable individual differences in anxiety proneness, as reflected in the frequency that anxiety states have been manifested in the past and the probability that S-Anxiety reactions will be experienced in the future.

### **Test Development**

The STAI was developed by Spielberger, Gorsuch, and Lushene (1970) to provide reliable, relatively brief, self-report scales for assessing both state and trait anxiety in research and clinical practice. When test construction began in 1964 (Spielberger & Gorsuch, 1966), the initial goal was to construct an inventory consisting of a single set of items that could be administered with different instructions to assess the intensity of anxiety as an emotional state and individual differences in anxiety as a personality trait. A large pool of items was adapted from existing anxiety measures. The essential psychological content of each adapted item was simplified and retained, but the format was modified so that the same items could be given with different instructions to assess either S-Anxiety or T-Anxiety. A number of new items were also written in keeping with the concepts of state and trait anxiety as previously defined.

The resulting pool of more than 60 anxiety items was administered to large samples of university students and psychiatric patients, first with state and then with trait instructions. When given with trait instructions, those items that correlated significantly with scores on two widely used anxiety measures, the Taylor (1953) Manifest Anxiety Scale (MAS) and Cattell and Scheier’s (1963) IPAT Anxiety Scale Questionnaire (ASQ), were retained for further study. On the basis of extensive item-validity research with more than 2,000 study participants, a final set of 20 items was selected for the preliminary form of the STAI.

Research with the preliminary STAI indicated that altering

the instructions could not overcome the strong state or trait psycholinguistic connotations of key words in some of the items (Spielberger et al., 1970). For example, “I worry too much” was stable over time and correlated highly with other T-Anxiety items, but scores on this item did not increase in response to stressful circumstances, nor did they decrease under relaxed conditions, as required for the construct validity of an S-Anxiety item. In contrast, “I feel upset” was a highly sensitive measure of S-Anxiety; item scores increased markedly under stressful conditions and were lower under relaxed conditions. However, when given with trait instructions, the scores for this item were unstable over time and correlations with other T-Anxiety items were relatively low. Given the difficulties encountered in measuring state and trait anxiety with the same items, the test-construction strategy for the STAI was modified, and separate sets of items were selected for assessing S-Anxiety and T-Anxiety. When given with trait instructions, the 20 items with the best concurrent validity, as indicated by the highest correlations with the MAS and the ASQ, and that were stable over time were selected for the original STAI (Form X) T-Anxiety scale (Spielberger et al., 1970). When given with state instructions, the 20 items with the best construct validity, as indicated by higher and lower scores, respectively, under stressful and nonstressful conditions, were selected for the S-Anxiety scale. Only 5 of the 40 items met the validity criteria for both scales. The remaining 30 items were relatively unique measures of either state or trait anxiety.

Following the publication of the STAI (Form X), insights gained in more than a decade of research stimulated a major revision in this inventory, with the goal of developing “purer” measures for assessing state and trait anxiety in adolescents and adults. Careful scrutiny of the content of the STAI items with the best psychometric properties resulted in clearer conceptual definitions of the state and trait anxiety constructs, which then guided the construction of potential replacement items for the revised STAI (Form Y). The item selection and validation procedures, described in detail in the STAI (Form Y) Test Manual (Spielberger, 1983), resulted in the replacement of 30% of the original STAI items.

In the construction and standardization of the STAI (Form Y), more than 5,000 additional subjects were tested. Factor analyses of the Form Y items identified distinct state and trait anxiety factors (Spielberger, Vagg, Barker, Donham, & Westberry, 1980), which were quite similar to those found in previous factor studies of the Form X items (Gaudry, Spielberger, & Vagg, 1975). When four factors were extracted in analyses of responses to Form Y, distinctive state and trait anxiety-present and anxiety-absent factors emerged that were similar to those reported in factor studies of Form X. However, the Form Y factors were more differentiated and had better simple structure than the corresponding Form X factors, reflecting a better balance in the number of T-Anxiety present and absent items (Spielberger et al., 1980).

#### **Psychometric Characteristics**

Detailed reliability data for the STAI (Form Y) are reported

in the Test Manual (Spielberger, 1983). Test-retest stability coefficients for the T-Anxiety scale were reasonably high for large groups of high school and college students, ranging from .73 to .86 over intervals of 20 to 104 days. In contrast, the test-retest coefficients for the S-Anxiety scale were relatively low, with a median  $r$  of only .33. However, this lack of stability for the S-Anxiety scale was both expected and considered desirable because valid measures of emotional states should reflect the influence of unique situational factors at the time of testing.

Since anxiety states vary in intensity as a function of perceived stress, internal consistency measures, such as alpha coefficients, provide a more meaningful index of reliability than test-retest correlations. Alpha coefficients for the STAI (Form Y) S-Anxiety scale, computed by Formula KR-20 as modified by Cronbach (1951), were .86 or higher with a median coefficient of .93 for large, independent samples of high school and college students, working adults, and military recruits (Spielberger, 1983). The alpha coefficients of the T-Anxiety scale for these groups were also uniformly high, with a median alpha of .90.

It should be noted that the distribution of scores for the STAI S-Anxiety scale, when given under neutral conditions, is positively skewed and approaches a normal distribution under stressful conditions. Consequently, alpha coefficients are generally higher when the S-Anxiety scale is given under conditions of psychological stress. For example, the alpha reliability was .94 when the S-Anxiety scale was administered to college students immediately after a distressing film with instructions to report how they felt while watching the film (Spielberger, 1983). For these same respondents, the alpha was .89 when the S-Anxiety scale was given following a brief period of relaxation training.

The individual STAI S-Anxiety and T-Anxiety items were required to meet stringent validity criteria at each stage of the test development process (Spielberger, 1983; Spielberger & Gorsuch, 1966; Spielberger et al., 1970). Although the Form X items were selected on the basis of significant correlations with the most widely used measures of anxiety at the time the inventory was developed (Spielberger et al., 1970), the content of several items that had been adapted from the MAS seemed to be more closely related to depression than anxiety (e.g., "I cry easily"; "I feel useless at times"; "I feel blue"). In developing the revised STAI (Form Y), the conceptual definitions of state and trait anxiety were improved, and items with depressive content were replaced by new items constructed in keeping with these improved definitions, which had better psychometric properties (Spielberger, 1983). Each item selected for the STAI S-Anxiety scale had to meet demanding criteria for construct validity. When compared with a neutral situation, the score for each S-Anxiety item had to increase significantly in a stressful situation and significantly decline in relaxing situations. Evidence of the construct validity of the S-Anxiety scale was demonstrated in findings that the scores of college students were significantly higher during examinations and lower after relaxation

training than when they were tested during a relatively nonstressful class period (Spielberger, 1983). Further evidence of the construct validity of the S-Anxiety scale was observed in military recruits whose S-Anxiety scores were much higher shortly after they began a highly stressful training program than those of high school and college students who were tested under nonstressful classroom conditions.

Relatively high correlations of scores on the revised STAI (Form Y) T-Anxiety scale with the ASQ and the MAS, ranging from .73 to .85 for college students and neuropsychiatric patients, provided evidence of concurrent validity and suggested that all three inventories measured trait anxiety. However, a major advantage of the 20-item STAI T-Anxiety scale is that it is less contaminated with depression items than the MAS, and requires less than half as much time to administer as compared with the 43-item ASQ and 50-item MAS.

#### **Range of Applicability, Computerization, and Accommodation for Diverse Populations**

In constructing and validating the STAI, more than 10,000 adolescents and adults have been tested. The major populations with whom the STAI has been used include high school and college students; working adults; psychiatric, psychosomatic, medical, surgical, and dental patients; military personnel; and prison inmates. Extensive normative data for most of these groups, including separate norms for females and males, are reported in the STAI (Form Y) Test Manual (Spielberger, 1983). The STAI may be administered with hand-scoreable test forms or with multiple-choice answer sheets that permit computer scoring. The machine-scoreable answer sheets can be read by an optical scanner, and test responses can be scored and evaluated by analysis software, such as spreadsheets and statistical packages. When the machine-scoreable test forms are used, special emphasis should be given to explaining differences in the instructions for the S-Anxiety and T-Anxiety scales. Most of the normative data reported in the Test Manual were obtained with machine-scoreable test forms.

The STAI has been found to have excellent psychometric properties for the assessment of anxiety in elderly persons (Patterson, O'Sullivan, & Spielberger, 1980). The inventory can be modified for persons with diminished visual acuity, which is common among the elderly, by printing the items in larger type (McDonald & Spielberger, 1983). For individuals who are completely visually impaired or persons with limited reading ability, the test items can be read to them by an examiner who marks the subject's responses on the test form.

The key words in most of the STAI (Form Y) items are at or below the sixth-grade reading level. Therefore, the inventory can be readily administered to high school students and younger adolescents. However, the State-Trait Anxiety Inventory for Children (STAIC), which was developed to measure anxiety in 9- to 12-year-old elementary school children (Spielberger, 1973), may be more effective for assessing anxiety in adolescents with emotional problems or reading difficulties (e.g., Finch, Kendall, Dannenburg, & Morgan, 1978;

Finch, Kendall, & Montgomery, 1976; Finch, Montgomery, & Deardorff, 1974). Extensive norms for fourth-, fifth-, and sixth-grade students are reported for the STAIC S-Anxiety and T-Anxiety scales in the STAIC Test Manual (Spielberger, 1973).

Since first introduced 35 years ago (Spielberger&Gorsuch, 1966), the STAI has been adapted in more than 60 different languages and dialects, and used extensively in research as reflected in citations in over 14,000 archival publications. Research with the STAI has stimulated a growing consensus among behavioral and medical scientists regarding the critical need to distinguish between the concepts of stress and anxiety. It has also contributed to recognizing the importance of distinguishing between anxiety as a transitory emotional state and individual differences in anxiety proneness as a relatively stable personality trait. **Use in Clinical Practice and Organizational Settings**

The STAI has been used in numerous clinical studies of substance abuse; psychiatric, psychosomatic, and medical disorders; and in experimental investigations of the effects of anxiety on a wide range of psychological processes, such as attention, memory, learning, perception, and academic aptitude and achievement (Spielberger, 1989). It has also been used extensively in research on situation-specific anxiety phenomena, such as test anxiety, speech anxiety, and anxiety in sports competition. The T-Anxiety scale has proved useful as an outcome measure in research on the effectiveness of relaxation training, systematic desensitization, biofeedback, and various forms of behavioral and cognitive treatment. Evidence of the construct validity of the STAI (Form Y) T-Anxiety scale is reflected in the high scores of various neuropsychiatric (NP) patient groups for whom anxiety is a major symptom (American Psychiatric Association, 1994). Except for personality disorders, all NP diagnostic groups have substantially higher T-Anxiety scores than normal subjects (Spielberger, 1983). General medical and surgical (GMS) patients with psychiatric complications also have higher T-Anxiety scores than GMS patients without such complications, indicating that the T-Anxiety scale can help to identify nonpsychiatric patients with emotional problems. The lower T-Anxiety scores of patients with personality disorders, for whom the absence of anxiety is an important defining characteristic, provides further evidence of the construct and discriminant validity of the STAI.

While most studies with the STAI have been conducted by psychologists or medical researchers, the inventory has also been widely used by investigators from other disciplines, which include education, counseling and guidance, speech and hearing, criminal justice, nursing, and sports psychology. In addition, the STAI has proved useful in research in anthropology, sociology, political science, government, fine arts, and musical performance. References to studies in these fields may be found in the Comprehensive Bibliography of Research with the STAI (Spielberger, 1989).

## **STATE-TRAIT ANGER EXPRESSION**

### **INVENTORY-2 (STAXI-2)**

#### **Test Description**

The 57-item STAXI-2 includes 42 of the 44 original STAXI items, plus 15 new items that were constructed for this measure. Brief descriptions of the 6 scales, 5 subscales, and the Anger Expression Index (AX Index) that are included in the STAXI-2 are briefly described in Table 6.1. The number of items and the range of scores for each scale and subscale are also reported in this table. Since the STAXI-2 items are all direct scored, the scale and subscale scores are obtained by simply adding the scores for the items comprising each measure.

The STAXI State-Anger (S-Anger) scale was expanded from 10 to 15 items, with three factorially derived 5-item subscales that assess: Feeling Angry (S-Ang/F, e.g., “I feel annoyed”); Feel Like Expressing Anger Verbally (S-Ang/V, e.g., “I feel like yelling at somebody”); and Feel Like Expressing Anger Physically (S-Ang/P, e.g., “I feel like hitting someone”). The 10-item STAXI-2 Trait-Anger scale (T-Anger, e.g., “I am a hotheaded person”), and the 8-item AX/Out and AX/In scales (e.g., “I argue with others”; “I withdraw from people”) remain the same as in the original STAXI. The 4-item subscales for assessing T-Anger Temperament (T-Ang/T, e.g., “I have a fiery temper”) and T-Anger Reaction (T-Ang/R, e.g., “I feel furious when criticized in front of others”) are also unchanged.

The STAXI-2 AX/Con-Out scale (e.g., “I control my temper”), which assesses the control of angry feelings by preventing the expression of anger toward other persons or objects in the environment, is comprised of 7 of the 8 items from the original STAXI AX/Con scale, plus 1 replacement item. An entirely new 8-item AX/Con-In scale (e.g., “I try to relax”) was constructed for the STAXI-2 to assess how often a person tries to control suppressed anger by reducing its intensity. The AX Index, which provides a measure of total anger expression, is computed by using the following formula in which the constant, 48, was included to eliminate negative numbers:  $AX\ Index = AX/Out + AX/In + (AX/Con-Out + AX/Con-In) - 48$ .

The STAXI-2 was designed to be self-administering and may be given individually or to groups of respondents. The S-Anger and T-Anger scales are printed on the front side of a single-page test form; the anger expression and control scales are on the back of the test form. The STAXI-2 has no set time limit and generally requires only 12 to 15 minutes for completion. In responding to the S-Anger and T-Anger scales, examinees first rate the *intensity* and then the *frequency* that they experience angry feelings on the same 4-point scales that are used to assess state and trait anxiety with the STAI. The response choices for the S-Anger scale are: (1) “Not at All”; (2) “Somewhat”; (3) “Moderately So”; and (4) “Very Much So.” For the T-Anger scale, the response choices are: (1) “Almost Never”; (2) “Sometimes”; (3) “Often”; and (4) “Almost Always.”

Although the four Anger Expression and Control scales are trait measures, the instructions for these scales are quite different from those for the T-Anger scale. Examinees are **State-Trait Anger Expression Inventory-2 (STAXI-2)**

**TABLE 6.1 Brief Overview of the STAXI-2 Scales and Subscales**

STAXI-2 Scale/Subscale	Number of Items	Scale/Subscale	Range	Description of Scale/Subscale
State Anger ( <i>S-Anger</i> )	15	15–60	Measures the intensity of angry feelings and the extent to which a person feels like expressing anger at a particular time	
Feeling Angry ( <i>S-Ang/F</i> )	5	5–20	Measures the intensity of the angry feelings the person is currently experiencing	
Feel Like Expressing Anger Verbally ( <i>S-Ang/V</i> )	5	5–20	Measures the intensity of current feelings related to the verbal expression of anger	
Feel Like Expressing Anger Physically ( <i>S-Ang/P</i> )	5	5–20	Measures the intensity of current feelings related to the physical expression of anger	
Trait Anger ( <i>T-Anger</i> )	10	10–40	Measures how often angry feelings are experienced over time	
Angry Temperament ( <i>T-Ang/T</i> )	4	4–16	Measures the disposition to experience anger without specific provocation	
Angry Reaction ( <i>T-Ang/R</i> )	4	4–16	Measures the frequency that angry feelings are experienced in situations that involve frustration and/or negative evaluations	
Anger Expression-Out ( <i>AX-Out</i> )	8	8–32	Measures how often angry feelings are expressed in verbally or physically aggressive behavior	
Anger Expression-In ( <i>AX-In</i> )	8	8–32	Measures how often angry feelings are experienced but not expressed (suppressed)	
Anger Control-Out ( <i>AX/Con-Out</i> )	8	8–32	Measures how often a person controls the outward expression of angry feelings	
Anger Control-In ( <i>AX/Con-In</i> )	8	8–32	Measures how often a person attempts to control angry feelings by calming down or cooling off	
Anger Expression Index ( <i>AX Index</i> )	32	0–96	Provides a general index of anger expression based on responses to the AX-Out, AX-In, AX/Con-Out, and AX/Con-In items	
			informed that “Everyone feels angry or furious from time to time, but people differ in the ways that they react when they are angry.” Respondents are then instructed to indicate how often they generally react or behave in the manner described by each item when they feel angry by rating themselves on the same 4-point frequency scale used in responding to the T-Anger items.	

**Theoretical Basis**

Anger (rage) was considered by Charles Darwin (1872/1965, p. 74) to be a powerful emotion that motivated “animals of all kinds, and their progenitors before them, when attacked or threatened by an enemy,” to fight and defend themselves. For Darwin, anger was a state of mind that differed “from rage only in degree, and there is no marked distinction in their characteristic signs” (1872/1965, p. 244). He also observed that rage, which often resulted in violent behavior, was

reflected in facial expressions (e.g., reddened face, clenched teeth, dilated nostrils), accelerated heart rate, and muscular tension. Thus, anger was implicitly defined as a psychobiological emotional state that varied in intensity, from mild irritation or annoyance to intense fury and rage (Spielberger, 1999).

Freud (1933/1959) considered aggression to be a biologically determined instinctual drive that motivated hatred and aggressive behavior. If aggression could not be expressed against external objects, it was turned back into the self, resulting in depression and other psychosomatic manifestations (Alexander & French, 1948; Freud, 1936). The maladaptive effects of anger, hostility, and aggression are traditionally emphasized as important contributors to the etiology of the psychoneuroses, depression, and schizophrenia. Research findings also indicate that anger and hostility contribute to the pathogenesis of hypertension (e.g., Crane, 1981; Harburg, Blakelock, & Roeper, 1979) and coronary heart disease (Friedman & Rosenman, 1974; Matthews, Glass, Rosenman, & Bortner, 1977; Spielberger & London, 1982).

Although much has been written about the negative impact of anger and hostility on physical and psychological well-being, definitions of these constructs are ambiguous and sometimes contradictory. Moreover, anger, hostility, and aggression are often used interchangeably in the research literature, resulting in conceptual confusion, which is reflected in a diversity of measurement operations of questionable validity (Biaggio, Supplee, & Curtis, 1981).

Given the substantial overlap in prevailing conceptual definitions of anger, hostility, and aggression, and in the variety of operational procedures that are used to assess these constructs, we have referred to them, collectively, as the AHA! Syndrome (Spielberger et. al., 1985).

The concept of anger generally refers to an emotional state that consists of feelings that vary in intensity, with associated activation or arousal of the autonomic nervous system. Although hostility is almost always accompanied by angry feelings, this concept also has the connotation of a complex set of attitudes and behaviors that include being mean, vicious, vindictive, and often cynical (Spielberger et. al., 1985). Aggression as a psychological construct is defined as destructive or punitive behavior directed toward other persons or objects in the environment (Buss, 1961). While hostility and the behavioral manifestations of aggression have been investigated in numerous studies, psychometric measures of hostility tend to confound angry feelings with aggressive behavior, and anger as an emotional state has been largely neglected in psychological research.

### **Test Development**

The STAXI was constructed to assess the experience, expression, and control of anger. The initial step in developing this inventory was the construction of the State-Trait Anger Scale (STAS; Spielberger, 1980; Spielberger, Jacobs, Russell, & Crane, 1983), which was modeled after the STAI. The STAS was designed to assess the intensity of anger as an emotional state (S-Anger) and individual differences in anger proneness as a personality trait (T-Anger). S-Anger was defined

as a psychobiological state or condition, consisting of subjective feelings that varied in *intensity* from mild irritation or annoyance to intense fury and rage, with associated activation of the autonomic nervous system. T-Anger was defined in terms of individual differences in the *frequency* that S-Anger was experienced over time. Guided by these working definitions, a pool of items was constructed to assess the intensity of angry feelings at a particular time and individual differences in anger proneness as a personality trait. The preliminary form of the STAS, consisting of 15 S-Anger and 15 T-Anger items, was administered to a large sample of university students (Spielberger et al., 1983). Alpha coefficients for the preliminary S-Anger and T-Anger scales were .93 and .87, respectively, indicating a high degree of internal consistency, and providing impressive evidence of the utility of the working definitions of anger that guided the item-construction process. Test-retest reliability coefficients for the STAS T-Anger scale over a 2-week interval were .70 and .77 for males and females, respectively (Jacobs, Latham, & Brown, 1988). In contrast, the stability coefficients for the STAS S-Anger subscale were much lower (.27 for males, .21 for females), as would be expected for a measure of a transitory emotional state.

Given the high internal consistency of the preliminary STAS S-Anger and T-Anger scales, it was possible to reduce the length of these scales from 15 to 10 items without unduly weakening their psychometric properties. The final set of 10 S-Anger and 10 T-Anger items was comprised of those items with the largest item-remainder correlations (.50 or higher) and the best content validity. Correlations between the 10- and 15-item S-Anger and T-Anger scales, ranging from .95 to .99 for college students and navy recruits, indicated that the 10-item S-Anger and T-Anger scales provided essentially the same information as the longer forms (Spielberger, 1988). Factor analyses of the STAS S-Anger items identified only a single underlying factor for both males and females, providing evidence that the S-Anger scale measured a unitary emotional state. In contrast, factor analyses of the T-Anger items consistently identified two substantially correlated factors, which provided the basis for developing subscales to measure Angry Temperament (T-Ang/T) and Angry Reaction (T-Ang/R). The items comprising the T-Ang/T subscale (e.g., "I have a fiery temper") described the experience of anger without specifying any provocation. The T-Ang/R subscale items (e.g., "It makes me furious when I am criticized in front of others") described angry reactions to situations that involved frustration and/or negative evaluations.

The concurrent validity of the STAS T-Anger scale was evaluated by administering this scale to large samples of college students and navy recruits, along with the Buss-Durkee (1957) Hostility Inventory (BDHI) and the Hostility (HO; Cook & Medley, 1954) and Overt Hostility (Hv; Schultz, 1954) scales. Moderately high positive correlations of the STAS T-Anger scale with the BDHI and the HO scales (*Mdn.*  $r = .63$ ) were found for males and females in both groups (Spielberger, 1988), providing evidence of a strong relationship

between T-Anger and hostility. Moderate positive correlations of the STAS T-Anger scale were also found with the Neuroticism scale of the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) and the T-Anxiety scale of the State-Trait Personality Inventory (Spielberger, 1979b). These results were consistent with clinical observations that neurotic individuals frequently experience guilt and anxiety about their angry feelings.

The results of a study of hypertensive patients clearly demonstrated that the two T-Anger subscales measured different facets of trait anger (Crane, 1981). The T-Anger scores of hypertensive patients were significantly higher than those of medical and surgical patients with normal blood pressure,

but this difference was due entirely to the substantially higher T-Ang/R scores of the hypertensives. The T-Ang/T scores of the hypertensives were essentially the same as those of the patients with normal blood pressure. Crane also found that the S-Anger scores of the hypertensives while performing on a mildly frustrating task were higher than the corresponding scores for patients with normal blood pressure.

In a series of studies, Deffenbacher (1992) found that individuals with high STAS T-Anger scores reported that they experienced intense angry feelings with greater frequency than persons low in T-Anger did. The high T-Anger individuals also reported experiencing anger-related physiological symptoms two to four times more often than those who were low in T-Anger, and a stronger tendency to both express and suppress their anger when provoked. In addition, negative events, such as failure experiences, seemed to have a more devastating (catastrophizing) impact on high T-Anger individuals (Story & Deffenbacher, 1985).

As our research on anger has progressed, the critical importance of differentiating between the *experience* of anger and the characteristic ways in which anger was *expressed* became increasingly apparent (Spielberger et al., 1985). In a study of anger expression, Funkenstein, King, and Drolette (1954) classified participants as either anger-in or anger-out on the basis of how they responded to being harassed. Thus, anger expression was implicitly defined as a unidimensional, bipolar construct that described individual differences in holding anger in or the expression of anger in aggressive behavior. The Anger Expression (AX) scale was constructed to assess this dimension.

The first step in constructing the AX scale was to review the relevant research literature and develop working definitions of anger-in and anger-out. Anger-in was defined in terms of how often an individual experienced but held in (suppressed) angry feelings. Anger-out was defined as the frequency with which an individual expressed angry feelings in verbal or physically aggressive behavior. However, rather than assigning subjects to dichotomous anger-in or anger-out categories as in previous research (Funkenstein et al., 1954), the AX scale was designed to measure a continuum of individual differences in how often anger was held in or expressed.

The rating format for the AX scale was the same as for the T-Anger scale, but the instructions differed markedly

from those used in assessing T-Anger. Instead of simply asking respondents to indicate how they generally feel, they were informed that “everyone feels angry or furious from time to time,” and instructed to report “how often you generally react or behave in the manner described *when you feel angry or furious*” (Spielberger, Reheiser, & Sydeman, 1995, p. 58). A preliminary form of the AX scale was administered to 1,060 high school students (Johnson, 1984). The results of separate factor analyses of the responses of females ( $N = 459$ ) and males ( $N = 601$ ) clearly indicated that the anger expression items were tapping two independent anger-in and anger-out dimensions. Given the strength and clarity of the anger-in and anger-out factors and the striking similarity of these factors for males and females, the items with the strongest loadings on each factor were selected to form 8-item Anger-In (AX/In) and Anger-Out (AX/Out) scales (Spielberger et al., 1985). The AX/In items (e.g., “I keep things in”; “I boil inside, but I don’t show it”) had uniformly high loadings for both sexes on the anger-in factor ( $Mdn. = .665$ ), and negligible loadings on the anger-out factor ( $Mdn. = .045$ ). Similarly, the items selected for the AX/Out scale (e.g., “I lose my temper”; “I strike out at whatever infuriates me”) had high loadings for both sexes on the anger-out factor and negligible loadings on anger-in; the median item loadings on these factors were .59 and .01.

The correlations between the AX/In and AX/Out scales were essentially zero for both males and females for large samples of high school and college students (Johnson, 1984; Pollans, 1983) and for other populations (Dembroski, MacDougall, Williams, & Haney, 1985; Knight, Chisholm, Paulin, & Waal-Manning, 1988; Spielberger, 1988). The internal consistency of the 8-item AX/In and AX/Out scales, when evaluated by alpha coefficients, ranged from .73 to .84; the test-retest stability of these scales over varying time periods ranged from .64 to .86 (Jacobs et al., 1988). Thus, the AX/In and AX/Out subscales are factorially orthogonal, empirically independent, and internally consistent. Clearly, these scales assess two distinct anger expression dimensions that are stable over time.

The original item pool for the AX scale included three items (“Control my temper”; “Keep my cool”; “Calm down faster”) that were included to measure the middle range of an anger-in/anger-out continuum. Since all three items had substantial loadings on *both* the anger-in and anger-out factors, they were retained in the 20-item AX scale, but were not included in the AX/In or AX/Out scales. In subsequent research, these items coalesced to form the nucleus of an anger control factor (Spielberger, 1983), which stimulated the development of the AX/Con scale. Dictionary and thesaurus definitions of “control” and idioms pertaining to the control of anger were consulted in constructing additional anger control items, which were administered to a large sample of university students along with the 20-item AX scale. In separate factor analyses for males and females, a strong anger control factor was identified. Those items with the highest loadings

on this factor for both males and females, which included the three original anger control items, were selected to form the 8-item AX/Con scale.

The revised 24-item AX scale, which included 8-item AX/In, AX/Out, and AX/Con scales, was administered to a large sample of university students (Spielberger, Krasner, & Solomon, 1988). In factor analyses of responses to the AX scale items, an anger control factor was the strongest to emerge for both males and females; all 8 AX/Con items had dominant salient loadings on this factor for both sexes. As in previous research, well-defined anger-in and anger-out factors were also found; all of the items in the AX/In and AX/Out scales had dominant salient loadings on the appropriate factor. The AX/Con scale correlated negatively with AX/Out for both males ( $r = -.59$ ) and females ( $r = -.58$ ). The correlations of the AX/In scale with the AX/Con and AX/Out scales were essentially zero for both sexes (Pollans, 1983; Spielberger, 1988).

Moderately high positive correlations of the AX/Out scale with the T-Ang/T subscale, along with small positive correlations of both the AX/Out and AX/In scales with the T-Ang/R subscale, suggested that persons with an angry temperament were more likely to express their anger outwardly rather than suppress it, whereas individuals who frequently experienced anger when they were frustrated or treated unfairly were equally likely to suppress or outwardly express their anger (Spielberger, 1988). Small but significant positive correlations of both the AX/In and AX/Out scales with the STPI T-Anxiety scale suggested that individuals who frequently suppressed or expressed their anger experienced anxiety more often than individuals with low anger expression scores. The correlations of all three anger expression scales with the STPI T-Curiosity scale were essentially zero, providing evidence of discriminant validity.

The 20-item STAS and 24-item AX scale were combined to form the *State-Trait Anger Expression Inventory* (Spielberger, 1988). Fuqua et al. (1991) administered the STAXI to a large sample of college students and factored their responses to the 44 individual items. Of the seven factors identified by Fuqua et al. (1991), the first six corresponded almost exactly with the STAXI S-Anger, AX/In, AX/Out, and AX/Con scales and the T-Ang/T and T-Ang/R subscales, which were defined by separate factors. Almost all of the STAXI scale and subscale items had salient loadings on the appropriate factors and negligible loadings on the other factors.

The relatively weak seventh factor identified by Fuqua et al. (1991) was defined by the salient loadings of three S-Anger items (Feel like . . . breaking things, banging on the table, hitting someone), which also had strong loadings on Factor I, along with the other seven S-Anger items. Van der Ploeg (1988) factored the responses of male military draftees to a Dutch adaptation of the 10-item S-Anger scale, and found two S-Anger factors that were similar to those reported by Fuqua et al. These findings suggested that there might be a second S-Anger factor, defined by items with content that reflected feeling like expressing anger. Additional factor analyses

of the 44 STAXI items provided further evidence of two distinct but highly correlated S-Anger factors (Forgays, Forgays, & Spielberger, 1998), which were labeled Feeling Angry (e.g., "I feel furious") and Feel Like Expressing Anger (e.g., "I feel like hitting someone") on the basis of the content of the items with dominant salient loadings on each factor. In developing the STAXI-2, as previously noted, the original STAXI S-Anger scale was expanded from 10 to 15 items, with three factorially derived 5-item subscales: Feeling Angry (S-Ang/F), Feel Like Expressing Anger Verbally (S-Ang/V), and Feel Like Expressing Anger Physically (S-Ang/P). The STAXI-2 T-Anger, AX/In and AX/Out scales are the same as in the original STAXI. The STAXI-2 AX/Con-Out scale is comprised of 7 of the 8 original STAXI AX/Con items plus 1 replacement item, and an entirely new 8-item scale was constructed for the STAXI-2 to assess how often a person tries to control anger-in (AX/Con-In) by reducing the intensity of suppressed anger.

### **Psychometric Characteristics of the STAXI-2**

The normative samples for the STAXI-2 are based on the responses of more than 1,900 individuals from two populations: a heterogeneous sample of 1,644 normal adults (977 females, 667 males) and 274 hospitalized psychiatric patients (103 females, 171 males). The sample of normal adults included managerial, technical, and clerical personnel; participants in stress management programs; health care managers and professionals; insurance company employees; and undergraduate and graduate students enrolled in a large urban university. The data for the psychiatric patients were obtained as part of routine psychological testing, which was completed at the time of their admission into a hospital program for treating psychiatric and substance abuse problems.

The internal consistency of the STAXI-2 scales and subscales for the normal adults and psychiatric patients, as measured by alpha coefficients, ranged from .73 to .95 (*Mdn. r* = .87), and was not influenced by either gender or psychopathology. Alpha coefficients for the AX Index, ranging from

.75 to .82, indicated satisfactory internal consistency for this measure, which is based on scores on the anger expression and control scales rather than computed directly from item ratings. Normative data for male and female adolescents and adults and psychiatric patients are reported in the STAXI-2 Test Manual (Spielberger, 1999). Norms are also reported in the original STAXI Test Manual (Spielberger, 1988) for the

following special interest groups: general medical and surgical patients, prison inmates, and military recruits.

In analyses of the data for STAXI-2 normative samples, significant gender differences were found for the AX Index and the AX/Out and AX/Con-In scales. Males had higher AX Index and AX/Out scores than females and lower AX/Con-In scores, indicating that the males were more likely than females to express anger toward other persons and less likely to control suppressed anger by reducing its intensity. As expected, psychiatric patients had significantly higher scores than normal adults on the AX Index and the S-Anger, T-Anger, and AX/In scales, and significantly lower scores on the AX/Con-

Out and AX/Con-In scales. These differences indicated that psychiatric patients experience and suppress anger more frequently than normal adults and have less control of their anger.

### **Range of Applicability, Computerization, and Accommodation for Diverse Populations**

The STAXI has proved useful for assessing the experience, expression, and control of anger in normal individuals and psychiatric patients (Deffenbacher, 1992; Moses, 1992; Spielberger, 1999), and for evaluating anger and its effects on a variety of disorders, including alcoholism, hypertension, coronary heart disease, and cancer (Spielberger, 1988, 1999). The STAXI-2 may be administered with multiple-choice answer sheets that permit machine scoring. When these are used, special attention should be given to explaining the differences in the instructions for the three parts of the inventory, especially the instructions for the anger expression and control scales in Part 3. In scoring the responses, the file created by an optical scanner can be read by analysis software, such as spreadsheets and statistical packages.

Since the STAXI anger expression and control scales were developed with high school students and the key words in most of the items are at the sixth-grade reading level or below, the inventory can be readily administered to junior high school students. The STAXI-2 can also be modified for persons with diminished visual acuity, which is a common problem among older persons, by reprinting the items in larger type. For persons who are severely visually impaired or have limited reading skills, the items can be read to them by an examiner, who can mark the subject's responses on the test form.

Comparing the STAXI scale and subscale scores with appropriate norms is an important first step in test interpretation.

Percentile ranks, corresponding to the scale and subscale scores that are reported in the Test Manual (Spielberger, 1999), provide information on how a particular person compares with other individuals who are similar in age and gender. Scores between the 25th and 75th percentiles fall in what may be considered the normal range.

While individuals with scale scores that approach the 75th percentile are more prone to experience, express, or suppress their anger than those with scores below the median, such differences are generally not sufficient to detect persons whose anger problems may dispose them to develop physical or psychological disorders (Spielberger, 1999). Individuals with anger scores above the 75th percentile are likely to experience and/or express angry feelings that interfere with optimal functioning and dispose them to develop psychological or physical disorders. General guidelines for interpreting high scores for each STAXI-2 scale and subscale are provided in Table 6.2.

### **Use in Clinical or Organizational Practice**

The STAS and the AX scales have been used extensively in research on the relationship between anger and health (Brooks, Walfish, Stenmark, & Canger, 1981; Johnson & Broman, 1987; Schlosser, 1986; Vitaliano, 1984). High scores on

the STAXI AX/In scale have been consistently found to be related to elevated blood pressure and hypertension (e.g., Crane, 1981; Johnson, 1984; Johnson, Spielberger, Worden, & Jacobs, 1987; Kearns, 1985; Schneider, Egan, & Johnson, 1986; Spielberger et al., 1988; van der Ploeg, van Buuren, & van Brummelen, 1988). High scores on the AX/Out scale (above the 90th percentile) for persons who are also high in anger control may place an individual at risk for arteriosclerosis and heart attacks (Spielberger, 1999).

Johnson (1984) investigated the relationship between anger expression and blood pressure (BP) in a large sample of high school students ( $N=1,114$ ). Measures of systolic (SBP) and diastolic (DBP) blood pressure were obtained during the same class period in which these students responded to the AX scale. The correlations of the AX/In scores with SBP and DBP were positive, curvilinear, and significant for both females and males. Students with very high AX/In scores had much higher SBP. Because the negative correlations of the AX/Out scores with BP were quite small, the overall pattern of correlations indicated that higher blood pressure was associated with holding anger in. After partialing out the influence of a number of variables related to BP in previous research (e.g., height, weight, dietary factors [salt intake], racial differences, family history of hypertension), the AX/In scores were still positively and significantly associated with elevated SBP and DBP, and were better predictors of blood

## **TABLE 6.2 Guidelines for Interpreting High STAXI Scores**

### **Scale Characteristics of Persons with High Scores**

**S-Anger:** Individuals with high scores are experiencing relatively intense angry feelings at the time the test was administered. If S-Anger is elevated relative to T-Anger, the individual's angry feelings are likely to be situationally determined. Elevations in S-Anger are more likely to reflect chronic anger if T-Anger and AX/In scores are also high.

**T-Anger:** High T-Anger individuals frequently experience angry feelings, especially when they feel they are treated unfairly by others. Whether persons high in T-Anger suppress, express, or control their anger can be inferred from their scores on the AX/In, AX/Out, and AX/Con scales.

**T-Anger/T:** Persons with high T-Anger/T scores are quick tempered and readily express their anger with little provocation. Such individuals are often impulsive and lacking in anger control. High T-Anger/T individuals who have high AX/Con scores may be strongly authoritarian and use anger to intimidate others.

**T-Anger/R:** Persons with high T-Anger/R scores are highly sensitive to criticism, perceived affronts, and negative evaluation by others. They frequently experience intense feelings of anger under such circumstances.

**AX/In:** Persons with high AX/In scores frequently experience intense angry feelings but tend to suppress these feelings rather than to express them

either physically or in verbal behavior. Persons with high AX/In scores who also have high AX/Out scores may express their anger in some situations, while suppressing it in others.

**AX/Out:** Persons with high AX/Out scores frequently experience anger, which they express in aggressive behavior. Anger-out may be expressed in physical acts such as assaulting other persons or slamming doors, or verbally in the form of criticism, sarcasm, insults, threats, and the extreme use of profanity.

**AX/Con-Out:** Persons with high AX/Con-Out scores tend to expend a great deal of energy in monitoring and preventing the outward experience and expression of anger. Although controlling outward or external manifestations of anger may be desirable, overcontrol can lead to passivity, depression, and withdrawal. Persons with high AX/Con-Out and T-Ang scores combined with low AX/Out scores are likely to experience these problems due to their chronic anger without an easy way to express that anger.

**AX/Con-In:** Persons with high AX/Con-In scores expend a great deal of energy in calming down and reducing their anger as soon as possible. The development of internal controls over the experience and expression of anger is generally seen in a positive light, but it can reduce the person's awareness of the need to respond with assertive behavior when this might facilitate a constructive solution to a frustrating situation. However, if a low AX/Con-In score is combined with high AX-Out and AX-In scores, there may be significant risk of developing medical problems.

pressure than any of the other measures for both males and females.

The STAXI has also been used to assess the anger experienced by patients undergoing treatment for Hodgkin's disease and lung cancer (McMillan, 1984); to investigate the relation of anger to type A behavior (Booth-Kewley & Friedman, 1987; Janisse, Edguer, & Dyck, 1986; Krasner, 1986; Spielberger et al., 1988); and to examine relationships among hardiness, well-being, and coping with stress (Schlosser & Sheeley, 1985). Kinder and his colleagues (Curtis, Kinder, Kalichman, & Spana, 1988; Kinder, Curtis, & Kalichman, 1986) used the STAXI scales in a series of studies of chronic pain, and Stoner (1988) has investigated the effects of marijuana use on the experience and expression of anger.

## **THE STATE-TRAIT PERSONALITY INVENTORY (STPI)**

### **Test Description**

The fundamental importance of anxiety, anger, depression, and curiosity as emotional states that motivate a wide range of behaviors makes clear the need to assess both the intensity of these emotions and individual differences in how frequently they are experienced as personality traits. The 80-item STPI is comprised of eight 10-item scales for assessing

anxiety, anger, depression, and curiosity as emotional states and personality traits. The STPI state and trait anxiety scales are comprised of the best items from the STAI (Form Y). The STPI state and trait anger scale and subscales are the same as in the original STAXI. The 40 state items are listed on the front of the single-page STPI test form; the 40 trait items are listed on the back of the test form. The theoretical basis and the construction and development of the state and trait depression and curiosity scales, which are unique to the STPI, are described in detail below.

Designed to be self-administering, the STPI has no time limits and may be given either individually or to groups of respondents. The instructions and rating format for the STPI are the same as for the STAI and STAXI state and trait scales, as previously described. The scoring procedure for the STPI items is also the same as that used in scoring the STAI and STAXI state and trait anxiety and anger scales. The STPI anxiety, depression, and curiosity items describe either the presence or absence of these emotional states and the corresponding personality traits; only the presence of angry feelings is evaluated by STPI state and trait anger items. The score for each STPI scale and subscale is the sum of the direct- and reverse-scored items comprising each measure.

#### **Theoretical Basis**

The theoretical basis for the STPI anxiety and anger scales, which is essentially the same as for the STAI and STAXI state and trait scales, was previously described. The concept of depression can be traced back to the fifth-century B.C. writings of Hippocrates, the father of modern medicine (Jackson, 1986, 1995). The Greek term *melancholia*, which had the connotation of both anxiety and depression, was used by Hippocrates to describe a “black mood” that involved prolonged fear and sadness (Jackson, 1995). Both Darwin and Freud considered depression to result from the interaction of anxiety and anger. Symptoms of depression vary in severity, from feeling sad or gloomy for a relatively short period of time, to deep despair, extreme guilt, hopelessness, and thoughts of death that could result in suicide. Persistent depression can also produce behavioral and physical symptoms such as fatigue, insomnia, impotence, frequent crying, chronic aches and pain, and excessive gain or loss in weight (Rosenfeld, 1999).

The insecurity caused by anxiety was viewed by Freud (1933/1959, 1936) as a major instigator of exploratory behavior, thus linking curiosity to anxiety. Although Freud did not directly address the origins of curiosity, he considered exploratory behavior to be determined by instinctive biological urges and ego mechanisms that served to reduce threat and insecurity (Aronoff, 1962). William James (1890), who was strongly influenced by Darwin’s (1872/1965) views on evolution, also proposed an instinct theory of curiosity, which posited that attraction to a novel stimulus was adaptive because it facilitated survival. However, fear (anxiety) aroused by novel situations inhibited curiosity but was also adaptive because the novel stimulus might prove to be dangerous. Thus, like Freud, James recognized a potentially antagonistic relationship between curiosity and fear, which

often resulted in the simultaneous arousal of these two emotions.

Curiosity and exploratory behavior have been linked to a variety of motivational constructs such as instincts, drives, and intrinsic motivation (Voss & Keller, 1983). Curiosity has been defined as a primary drive that guides and directs exploratory behavior (Cofer & Appley, 1964; Harlow, 1953), and as an acquired or secondary drive that is learned as a result of the reduction of primary drives such as hunger or thirst (Dashiell, 1925; Dollard & Miller, 1950). Although the literature on curiosity is characterized by diverse theoretical perspectives and contradictory empirical findings, curiosity clearly influences exploratory behavior and may be regarded as positive indicator of psychological health.

### **Test Development**

The STPI State (S-Dep) and Trait (T-Dep) scales were constructed to assess the presence and absence of affective feelings of depression. A pool of 40 items that described depressive feelings and cognitions was adapted from four widely used measures of depression: the Beck Depression Inventory (BDI; Beck & Steer, 1987; Beck, Steer, & Brown, 1996), the Zung Self-Rating Depression Scale (ZUNG; Zung, 1965, 1986), the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977), and the Depression scale of Zuckerman and Lubin's (1985) Multiple Affect Adjective Check List. The content and wording of each item followed, as closely as possible, the description of manifestation of depression in the measure from which it was adapted.

The depression items were administered to a large sample of university students. In responding to each item, the study participants first rated the intensity of their feelings of depression at a particular time, then rated how often depression was experienced on the same 4-point scales that are used in assessing state and trait anxiety and anger. Factor analysis of responses to the state and trait depression items identified strong depression-present and depression-absent factors for both females and males. The items with the strongest loadings on each factor for both sexes provided the basis for forming five-item state (S-Dep) and trait (T-Dep) depression-present and depression-absent subscales, which are referred to as Dysthymia and Euthymia.

The alpha coefficients for the S-Dep and T-Dep scales, and the S-Dep and T-Dep Dysthymia and Euthymia subscales, were .81 or higher for both sexes (*Mdn.  $r_{.90}$* ). The correlations of the T-Dep scale with the BDI, ZUNG, and CES-D for females and males ranged from .72 to .85 (*Mdn.  $r_{.78}$* ) and were significantly higher than the corresponding correlations of the S-Dep scale with these measures (*Mdn.  $r_{.66}$* ). While the BDI, Zung, and CES-D assess both state and trait depression (Ritterband & Spielberger, 1996), more persistent trait-like depressive characteristics appear to be assessed by all three of these widely used depression measures.

The T-Dep scale and all three widely used depression measures correlated highly with STPI T-Anxiety scale, as was expected, given the high rate of comorbidity for depression and anxiety (Gotlib & Crane, 1989; Mineka, Watson, &

Clark, 1998). The S-Dep scale also correlated substantially with T-Anxiety, but to a lesser degree. In addition, all five depression measures correlated positively and significantly with the STPI T-Anger scale, but these correlations were substantially smaller than with T-Anxiety. Significant negative correlations of the depression scales with the STPI T-Curiosity scale suggested that depression inhibits curiosity, which was consistent with observations of an antagonistic relation between curiosity and fear (Freud, 1936; James, 1890).

The State-Trait Curiosity Inventory (STCI) developed by Spielberger and Butler (1971), and subsequently revised (Spielberger et al., 1979; Spielberger, Peters, & Frain, 1981), measures the intensity of curiosity as a transitory emotional state (S-Curiosity) and individual differences in curiosity as a relatively stable personality trait (T-Curiosity). Subjects respond to the STCI S-Curiosity and T-Curiosity scales by rating the intensity and the frequency of their thoughts and feelings related to curiosity and exploratory behavior on the same 4-point rating scales that are used to assess state and trait anxiety, anger, and depression.

High levels of S-Curiosity reflect an intense desire to seek out, explore, and understand new things in the environment. The STCI T-Curiosity scale assesses individual differences in the disposition to experience S-Curiosity when responding to novel or ambiguous stimuli. Persons high in T-Curiosity feel curious more frequently and with higher levels of intensity than those who are low in T-Curiosity. The results of factor analyses of the STCI S-Curiosity and T-Curiosity items confirmed that these scales assessed independent but substantially correlated factors (Spielberger & Starr, 1994).

#### **Use in Clinical Practice: Assessing Psychological Vital Signs**

Emotions motivate behavior and have a significant impact on health and psychological well-being. It is therefore essential to evaluate and monitor emotional states in diagnosis and treatment, just as physicians in medical examinations routinely measure pulse rate, blood pressure, and temperature, the vital signs that provide essential information about physical health (Spielberger, Ritterband, Sydeman, Reheiser, & Unger, 1995). When a physician detects an abnormal pulse during a physical examination, this signals a potentially significant problem in the functioning of the cardiovascular system.

Running a high fever may indicate that the immune system is not protecting the person from harmful viruses. Intense anxiety and anger are analogous to elevations in pulse rate and blood pressure, while the presence of a fever, as indicated by abnormally high body temperature, may be considered as roughly analogous to depression. Elevations in temperature that define a fever are interpreted by physicians as a strong indication of the presence of an infection or metabolic problem that requires immediate attention (Guyton, 1977). Similarly, symptoms of depression often indicate the presence of pervasive unresolved conflicts that result in an emotional fever. While fevers can usually be reduced in patients with colds or the flu by aspirin or acetaminophen, the emotional conflicts that contribute to clinical depression are

not as readily relieved by either drugs or brief behavioral interventions.

Anxiety, anger, and depression are the emotional vital signs that are most critical to an individual's psychological well-being. Variations in the intensity and duration of these emotional states provide essential information about a person's mental and physical health, and can help to identify recent events that have particular meaning and impact on an individual's life, as well as long-standing conflicts. Assessing emotional vital signs and providing timely and meaningful feedback during treatment will enhance a patient's awareness and understanding of his or her feelings. Helping patients to cope more effectively with these feelings early in treatment will also minimize dropouts.

The clinical assessment of emotional vital signs can provide essential information for diagnosis and treatment planning, and for monitoring the treatment process. Since management of anxiety, anger, and depression are major concerns of most counselors and psychotherapists, the continuous assessment of these emotions can facilitate the treatment process (Deffenbacher, Demm, & Brandon, 1986; Novaco, 1979). Dealing with intense feelings of anxiety (SAnxiety), according to de la Torre (1979), should be a major priority in all forms of psychotherapy, especially in crisis interventions that focus on the specific problems of a patient or client. Barlow (1985) has consistently emphasized the importance of utilizing measures that differentiate between anxiety and depression during the course of treatment. Although less attention has been given to the assessment of anger, Deffenbacher's (1992) research demonstrates that anger can be readily measured during treatment, and that it is important to do so.

What does curiosity add to the assessment of emotional vital signs? As a motivator of exploratory behavior, curiosity may be considered as a positive vital sign that contributes to personal adjustment and successful adaptation to environmental stimuli. Research has shown that anxiety, anger, and depression are inversely related to curiosity, providing evidence that negative emotions, especially when relatively intense, may interfere with and inhibit the positive and adaptive effects of curiosity. Thus, curiosity may be considered as a positive vital sign and an important indicator of an individual's psychological well-being. Ideally, the assessment of emotional vital signs should begin immediately prior to the initial treatment interview so that feedback and crisis-oriented intervention can be given immediately if needed. The STPI can be rapidly and easily administered and scored, either by computer or manually, to assess anxiety, anger, depression, and curiosity, while the patient is waiting to be seen. If a patient is depressed or experiencing intense anxiety or anger, it is imperative for the therapist to deal immediately and directly with these feelings, which can greatly interfere with judgment and reality testing and result in injuries to the patient or other persons. Feedback concerning emotional vital signs can also help patients to recognize and report relationships between their thoughts and feelings and the events that give rise to them, thus facilitating

the therapeutic process.

Charting emotional vital signs over the course of treatment and providing patients with feedback about their feelings can help to identify significant problem areas, and thereby facilitate a better understanding of how specific problems influence the patients' emotions. The continual assessment of anxiety, anger, depression, and curiosity as emotional vital signs can also provide information regarding the effectiveness of the treatment process. Periodically obtaining measures of anxiety, anger, depression, and curiosity as personality traits will also provide evidence of treatment effectiveness. A decrease in anxiety, anger, and depression as personality traits should follow periods in which intense levels of these emotional states are experienced. Enhanced curiosity will generally be associated with improved psychological wellbeing.

### **SUMMARY AND CONCLUSIONS**

The historical background of theory and research on anxiety, anger, depression, and curiosity were briefly reviewed, and the nature of these emotional states and personality traits were examined. Darwin and Freud regarded fear (anxiety), rage (anger), and aggression as universal characteristics of human experience, and noted that the interaction of anxiety and anger contributed to depression. Both Freud and William James recognized the unique and positive role of curiosity in stimulating exploratory behavior that was essential to survival, and observed that curiosity was often inhibited by anxiety.

The construction and development of the State-Trait Anxiety Inventory, the State-Trait Anger Expression Inventory, and the State-Trait Personality Inventory to assess anxiety, anger, depression, and curiosity were described in detail in this chapter. Measuring these psychological vital signs is of critical importance in the diagnosis and treatment of patients with emotional problems, and can provide timely feedback that contributes to more effective psychological interventions. Assessing a patient's emotional vital signs can also facilitate treatment by linking intense emotional feelings to the events and the experiences that give rise to them. Anxiety, anger, depression, and curiosity, as indicators of psychological distress and well-being, should be carefully assessed in diagnostic evaluations, continuously monitored in counseling and psychotherapy, and evaluated as outcome measures in treatment interventions.

### **REFERENCES**

- Alexander, F.G., & French, T.M. (Eds.). (1948). *Studies in psychosomatic medicine: An approach to the cause and treatment of vegetative disturbances*. New York: Ronald.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Aronoff, J. (1962). Freud's conception of the origin of curiosity. *Journal of Psychology*, *54*, 39–45.
- Auerbach, S.M. (1973). Trait-state anxiety and adjustment to surgery. *Journal of Consulting and Clinical Psychology*, *40*, 264–271.
- Barlow, D.H. (1985). The dimensions of anxiety disorders. In A.H.

- Tuma and J.D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 479–500). Hillsdale, NJ: Erlbaum.
- Beck, A.T., & Steer, R.A. (1987). *Beck Depression Inventory: Manual*. New York: Psychological Corporation.
- Beck, A.T., Steer, R.A., & Brown, G.K. (1996). *Beck Depression Inventory: Manual BDI-II*. New York: Psychological Corporation.
- Biaggio, M.K., Supplee, K., & Curtis, N. (1981). Reliability and validity of four anger scales. *Journal of Personality Assessment*, *45*, 639–648.
- Booth-Kewley, S., & Friedman, H.S. (1987). Psychological predictors of heart disease: A quantitative review. *Psychological Bulletin*, *101*, 343–362.
- Brooks, M.L., Walfish, S., Stenmark, D.E., & Canger, J.M. (1981). Personality variables in alcohol abuse in college students. *Journal of Drug Education*, *11*, 185–189.
- Buss, A.H. (1961). *The psychology of aggression*. New York: Wiley.
- Buss, A.H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, *21*, 343–349.
- Cattell, R.B. (1966). Patterns of change: Measurement in relation to state-dimension, trait change, lability, and process concepts. *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally.
- Cattell, R.B., & Scheier, I.H. (1958). The nature of anxiety: A review of thirteen multivariate analyses comprising 814 variables. *Psychological Reports*, *4*, 351.
- Cattell, R.B., & Scheier, I.H. (1963). *Handbook for the IPAT Anxiety Scale* (2nd ed.). Champaign, IL: Institute for Personality and Ability Testing.
- Cofer, C.N., & Appley, M.H. (1964). *Motivation: Theory and research*. New York: Wiley.
- Cook, W.W., & Medley, D.M. (1954). Proposed hostility and pharisaic-virtue scales for the MMPI. *Journal of Applied Psychology*, *38*, 414–418.
- Crane, R.S. (1981). The role of anger, hostility, and aggression in essential hypertension (Doctoral dissertation, University of South Florida, Tampa, 1981). *Dissertation Abstracts International*, *42*, 2982B.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–335.
- Curtis, G., Kinder, B., Kalichman, S., & Spana, R. (1988). Affective differences among subgroups of chronic pain patients. *Anxiety Research: An International Journal*, *1*, 65–73.
- Darwin, C. (1965). *The expression of emotions in man and animals*. Chicago: University of Chicago Press. (Original work published 1872.)

## 17 Reynolds

## The Behavior Assessment System for Children

JENNIFER THORPE RANDY W. KAMPHAUS CECIL R. REYNOLDS

## V. OBJECTIVE METHODS

Bedecked in a long flowered dress, Anna came skipping through the clinic door, smiling from ear to ear, with her pigtails flopping side to side. She was full of energy, had trouble taking her gaze off her own reflection in the observation mirror, and was immediately conversant with the examiner. During the parent interview, her mother explained that Anna had difficulty paying attention in school, daydreamed, and sometimes did not complete schoolwork. Despite being generally a happy child, Anna was easily upset and this tendency sometimes interfered with cooperative play with friends. Her mother said she was worried that Anna's standardized reading scores were well below that of her older sister at her age and was concerned that she might suffer from attention-deficit/hyperactivity disorder (ADHD). Anna's older sister had always been one of the top students in her class.

Her mother reported that it was difficult to avoid comparisons of her daughters, which she said were often "inevitable."

The results of Anna's evaluation revealed scores in the average range on the Wechsler Intelligence Scale for Children—Third Edition (WISC-III Verbal IQ = 97; WJSC-III Performance IQ = 91) with achievement commensurate with these estimates of cognitive ability (WJ-R Reading Composite = 99; WJ-R Math Composite = 93). Anna's mother and father reported some attentional difficulties on the Behavior Assessment System for Children that fell in the borderline range (mother's BASCPRS *T*-score = 63; father's BASC-PRS *T*-score = 61). However, when her mother was queried using the ADHD module of the Structured Interview for the Diagnostic Assessment for Children, Anna's symptoms did not meet criteria for diagnosis under the fourth edition of *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV; American Psychiatric Association, 1994). In addition, although her teacher reported some daydreaming in class during the interview, she did not report any elevations in Attentional Problems on the BASC (TRS *T*-score = 56), instead endorsing items on the Anxiety scale approaching clinical significance (TRS *T*-score = 69). No elevations were apparent on the Achenbach Teacher Report Form, which combines anxious and depressed symptoms in a single Anxious/Depressed scale (*T*-score = 50). Although still in the second grade, Anna had perfect grades in

school and exhibited no behavioral problems at home or in school. Her teacher also reported that Anna had good social skills relative to her peers (TRS Social Skills *T*-score = 61) and average scores on other Adaptive composite scales (BASC TRS *T*-scores: Leadership = 57; Adaptability = 51; Study Skills = 59). Such behavioral strengths are unexpected in children with ADHD. On the Parenting Stress Index Anna's mother reported elevated levels of stress associated with Anna's moods, distractibility, and acceptability.

Anna did not meet criteria for ADHD diagnosis under DSM-IV; however, her mother had valid concerns regarding her ability to remain focused on academic tasks and to follow through on chores at home. The evaluation results revealed subclinical elevations in attention problems and possibly emergent anxiety that warranted continued monitoring, reevaluation within a year, and low-level behavioral interventions but did cross the categorical threshold for ADHD diagnosis. The clinician recommended that Anna be seen again in a year to evaluate her academic progress and short-term family therapy to address behavioral problems and expectations regarding Anna's performance relative to that of her sister.

#### **DIMENSIONAL ASSESSMENT WITH THE BEHAVIOR ASSESSMENT SYSTEM FOR CHILDREN**

Every clinician has experienced the quandary of children such as Anna who present with problem symptoms that nevertheless do not warrant DSM diagnosis: a child who cannot sit still but manages to make average grades in school, a child who seems to cry at the least provocation but does not meet criteria for mood disorder, a child who appears withdrawn but does not exhibit other identifiable difficulties. Behavior rating scales, such as those included in the Behavior Assessment System for Children (BASC), provide dimensional information on whether this child lies at the low or high range of the behavioral continuum. Because many or most highly prevalent disorders of childhood represent extremes of a continuum (Fergusson & Horwood, 1995; Scahill et al., 1999) (e.g., some children are more or less anxious, sad, social, inattentive, or active than others), diagnosis and treatment are highly dependent on determining that child's location on the continuum relative to other children at his or her developmental level. Dimensional approaches ensure that children with subthreshold impairments can

be monitored. Thus, elevations or a pattern of subclinical scale elevations can be a red flag to the clinician that the child may not meet categorical criteria but is still experiencing impairment in particular domains that require tracking over time (Cantwell, 1996).

## BRIEF HISTORY

The BASC represents a blend of traditional behavioral methods and contemporary cognitive-behavioral approaches. At one time, behavioral assessment only dealt with clearly observable, overt behavior. With the rise of popularity of cognitive behaviorism throughout the 1980s, reports of covert behavior such as thoughts, feelings, and desires have come to be included as a significant component of behavioral assessment (e.g., Kratochwill, Sheridan, Carlson, & Laseck, 1999). However, in the use of selfreports of internal or otherwise unobservable phenomena, behavioral assessment does not draw deep-seated psychodynamic inferences nor does it seek determinants of character; the responses are viewed directly for what they represent—samples of behavior and reports of the frequency or occurrence of specific behaviors. In contrast to early conceptualizations of the behavioral assessment paradigm, however, clinicians do now recognize the appearance of chronic, long-standing characteristics such as anxiety and locus of control that generalize beyond highly specific settings. In fact, many traditional measures in use by psychologists are now used as components of behavioral assessments but with lower levels of inference involved in the interpretation of the results. (See Ramsay, Reynolds, & Kamphaus, 2002, for a review of the many different methods of behavioral assessment.) This chapter discusses the BASC, which offers practitioners a practical tool kit with complementary scales and interviews for assessing both positive and negative behaviors across home and school settings, using different informants and methods for measuring behavior. Interpretation and synthesis of multisource information is simplified through co-normed scales consistent across age levels and teacher and parent forms. Thus, use of the full Assessment System can capitalize on empirically supported strengths in raters, such as the general superiority of teachers to rate attention and hyperactivity (Loeber, Green, & Lahey, 1990;

Verhulst, Koot, & Van der Ende, 1994); of children to report on their own internal moods such as depression or anxiety (Bird, Gould, & Staghezza, 1992), and of parents to describe their child's behavior specifically and differentially (Daniel, 1993). The provision of comprehensive, unique information using a variety of measurement techniques and sources is a strength of the system (Adams & Drabman, 1994). The BASC also offers a structured developmental history form to place behavior in context and an efficient Systematic Observation System (SOS) to allow for the real-time coding of classroom behaviors by trained observers.

### **SYSTEMATIC MEASUREMENT OF BEHAVIORAL STRENGTHS AND WEAKNESSES**

The BASC represents a departure from many typical rating scales in that it provides normatively referenced information on the child's adaptive behaviors or strengths as reported by parents, teachers, and the children themselves. The lack of positive behavioral dimensions in the past has been a key limitation of behavior rating scales (e.g., see Kratochwill et al., 1999). Not only are these behaviors indicators of the child's functional status, many states mandate that they be included in a diagnostic assessment of learning disabilities or emotional and behavioral disorders. Without the ability to reference the child to his or her peers on these dimensions, the practitioner must rely solely on subjective narrative reports of caregivers and teachers or his or her own one-time impression of the child in the clinic. In addition, the inclusion of positive items counterbalances negatively worded items, mitigates against response sets (Kamphaus & Frick, in press), and provides the child's caregivers and teachers the opportunity to report favorably about the child. Thorpe, Kamphaus, Rowe, and Fleckenstein (2000) found that scores on the Adaptive Composite scale of the BASC, which includes adaptability (response to change in environment), study skills, social skills, and leadership subscales, predicted children's behavioral and academic status in school as many as 2 years later. The Adaptive composite was as good or better than the Externalizing and Internalizing Composite scores in predicting children's school adjustment, adding complimentary predictions above and beyond these traditional indicators. These findings and others (e.g., diSibio, 1993) strongly support the standardized measurement of adaptive behaviors

in making predictions regarding future prognosis (Thorpe et al., 2000). In Anna's case her mother and teacher's report of good adaptive skills on the BASC Parent Rating Scale (BASC-PRS) and the BASC Teacher Rating Scale (BASC-TRS) suggested little functional impairment in these domains, which is vital information in making diagnostic determinations.

#### **EFFECTIVE DIFFERENTIAL DIAGNOSTIC TOOL**

The range of behavioral dimensions assessed by the BASC aids in making a differential diagnosis of specific categories of a disorder as denoted in the DSM-IV (American Psychiatric Association, 1994). The BASC rating scales are constructed to provide separate information on Attention Problems and Hyperactivity as well as separate information on Depression and Anxiety. This delineation allows for a better differential diagnosis and is particularly useful in making decisions regarding inattentive, hyperactive, or combined subtypes of ADHD that have very different implications for treatment (Doyle, Ostrander, Skare, Crosby, & August, 1997; Vaughn, Riccio, Hynd, & Hall, 1997). In Anna's case, her teacher's report of elevations on anxiety reached significance on the BASC but not on the Achenbach Child Behavior Checklist (CBCL) Anxious/Depressed subscales, perhaps due to the blending of these two constructs. In addition, although her mother reported elevations only on the Attention subscale of the BASC, she did not report problems on the Hyperactivity subscale (*T*score = 47), suggesting that Anna's difficulties were restricted to attentional lapses. With the exception of item-level analysis, no differentiation of the two behaviors was possible using Achenbach CBCL, which provided a single Attention Problems subscale (*T*-score = 68) and also includes items relating to impulsivity. Research supports the advantage of the BASC subscales in accurately discriminating children with primary inattentive type from combined type (Vaughn et al., 1997).

#### **COMPLIANCE WITH FEDERAL/STATE EDUCATIONAL ASSESSMENT REQUIREMENTS**

The BASC offers clinicians working in school settings or treating children with learning problems or behavioral problems an array of important measurement and assessment

techniques to target these problems and to comply with standards for behavioral analysis necessary for tailoring individualized education plans (IEPs) to students with special needs as required under the Individuals with Disabilities Education Act (IDEA) (Flanagan, 1995). A unique School Problems composite on the BASCTRS provides clinicians with well-normed information regarding that child's adjustment in a school setting. The BASC SOS facilitates effective functional behavioral analysis, also required in recent revisions to IDEA, as well as an easy-to-use method for tracking changes in frequency and duration of such behaviors over time. The BASC ADHD Monitor provides a shortened format for monitoring changes in Attention and Hyperactivity subsequent to medication and or targeted behavioral interventions. These components are described more fully later in the chapter. Components of the BASC also assess aspects of the federal definition of Emotional Disturbance (Flanagan, 1995; Reynolds & Kamphaus, 1992). In their review, Sandoval and Echandia (1994) call the BASC "one of the most useful and sophisticated of all the new measures available to those wishing to assess school-age children" (p. 425).

Referring back to Anna's case, the BASCTRS was used to reevaluate her school behavior 3 months later. The results, presented here, revealed both consistency and stability on the majority of scales as well as sensitivity to change. While Anna's scores on Attention Problems scaled were identical to the earlier evaluation and her elevations on the Anxiety subscale are still apparent, her teacher's report at follow-up reveals a sharp increase in somatic complaints, suggesting that Anna's school performance concerns might be manifesting as physical ailments. Scores in bold are composites that represent an overall index of their component subscales.

Anna's Case

BASC-TRS *T*-scores

Oct. 1999 Jan. 2000

Hyperactivity 44 44

Aggression 41 43

Conduct Problems 43 43

**Externalizing Problems 42 43**

**composite**

Anxiety 69 62

Depression 41 46

Somatization 46 64

## **Internalizing Problems 52 59**

### **composite**

Attention Problems 56 56

Learning Problems 61 53

## **School Problems 59 55**

### **composite**

Atypicality 52 57

Withdrawal 51 48

## **Behavior Symptoms 51 52**

### **Index**

## **COMPONENTS AND USES**

The BASC is a multimethod, multidimensional approach to evaluating the behavior and self-perceptions of children ages 2½–18 years, and includes its new variant, the BASC ADHD Monitor (Kamphaus & Reynolds, 1998). The original BASC is *multimethod* in that it has five components, which may be used individually or in any combination:

\_ The *Teacher and Parent Rating Scales (TRS and PRS)*, which are separate instruments that gather descriptions of the child's observable behavior at home, in the community, and at school.

\_ The *Self-Report of Personality (SRP)*, which the child uses to describe his or her behaviors, emotions and self-perceptions.

\_ The *Structured Developmental History (SDH)*, which is used to collect biographical, demographic, historical, and developmental information from parents or other primary caregivers, and which can serve as the basis for a parent interview.

### **390 V. OBJECTIVE METHODS**

\_ The *Student Observation System (SOS)*, a form for recording and classifying directly observed classroom behavior.

The BASC components not only provide different sources of information but, in fact, use different methods, a factor important to assessing generalizability of results and validation in diagnosis. The SRP, which can be used with children from 8 to 18 years of age, provides an introspective report of behavior, feelings, attitudes, and cognitions.

The BASC-TRS and -PRS provide holistic summaries of that child's "typical" behavior on an array of positive and negative indicators as seen through the eyes of behavioral experts specific to the child in question. The SOS provides direct observation and counting of behavior, believed by many to be the sine qua non of behavioral assessment (e.g., see Ramsay, Reynolds, & Kamphaus, in press, Ch. 1, for a review).

The SDH helps provide a context for the presenting problem and provides for a structured interview as an additional method of assessment.

The BASC is *multidimensional* in that it measures numerous dimensions of behavior and personality, including positive (adaptive) as well as negative (clinical) dimensions and both internalizing and externalizing problems. The BASC Adaptive scales include Social skills, Study skills, Leadership, and Adaptability from the standpoint of the child's parent or teacher and Interpersonal relations, Self-esteem, Self-reliance, and parental relations from the point of view of the child. The BASC assesses both overt and covert behavior along with attitudes, feelings, and cognitions as well as certain affective states (e.g., anxiety, depressed mood, and attributional states), giving a range of dimensions heretofore unavailable in a single system.

Scales were developed based on comprehensive theoretical and empirical considerations (Reynolds & Kamphaus, 1992) and represent a synthesis of what is known about developmental psychopathology (Sandoval & Echandia, 1994). In their review of technical qualities of the BASC, Sandoval and Echandia (1994) remark that "authors have set the standard for test construction for this kind of scale to be used with the childhood population" (p. 421).

### Teacher Rating Scales

The TRS has three forms with items designed for three age levels: preschool (2½–5), child (6–11), and adolescent (12–18). The forms contain descriptors of behaviors that the respondent rates on a 4-point scale of frequency, ranging from *Never* to *Almost always*.

The TRS takes 10–20 minutes to complete, although teachers who are familiar with the TRS seldom require more than 10 minutes.

Evidence for temporal stability and convergent validity (Merydith & Joyce, 1998) of the TRS has been presented.

The TRS assesses clinical problems in the broad domains of Externalizing Problems, Internalizing Problems, and School Problems.

It also measures Adaptive Skills. Table 17.1 shows the scales for all levels of the TRS. The slight differences between levels are due to developmental changes in the behavioral manifestations of child problems.

Nevertheless, scales and composites with

the same name contain essentially the same conceptual content at all age levels, even though specific items change across age. Children simply do not show their problems the same way at all developmental levels. In addition to scale and composite scores, the TRS provides a broad composite, the Behavioral Symptoms Index (BSI) that assesses the overall level of problem behaviors. While we recommend that a teacher know a child at least 4 to 6 weeks before using the TRS, a recent study found little difference between ratings of a new teacher and the previous year's teacher who had known the child for most of a school year. Hoover, Braver, Wolchik, and Sandler (2000) found that teacher ratings on the Teacher-Child Rating Scale (T-CRS; Hightower, 1986) were similar for a group of 240 elementary grade children who were part of a divorce intervention study. They concluded that neither the previous teachers' nor the current teachers' fall ratings were significantly different from the spring teachers' pretest ratings. Thus, school psychologists may elect to obtain ratings from either the previous or current teacher early in the fall of a new academic year. The TRS may be interpreted with reference to national age norms (General, Female, or Male) or to Clinical norms. In addition, selected critical items may be interpreted individually. The TRS includes a validity check in the form of an *F* ("fake bad") index designed to detect an excessively negative response set on the part of the teacher completing the rating. The BASC software programs also yield a Patterning validity index that assesses deviant patterns such as alternating between choices on a consistent basis. This validity index is not typically in question because teachers and parents have little incentive to complete a rating scale carelessly. The consistency index produced by the software is of greater value in that it detects agreement among highly similar items. By doing so this index assesses more subtle response bias or may detect unreliability of a specific rater.

### **Parent Rating Scales**

The PRS is a comprehensive measure of a child's adaptive and problem behaviors in community and home settings. The PRS uses the same four-choice response format as the TRS and also takes 10–20 minutes to

complete. Like the TRS, the PRS has three forms at three age levels: preschool, child, and adolescent. The age levels of the PRS are similar in content and structure. Table 17.2 shows the scale definitions of the PRS. The PRS assesses almost all the clinical problem and adaptive behavior domains that the TRS measures. However, the PRS does not have a School Problems composite, nor does it include the two TRS scales that are best observed by teachers (Learning Problems and Study Skills). The PRS offers the same norm groups as the TRS: national age norms (General, Female, and Male) and Clinical norms. Like the TRS, the PRS includes an *F* index, patterning, and consistency indexes as checks on the validity of the parent ratings and critical items that may signify behaviors that should be interpreted individually.

**Self-Report of Personality**

The SRP is an omnibus behavioral and personality inventory specially designed for children and adolescents to report an array

392 V. OBJECTIVE METHODS

**TABLE 17.1. Composites and Scales in the TRS and PRS**

Teacher Rating Scales Parent Rating Scales

Composite/Scale	Preschool	Child	Adolescent	Preschool	Child	Adolescent
<b>Externalizing Problems</b>	*	*	*	*	*	*
<i>Aggression</i>	*	*	*	*	*	*
<i>Hyperactivity</i>	*	*	*	*	*	*
Conduct Problems	*	*	*	*	*	*
<b>Internalizing Problems</b>	*	*	*	*	*	*
<i>Anxiety</i>	*	*	*	*	*	*
<i>Depression</i>	*	*	*	*	*	*
Somatization	*	*	*	*	*	*
<b>School Problems</b>	*	*				
<i>Attention Problems</i>	*	*	*	*	*	*
Learning Problems	*	*				
<b>Other Problems</b>						
<i>Atypicality</i>	*	*	*	*	*	*
<i>Withdrawal</i>	*	*	*	*	*	*
<b>Adaptive Skills</b>	*	*	*	*	*	*
Adaptability	*	*	*	*	*	*
Leadership	*	*	*	*	*	*
Social Skills	*	*	*	*	*	*
Study Skills	*	*				
<b>Behavioral Symptoms Index</b>	*	*	*	*	*	*

*Note.* Italicized scales compose the Behavioral Symptoms Index. From Reynolds and Kamphaus (1992). Copyright 1992 by American Guidance Service, Inc. Reprinted by permission.

of thoughts, feelings, and behaviors relevant to psychological and interpersonal adjustment. The SRP, which takes about 20–30 minutes to complete, consists of a list of True/False statements to be completed by the child or adolescent. The two forms, one for children (ages 8–11) and one for adolescents (ages 12–18), have considerable overlap

in scales, in structure, and in individual items. Both levels have identical composite scores: School Maladjustment, Clinical Maladjustment, Personal Adjustment, and an overall composite score, the Emotional Symptoms Index (ESI). The child level (SRPC) has 12 scales and the adolescent level (SRP-A) has 14 scales arranged into composites (see Table 17.3). Unlike the BSI for the rating scales, the ESI is composed of both negative (clinical) scales and positive (adaptive) scales whose scoring has been reversed, because these are the scales that load highest on a general psychopathology factor. Like the rating scales, the SRP may be interpreted with reference to national age norms (General, Female, and Male) or to Clinical norms. Special indexes are incorporated to assess the validity of the child's responses: the *F* index, patterning index, consistency index, the *L* ("fake good") index for the SRP-A only, and the *V* index designed to detect invalid responses due

#### 17. The Behavior Assessment System for Children 393

**TABLE 17.2. BASC TRS and PRS Definitions**

Scale Definition

Adaptability The ability to adapt readily to changes in the environment

Aggression The tendency to act in a hostile manner (either verbal or physical) that is threatening to others

Anxiety The tendency to be nervous, fearful, or worried about real or imagined problems

Attention Problems The tendency to be easily distracted and unable to concentrate more than momentarily

Atypicality The tendency to behave in ways that are immature, considered "odd," or commonly associated with psychosis (such as experiencing visual or auditory hallucinations)

Conduct Problems The tendency to engage in antisocial and rule-breaking behavior, including destroying property

Depression Feelings of unhappiness, sadness, and stress that may result in an inability to carry out everyday activities (neurovegetative symptoms) or may bring on thoughts of suicide

Hyperactivity The tendency to be overly active, rush through work or activities, and act without thinking

Leadership The skills associated with accomplishing academic, social, or community goals, including, in particular, the ability to work well with others

Learning Problems The presence of academic difficulties, particularly in understanding or completing schoolwork

Social Skills The skills necessary for interacting successfully with peers and adults in home, school, and community settings

Somatization The tendency to be overly sensitive to and complain about relatively minor physical problems and discomforts

Study Skills The skills that are conducive to strong academic performance, including organizational skills and good study habits

Withdrawal The tendency to evade others to avoid social contact

*Note.* The PRS does not include TRS composite scales of Learning Problems, Study Skills, or School problems. From Reynolds and Kamphaus (1992). Copyright 1992 by American Guidance Service, Inc. Reprinted by permission.

to poor reading comprehension, failure to follow directions, refusal to respond seriously to the task, or poor contact with reality.

Table 17.4 lists scale definitions of the SRP.

### **Structured Developmental History**

The SDH is an extensive history and background survey that may be completed by a clinician during an interview with a parent or guardian, or may be completed as a questionnaire by a parent, either at home or in the school or clinic.

The SDH systematically gathers information that is crucial to the diagnostic and treatment process. Many developmental events and medical or related problems in the family may have an impact on a child's current behavior. The SDH structures the gathering of the child and family history, both social and medical. Because it is comprehensive, the SDH should be an asset to any evaluation of a child, whether or not other BASC components are used. Areas addressed in the SDH are noted in Table 17.5.

### **Student Observation System**

The SOS is a form for recording a direct observation of the classroom behavior of a child. The SOS uses the technique of momentary time sampling (i.e., systematic coding during 3-second intervals spaced 30 seconds apart over a 15-minute period) to record a wide range of children's behaviors, including positive behaviors (such as teacher–student interaction) as well as negative behaviors (such as inappropriate movement or inattention).

The BASC SOS may be used appropriately in regular and special education classes. It can be used in the initial assessment as part of the diagnostic process. It can also be used repetitively to evaluate the effectiveness of educational, behavioral, psychopharmacological, or other treatments.

### **Forms**

The TRS, PRS, and SRP forms come in two formats: handscoring or computer entry.

The handscoring forms are printed in a convenient self-scoring format, allowing them to be scored rapidly without using templates or keys (requiring about 5 minutes each to score after practice with the forms). Each form includes a profile of scale and composite scores. The computer entry forms, which are simpler one-part forms, are designed to allow the user to key item responses into a microcomputer in about 5 minutes.

### **Computer Software**

A microcomputer program, BASC Plus, is available that offers on-line administration of the TRS, PRS, and SRP and computer scoring of a completed computer-scored or handscored form. The manual for BASC Plus explains how to use the program to administer, score, and report the TRS, PRS, and SRP. It includes additional interpretive text and a listing of target behavior not available on other computer programs. The BASC Enhanced ASSIST program offers users a simpler computer program that produces all possible scores, a graphical display

#### 394 V. OBJECTIVE METHODS

#### **TABLE 17.3. Composites and Scales in the SRP**

Composite/Scale Child Adolescent

#### **Clinical Maladjustment \* \***

*Anxiety* \* \*

*Atypicality* \* \*

Locus of Control \* \*

*Social Stress* \* \*

Somatization \* \*

#### **School Maladjustment \* \***

Attitude to School \* \*

Attitude to Teachers \* \*

Sensation Seeking \*

#### **Other Problems**

*Depression* \* \*

*Sense of Inadequacy* \* \*

#### **Personal Adjustment \* \***

Relations with Parents \* \*

*Interpersonal Relations* \* \*

*Self-Esteem* \* \*

Self-Reliance \* \*

#### **Emotional Symptoms Index \* \***

*Note.* Italicized scales compose the Emotional Symptoms Index. From Reynolds and Kamphaus (1992).

Copyright 1992 by American Guidance Service, Inc.

Reprinted by permission.

of results, and item responses, but does not allow on-line administration.

#### **General Norms**

The General norms are based on a large national sample that is representative of the general population of U.S. children with regard to sex, race/ethnicity, clinical or special education classification, and, for the PRS, parent education. These norms are subdivided by age and, therefore, indicate how the child compares with the general population of children that age. For many applications, these norms (combining females and males) will be the preferred norms, and they are recommended for general use.

Several of the scales of the TRS, PRS, and SRP show gender differences. Males tend to obtain higher raw scores on the Aggression, Conduct Problems, Hyperactivity, Attention

**TABLE 17.4. SRP Scale Definitions**

Scale Definition

Anxiety Feelings of nervousness, worry, and fear; the tendency to be overwhelmed by problems

Attitude to School Feelings of alienation, hostility, and dissatisfaction regarding school

Attitude to Teachers Feelings of resentment and dislike of teachers; beliefs that teachers are unfair, uncaring, or overly demanding

Atypicality The tendency toward gross mood swings, bizarre thoughts, subjective experiences, or obsessive-compulsive thoughts and behaviors often considered "odd"

Depression Feelings of unhappiness, sadness, and dejection; a belief that nothing goes right

Interpersonal Relations The perception of having good social relationships and friendships with peers

Locus of Control The belief that rewards and punishments are controlled by external events or other people

Relations with Parents A positive regard toward parents and a feeling of being esteemed by them

Self-Esteem Feelings of self-esteem, self-respect, and self-acceptance

Self-Reliance Confidence in one's ability to solve problems; a belief in one's personal dependability and decisiveness

Sensation Seeking The tendency to take risks, to like noise, and to seek excitement

Sense of Inadequacy Perceptions of being unsuccessful in school, unable to achieve one's goals, and generally inadequate

Social Stress Feelings of stress and tension in personal relationships; a feeling of being excluded from social activities

Somatization The tendency to be overly sensitive to, experience, or complain about relatively minor physical problems and discomforts

*Note.* From Reynolds and Kamphaus (1992). Copyright 1992 by American Guidance Service, Inc. Reprinted by permission.

**TABLE 17.5. Areas Addressed in the SDH**

1. Person Answering Questions

2. Referral Information

3. Parents

4. Primary Caregivers

5. Child Care

6. Family History

7. Brothers/Sisters

8. Child's Residence

9. Family Relations

10. Pregnancy

11. Birth

12. Development

13. Medical History

14. Family Health

15. Friendships

16. Recreation/Interests

17. Behavior/Temperament

18. Adaptive Skills

19. Educational History

20. Additional Comments

Problems, and Learning Problems scales of the TRS and PRS and on the Sensation Seeking, Attitude to School, Attitude to Teachers, and Self-Esteem scales of the SRP.

Females tend to score higher than males on the Social Skills, Study Skills, Leadership, and Depression scales of the TRS and PRS and on the Anxiety and Interpersonal Relations scales of the SRP. These differences in scores likely reflect real differences between males and females in the incidence of the indicated behavioral or emotional problems

or strengths in adaptive skills. For these gender differences to be reflected in the normative scores, a common set of norms must be used for both males and females. The General combined-sex norms serve this purpose. General norms answer the question: How commonly does this level of rated or self-reported behavior occur in the general population at this age? Using General norms, more males than females will show high *T*-scores on Aggression, for example, and more females than males will have high *T*-scores on Social Skills. The combined gender or general norms preserve any observed gender difference in the shape and level of the raw score distributions. This is appropriate, and the general norms should be used if one believes that boys and girls are in fact different on various behavioral characteristics (i.e., observed differences are not due to psychometric artifacts). For example, girls score higher than boys on the SRP Anxiety scale (a common finding in research on anxiety, e.g., see Reynolds & Richmond, 1985). In determining which set of norms to use, the clinician must answer the question, "Are girls more anxious than boys, or are they simply more willing to admit to symptoms of anxiety?" If the former is true, the general norms are more appropriate, but in the latter case, the gender-specific norms are the correct choice. Reynolds and Kamphaus (1992) recommend the use of the general norms, a decision with which we continue to concur, but the individual clinician may disagree and opt for the other norms. This allows the clinician more latitude than typically occurs on other behavioral and self-report scales.

#### **Female Norms and Male Norms**

These norms are based on subsets of the General norm sample; each is representative of the general population of children of that age and gender. The effect of using these separate-sex norms is to eliminate differences between males and females in the distribution of *T*-scores or percentiles. For example, although raw score ratings on the Aggression scale tend to be higher for males than females, use of separate-sex norms removes this difference and produces distributions of normative scores that are the same for both genders.

#### **Indexes of Validity and Response Set**

Several indexes are provided to help the BASC user judge the quality of a completed form. Validity may be threatened by any of

several factors including failure to pay attention to item content, carelessness, an attempt to portray a child in a highly negative or positive light, lack of motivation to respond truthfully, or poor comprehension of the items. Information on the development of these indexes and the setting of cutoff scores is provided in Reynolds and Kamphaus (1992).

### F Index

The *F* index, included on all of the BASC rating-scale and self-report forms, is a measure of the respondent's tendency to be excessively negative about the child's behaviors or self-perceptions and emotions. The *F* index was developed using traditional psychometric methods associated with Infrequency scales (e.g., see Reynolds, 2001).

On the PRS and TRS, the *F* index is scored by counting the number of times the respondent answered *Almost always* to a description of negative behavior or *Never* to a description of positive behavior. Because responses on the SRP are limited to *True* and *False*, items selected for that *F* index are either extremely negative items to which the child responded *True* or positive items to which the response was *False*. Items were selected for these scales that have a low probability of co-occurrence (i.e., they are seldom endorsed in concert with one another). The TRS, PRS, and SRP record forms show what levels of *F* index scores are high enough to be of concern. Detailed guidance to interpretation of the *F* index is given in Reynolds and Kamphaus (1992).

## 396 V. OBJECTIVE METHODS

### L Index

The *L* index, offered for the adolescent level of the SRP, measures an adolescent's tendency to give an extremely positive picture of him- or herself—what might be called “faking good.” The index consists of items that are unrealistically positive statements (such as “I like everyone I meet”) or are mildly self-critical statements that most people would endorse (such as “I sometimes get mad”). Individuals scoring high on this scale may also be giving the most socially desirable response or possibly are psychologically naive relative to their peers denying common, everyday problems or concerns. The SRP-A record form shows which *L* scores should be of concern.

### V Index

Each level of the SRP includes a *V* index

made up of five or six nonsensical or highly implausible statements (such as “Superman is a real person”). The *V* index serves as a basic check on the validity of the SRP scores in general. If a respondent marks two or more of these statements as *True*, the SRP may be invalid.

### **BASC ADHD MONITOR**

The BASC ADHD Monitor fills a unique role in the assessment of children who are diagnosed with ADHD. The Monitor is the second step in an assessment regimen that is designed to enhance treatment planning and evaluation by thoroughly assessing the primary symptoms of ADHD on a continuing basis. Attention Problems and Hyperactivity constitute the core symptoms used by DSM-IV to define the ADHD syndrome (Kamphaus & Frick, in press). Problems in one or both of these areas are used to differentiate the three subtypes of ADHD: ADHD, predominantly inattentive type; ADHD, predominantly hyperactive-impulsive type; and ADHD, combined type.

Components of the original BASC system serve as the first step in the comprehensive assessment of children suspected of having ADHD. The BASC takes a broad sampling of child behavior to identify the full range of child problems, especially those that may mimic the symptoms of ADHD. If the initial administration of the BASC reveals problems on the Attention Problems and/or Hyperactivity scales, the diagnosis of ADHD becomes a possibility. Of greater importance, however, is the necessity to use the BASC Teacher, Parent, and Self-Report Forms to rule out co-occurring problems, which can only be done with the initial use of a broad-based measure such as the BASC PRS or TRS (Kamphaus & Frick, in press). This process of ruling out other problems is particularly important for the diagnosis of ADHD, where so many comorbid disorders occur and where other disorders (e.g., childhood depression) may superficially appear to be ADHD. In fact, the use of narrowband scales of inattention or hyperactivity may result often in overdiagnosis of ADHD. The Monitor is concerned with treatment design for and evaluation of children with ADHD. The narrowly focused Monitor is designed to assess an expanded range of Attention Problems and Hyperactivity symptoms in a time efficient and practical manner. This additional detail allows the

clinician to refine the diagnosis of ADHD and, of greater importance, to design a comprehensive treatment program aimed at reducing the core behavioral problems of inattention and hyperactivity. The Monitor also provides Internalizing and Adaptive Skills scales that further encourage comprehensive treatment planning and evaluation of treatment effectiveness by allowing clinicians to include these important constructs easily in the treatment plan.

The BASC SOS and BASC ADHD Monitor represent a coordinated multiple-step assessment system that allows the clinician to proceed from referral for ADHD to diagnosis, treatment design, and treatment evaluation with greater ease and precision. In order to achieve these assessment objectives, the Monitor utilizes information provided by parents, teachers, and a classroom observer to assess the constructs listed in Table 17.6.

Few tests are designed in a manner that facilitates the repeated collection and dissemination of child information to treatment providers (Kratochwill et al., 1999).

The Monitor is designed to meet the unusual practical demands dictated by the need for the repeated assessment of the core symptoms of ADHD. The original BASC is quite sensitive to behavioral changes in individual children and it may be used repeatedly to evaluate treatment effects (Conoley et al., 2001), particularly if a child is found to have multiple problems (e.g., ADHD, depression, anxiety, and conduct disorder) that cannot be fully assessed by the Monitor. In the case of ADHD and its subtypes, however, the Monitor is constructed to allow clinicians to evaluate treatment with greater focus and time efficiency. The BASC ADHD Monitor is designed to:

1. *Provide accurate and frequent feedback to the prescribing physician.* The physician and other health care workers need accurate information to ensure that a child is receiving the most accurate psychotropic regimen and to adjust dosage. Information about the effects of medication on Hyperactivity, Attention Problems, Internalizing Problems, and Adaptive Skills can aid the physician in making crucial medical treatment decisions.
2. *Ensure that the ongoing assessment of*

*ADHD problems is efficient, timely, and cost-effective.* Given the multiple time demands on parents, teachers, and others, little time remains to complete lengthy or unnecessarily complex rating scales that are not specifically targeted to the needs of the child with ADHD. On the other hand, the Monitor is designed to be adequately thorough in order to allow for the assessment of constructs in addition to the core dimensions of ADHD—Internalizing Problems and Adaptive Skills (Kamphaus & Frick, in press). All these assessment objectives must be achieved in an efficient way given the exigencies of health care. Accordingly, the Monitor is brief, yet it provides coverage of four important domains related to the functioning of the child with ADHD: Attention Problems, Hyperactivity, Internalizing Problems, and Adaptive Skills.

3. *Provide a system of devices that allows for input from multiple informants.* Teacher, parent, and clinician observations are all of potential importance for the treatment process, and communication among these individuals is crucial for effective treatment (Bender, 1997). Each Monitor form is designed to meet the specialized needs of each of these informants.

4. *Emphasize the assessment of specific behavioral outcomes in order to demonstrate accountability for services.* Increasingly, the effectiveness of child services is being challenged, thereby creating the need to assess outcomes. The Monitor assesses the DSM-IV criteria for ADHD and includes items that are written in clear behavioral terms. In addition, the Monitor software is designed to produce output that gives providers and administrators a clear indication of response to treatment. The Monitor is designed to provide clinicians with the information needed to adjust treatment whenever response to intervention is not optimal.

5. *Link assessment to treatment planning and evaluation.* The Monitor is designed to be practical enough to be considered central to the treatment process. Heretofore, physicians and other clinicians have often had difficulty acquiring the feedback needed to adjust treatment. The test and software design of the Monitor was guided throughout by the need to provide information relevant

## **BASC, ADHD Monitor, and Student Observation System**

### Component Scales

Parent Monitor Attention Problems

Hyperactivity

Internalizing Problems

Adaptive Skills

Teacher Monitor Attention Problems

Hyperactivity

Internalizing Problems

Adaptive Skills

BASC SOS Response to Teacher/Lesson

Peer Interaction

Work on School Subjects

Transition Movement

Inappropriate Movement

Inattention

Inappropriate Vocalization

Somatization

Repetitive Motor Movements

Aggression

Self-Injurious Behavior

Inappropriate Sexual

Behavior

Bowel/Bladder Problems

to treatment. The selection of items and scales, test length, scoring and reporting systems, graphical output, and other Monitor characteristics were all guided by this central objective.

ADHD Monitor interpretation can take several forms depending on the instrument(s) used, theoretical orientation of the clinician, the nature of the evaluation questions posed, and other factors. It is also important to keep in mind that the Monitor is designed to create and evaluate treatment plans. Therefore, interpretation of the scales as diagnostic devices is of considerably lesser importance.

The initial step in evaluating monitor results is that the individual clinician asks whether or not significant change has occurred in response to treatment. For the Parent and Teacher Monitors four questions are generally posed:

1. Is treatment affecting symptoms of inattention?
2. Is treatment affecting symptoms of hyperactivity?
3. Is treatment affecting internalizing symptoms?
4. Is treatment affecting adaptive skills?

The questions related to change are multitudinous and parallel for the SOS, where one may be assessing change at either the item or scale level (discussed in the last half of this chapter). Keep in mind that when scores change, they may show deterioration in some areas, not just improvement. For example, as a child's symptoms of overactivity

and inattention, come under control, comorbid symptoms of depression may become more prominent causing scores on the internalizing dimension to elevate.

Even if a significant change in *T*-scores is apparent there are additional questions to consider.

5. Which scales have changed?

6. Is there a temporal (and potentially causal) relationship between the delivery (or lack thereof) treatment and the observed change?

7. Is the change of “clinical significance”?

In other words, regardless of the amount of *T*-score change are parents or teachers reporting change that is adequate to reduce functional impairment in their eyes?

We think that establishing the temporal relationship between *T*-score change and delivery or withdrawal of treatments is of greatest importance. It is our experience that often this relationship is assessed via the speculation or supposition of the clinician.

We think that a better way to draw a conclusion regarding the relationship of treatment to behavioral change is to conduct repeat assessments until the relationship is clear. For example, one could see a reduction in attention problems subsequent to the first month’s administration of medication.

While this change represents a hopeful sign, this pretest–posttest design is probably insufficient to draw such a conclusion definitively. A third set of Monitor ratings taken a few months later that show the same trend would provide more assurance that the conclusion that medication is having an effect is warranted. A set of ratings yielding more symptoms when the child is not taking medication in the summer months or some other time would lend further support for medication effectiveness.

We often find that an additional brief assessment clarifies our conclusions to a much greater extent than prolonged theorizing based on more limited data.

Use of the BASC often precedes administration of the ADHD Monitor. There is, however, one important area of interpretive overlap between the BASC and the BASC ADHD Monitor Parent and Teacher Forms. A *T*-score baseline for treatment evaluation can be obtained from either set of measures. There are two administration scenarios that are most likely.

1. A clinician may administer either or

both of the BASC Parent and Teacher Forms during the initial diagnostic evaluation. The obtained *T*-scores for the Hyperactivity, Attention Problems, Internalizing Problems, and Adaptive Skills scales may be entered into the BASC ADHD software and be used as the baseline against which subsequent administrations of the ADHD Monitor will be compared.

2. A clinician may administer either or both of the BASC ADHD Parent and Teacher Monitor Forms during the initial diagnostic evaluation. The obtained *T*-scores for the Hyperactivity, Attention Problems, Internalizing Problems, and Adaptive Skills scales will then be used as the baseline against which subsequent administrations of the ADHD Monitor Forms will be compared.

It is important to establish a *T*-score baseline in a timely fashion regardless of the method used. In other words, we advise that a *T*-score baseline be collected during the evaluation phase and prior to implementation of treatment. The ADHD Monitor *T* scores for Parent and Teacher rating scales serve as the most reliable indicator of behavioral change over time (see Kamphaus & Reynolds, 1998).

## SOS

### Functional Behavioral Assessment with the SOS

The SOS addresses some of the shortcomings inherent in the use of classroom observation techniques. Specifically, the SOS was developed to make practical the use of a momentary time-sampling procedure that adequately samples the full range of a child's behavior in the classroom (Reynolds & Kamphaus, 1992). Several characteristics of the SOS exemplify this effort, including the following:

1. Both adaptive and maladaptive behaviors are observed (see Table 17.1);
2. Multiple methods are used including clinician rating, time sampling, and qualitative recording of classroom functional contingencies;
3. A generous time interval is allocated for recording the results of each time sampling interval (27 seconds);
4. Operational definitions of behaviors and time-sampling categories are included in the BASC manual (Reynolds & Kamphaus,

1992); and

5. Interrater reliabilities for the time-sampling portion are high which lends confidence that independent observers are likely to observe the same trends in child's classroom behavior (Lett & Kamphaus, 1997).

These characteristics of the SOS have contributed to its popularity as a functional behavioral assessment tool. It is crucial, for example, to have adequate operational definitions of behaviors that, in turn, contribute to good interrater reliability. Without such reliability, clinicians will never know whether their observations are unique and potentially influenced by their own biases or idiosyncratic definitions of behavior.

We also think that it is central for observations to simultaneously account for a child's adaptive skills in the classroom. It is only by doing so that a clinician is able to recommend behaviors that should be targeted for instruction, intervention, or strengthening. Specifically, the BASC SOS Parts A, B, and C, and other components, can contribute to the functional assessment of behavior from multiple perspectives:

\_ Behavior *Frequency*. SOS Part A ratings of "never observed," "sometimes observed," and "frequently observed." SOS

Part B assesses frequencies by category of behavior problem and PRS and TRS ratings tally the frequency of behavior problems.

\_ Behavior *Duration*. SOS Part B ratings of percentage of time engaged in a particular behavior by category.

\_ Behavior *Intensity*. SOS Part A ratings of "disruptive." SOS Part B ratings of frequency by category.

\_ *Antecedent events to Behavior*. SOS Part C descriptions of teacher position, behavior and other variables that precede misbehavior.

\_ *Consequences of Behavior*. SOS Part C descriptions of teacher behavior, peer behavior, and other variables that follow a behavior.

\_ *Analysis of Behavior across Settings*. SOS observations made at various times of day and classroom setting. The PRS may be used for the assessment of behavior in the community and home environments. Other components of the BASC, such as the PRS and TRS, may also be used as part of a functional behavioral assessment paradigm. Given the time-consuming nature of

observations, it may be more practical to collect teacher ratings from classrooms where an observation is not practical and parent ratings in order to assess differences across settings. Observations are central to the ongoing classroom problem-solving and consultation process that is frequently concerned with the ongoing assessment of a child's behavioral adaptation in school as is discussed next (additional functional behavioral assessment guidance may be obtained at [www.air.org/cecp/fbalproblembehavior/strategies.htm#direct](http://www.air.org/cecp/fbalproblembehavior/strategies.htm#direct)).

### **Monitoring with the SOS**

The SOS is the one component of the BASC ADHD Monitor that may be applied to all children regardless of their diagnosis or classification. In fact, we know of school districts that use the SOS and Monitor Software to evaluate progress toward TEP objectives, assess effects of prereferral intervention, and assess the effectiveness of various special education programming decisions. Some have used the SOS to assess the impact of social work or the services on classroom behavior. Perhaps more than any other BASC component, the SOS was specifically designed to serve the behavioral intervention and evaluation process in the classroom. We now discuss some possible scenarios and examples of applications of the SOS.

### **Medical Effects**

Mary's parents are opposed to the use of medication with their child in spite of the fact that numerous behavioral (psychotherapy, play therapy, token economy, etc.) and educational interventions (peer tutor, after school tutor, summer school, preferential seating, etc.) have failed. The SOS may help such reluctant parents gauge the effects of pharmacological interventions on Mary's classroom behavior in a manner that they perceive as more objective than teacher ratings. In this example an independent, perhaps even case-blind, observer may take SOS observations presomatic therapy, at two or more points after initiation of somatic therapy (perhaps in as few as a couple of weeks to a month after the initiation of medication such as methylphenidate which reaches therapeutic levels rather quickly), or whenever dosage or medication is changed. The BASC Monitor Software can then graph Part B (momentary time sampling) results

that can be shared with parents, physician, or other service providers and caregivers. Specific behaviors from Part A can be graphed as well but we would expect individual behaviors to be less reliable indicators of change overall.

In this scenario it is crucial to be able to link somatic therapy to change. To do so, the SOS should be collected concurrently with changes in regimens. We think that the 15-minute time sampling is adequate for this purpose based on our experience and the fact that interobserver reliability did not differ for 15- or 45-minute observations (Lett & Kamphaus 1997). In addition, children receiving a variety of medications including psychostimulants, anxiolytics, antidepressants, and antipsychotic medications require careful monitoring of the effects of these drugs on classroom behavior.

#### **IEP Objectives**

Part A was designed specifically to enhance the development of IEP objectives. Behavior from Part A may then be tracked with the repeated rating of Part A and change graphed by Monitor Software. In fact, some statisticians who have expressed concern about the overreliance on significance testing have noted that graphing is one powerful alternative method for data analysis. We have noted how convincing a graph is to teachers, parents, and others.

We suggest, however, that the clinician observe using Part B prior to completing Parts A and C. We think that the vigilance required to complete the momentary time sampling ensures careful observation that leads to a more accurate rating of the behavior intervention plans in an ongoing fashion.

Finally, because 3 data points are advised to obtain a reliable trend line (Francis, Fletcher, Stuebing, Davidson, & Thompson, 1991), we recommend that, as a minimum, observations be collected at the outset of the school year (after the child has had 1 month to adjust to teachers, peers, etc.), at a midway point when it may be convenient to adjust intervention (certainly March or April of the academic year would be too late), and just prior to the annual evaluation of IEP goals.

#### **Prereferral Intervention**

The evaluation of such intervention can occur in the same framework advised for the

annual evaluation of IEP objectives but on a shorter timetable. Again, a minimum of 3 data points are advised even if the intervention is designed to be brief (e.g., 1 or 2 months). Consider the following example: Shane is a victim of physical abuse by his mother, resulting in his being placed in foster care for 3 months. At the same time his mother is receiving treatment. He is initiating routine counseling sessions at school for the first time. Shane also has a history of distractibility and truancy at school. Shane's truancy could be tracked by event recording during this period, and the SOS could assess his classroom behavior during monthly intervals. SOS results could be of some additive value in assessing the value and the effects of the foster care placement and counseling on his classroom behavior.

#### **Schoolwide Interventions**

While recognizing the impracticality of using the SOS on a large scale, we do think that it could be used for sampling purposes. For example, one or two children deemed to be at risk for aggression could be sampled from each classroom to evaluate the effects of the school's violence prevention program. Good evaluation data are crucial for such programs as some evidence of iatrogenic effects have been noted.

The SOS is designed specifically for classroom-based intervention. SOS results then should not be considered when evaluating home-based intervention unless home and school-based interventions are linked. For example, a homebound reinforcement program may be used to improve behavior at school.

The SOS assesses the frequency of classroom behavioral problems. Consequently, SOS results from Parts A and B may be used to identify behaviors in need of intervention. Specifically, any behavioral problem that is exhibited or adaptive skill that is not exhibited becomes a potential candidate for intervention. Within these groups, problem behaviors of higher frequency can be given priority for intervention. Analogously, low-frequency adaptive skills also become candidates for intervention.

The SOS is unique among Monitor components in that it allows clinicians to prioritize behaviors for classroom-based intervention.

The SOS also measures the "bothersomeness" of a child's behavioral problems via the disruptive category of Part

A. Often children display a number of behavioral problems making it difficult to prioritize behaviors for intervention (Schwanz & Kamphaus, 1997). The ratings of disruptiveness can be used to identify behaviors that should be targeted first for treatment.

#### OTHER APPLICATIONS OF THE BASC Longitudinal Outcome Research

Various components of the BASC are being used in longitudinal investigations to study the risk, onset, course, and progress of behavioral problems and psychopathology in childhood. Some studies have used the BASC as a measure of child outcomes or as the criterion variable of interest.

Nelson, Martin, Hodge, Havill, and Kamphaus (1999) used the BASC TRS and PRS as outcome criteria to assess the predictive validity of early temperament. Nelson and colleagues predicted that preschool temperament would predict later functional behavioral status as assessed by the BASC. Their hypothesis was supported. They found that three temperament constructs rated by parents at age 3 are associated with BASC-TRS-C ratings at age 8. The assessment of problems early in development via teacher ratings may indicate early risk.

These and related findings have considerable impact for the support of early screening and targeted prevention. As teachers use the BASC-TRS, they become quite adept at completing the forms, commonly completing them in 10 or so minutes. A teacher can then complete the BASC-TRS-P for an entire class in about 3 hours. There are few such efficacious approaches to screening for children at high risk for the development of behavioral and emotional difficulties at prereading age levels. The BASC-TRS-P and the -PRS-P are well suited to efficient

402 V. OBJECTIVE METHODS  
screening for identification of high-risk children in the 2½- to 5-year age range.

CHAMPUS, the U.S. military civilian and retiree health care system, began a longitudinal study of adolescents placed in residential treatment centers (RTCs) in 1997. The study is under the direction of Dr. Richard Gaines. Although the data analyses are not yet complete, preliminary analyses and results are reported as being quite good. In this study, CHAMPUS was interested in predicting which adolescents referred for placement would actually benefit from the expensive RTC setting. Gaines (personal

communication, 2001) reports that the BASC has been found to have “good predictive power” in this study, although details remain scant at this writing. We expect the BASC will continue to be used in such situations and that it will perform well due to its integrative development process (i.e., a combination of rational, theory-driven, and empirical methods).

Merydith (2000) used the BASC-TRS-A to assess the effects of violence prevention programs conducted in schools. On the basis of the TRS-A and principal’s nominations, aggressive adolescents were identified for specific intervention programs. After treatment, the BASC scores showed substantial reductions in means for the treatment versus the control group on relevant scales—some exceeding a full standard deviation. These improvements are viewed as remarkable and suggest the TRS-A is sensitive to the effects of even brief intervention programs such as implemented here. The BASC has also been noted to be sensitive to interventions with Head Start children in the younger age range (see Reynolds & Kamphaus, 2002, for a review).

#### **Forensic Applications of the BASC**

Reynolds and Kamphaus (2002) describe a variety of forensic or court-related applications of the BASC. They also note the many special features of the BASC that make it desirable in forensic settings. One key feature is the various validity scales on the BASC components and the ability to triangulate on behavior. An important factor for clinicians to consider in choosing instruments for forensic evaluations is the presence of scales designed to detect dissimulation (Reynolds, 1997). Dissimulation is the act of making oneself (or in the case of rating scales, the person being rated) appear dissimilar or different in some way from one’s actual state. In the legal arena, individuals may have much to gain by appearing to have more or fewer problems than actually exist. Almost any behavioral or emotional disorder can be the subject of dissimulation. As Sattler (1998) notes in his extensive review, dissimulation, especially negative dissimulation or malingering, is difficult to identify. Objective methods are absolutely necessary for the accurate identification of dissimulation and the BASC provides the clinician with one of the few sets of measures for children to detect such

problematic responding.

Reynolds and Kamphaus (2002) review many applications of the BASC in the forensic arena. They suggest the BASC is especially useful in child custody, personal injury (particularly when posttraumatic stress disorder, traumatic brain injury, or emotional pain and suffering are at issue), juvenile certification, determining the needs of adjudicated delinquents, and documenting the need for special educational services.

#### **SUMMARY**

From its SOS to the suite of behavior rating scales, the BASC provides multiple methods for gathering important information for making accurate assessments of children suffering from a wide range of diagnostic as well as subthreshold developmental difficulties.

In the absence of a specific diagnostic determination, the BASC provides the ability to determine a child's placement on a continuum of behavior relative to his or her peers, allowing clinicians to make judgments regarding probability of future problems.

Because it provides a spectrum of information beyond that necessary for identifying clinical pathologies, the BASC is a useful instrument for professionals called on to make recommendations for children who require intervention plans tailored to both nurture their strengths and buttress their weaknesses. The BASC and other similar rating scales provide important dimensional information on profiles of child behavior that provide a more complete

#### **17. The Behavior Assessment System for Children 403**

understanding of a child who may suffer functional impairment but may not meet strict diagnostic criteria under DSM-IV. Children in hospital, school, special education, and other similar settings often fit this profile, requiring accommodations or services without meeting categorical criteria for diagnosis.

The BASC offers a variety of data gathering avenues for clinicians working in school settings who must comply with federal and state standards for educational assessment and monitoring of changes subsequent to intervention. Computerized programs and co-normed scales make cross-informant and multimethod information easy to compare, assimilate, and present to parents and educators. Importantly, the BASC allows clinicians to objectively assess a child's adaptive strengths relative to peers, filling a large gap

in available behavioral measurement tools. The BASC ADHD Monitor is a timely and efficient method for measuring medication and behavioral intervention effects in children with ADHD and is the newest element of this comprehensive assessment system. The BASC rating scales have also been used effectively in research, providing sensitive, accurate measurement in a number of longitudinal studies and proven application in forensic evaluations.

## ACKNOWLEDGMENTS

This chapter is adapted in part from Reynolds and Kamphaus (2002). Copyright 2002 by The Guilford Press. Adapted by permission. Portions of the chapter are reprinted or adapted from Reynolds and Kamphaus (1992) and Kamphaus and Reynolds (1998). Copyright 1992, 1998 by American Guidance Service, Inc. Reprinted/adapted by permission.

## REFERENCES

- Adams, C. A., & Drabman, R. S. (1994). BASC: A critical review. *Child Assessment News*, 4, 1–5.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Bender, W. N. (1997). Medical interventions and school monitoring. In W. N. Bender (Ed.), *Understanding ADHD: A practical guide for teachers and parents* (pp. 107–122). Upper Saddle River, NJ: Merrill.
- Bird, H., Gould, M., & Staghezza, B. (1992). Aggregating data from multiple informants in child psychiatry epidemiological research. *Journal of the American Academy of Child and Adolescent Psychiatry*, 31, 78–85.
- Cantwell, D. P. (1996). Classification of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 37, 3–12.
- Daniel, M. H. (1993, August). *Diagnostic specificity of parents' vs. teachers' behavior ratings*. Paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC.
- diSibio, M. (1993). Conjoint effects of intelligence and adaptive behavior on achievement in a nonreferred sample. *Journal of Psychoeducational Assessment*, 11, 304–313.
- Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and criterion related validity of the behavior assessment system for children–parent rating scale. *Journal of Clinical Child Psychology*, 26, 276–284.
- Fergusson, D., & Horwood, L. J. (1995). Early disruptive behavior, IQ, and later school achievement and delinquent behavior. *Journal of Abnormal Child Psychology*, 23, 183–199.
- Flanagan, R. (1995). A review of the Behavior Assessment System for Children (BASC): Assessment consistent with the requirements of the Individuals with Disabilities Education Act (IDEA). *Journal of School Psychology*, 33, 177–186.
- Flanagan, D. P., Alfonso, V. C., Primavera, L. H., Povall, L., & Higgins, D. (1996). Convergent validity of the BASC and SSRS: Implications for social skills assessment. *Psychology in the Schools*, 33, 13–23.
- Francis, D. J., Fletcher, J. M., Stuebing, K. K., Davidson, K. C., & Thompson, N.M., (1991). Analysis of change: Modeling individual growth. *Journal of*

- Consulting and Clinical Psychology*, 59, 27–37.
- Hightower, A. D. (1986). The teacher–child rating scale: A brief objective measure of elementary children’s school problem behaviors and competencies. *School Psychology Review*, 15(3), 393–409.
- Hoover, H. V. A., Braver, S. L., Wolchik, S. A., & Sandler, I. N. (2000, August). *Teachers’ ratings of children’s classroom behaviors: Time of year effects?* Poster session at the annual meeting of the American Psychological Association, Washington, DC.
- Kamphaus, R. W., & Frick, P. J. (in press). *Clinical assessment of child and adolescent personality and behavior* (2nd ed.). Needham Heights, MA: Allyn & Bacon.
- Kamphaus, R. W., & Reynolds, C. R. (1998). *BASC ADHD Monitor*. Circle Pines, MN: American Guidance Service.
- Kratochwill, T., Sheridan, S., Carlson, J., & Laseck, K. (1999). Advances in behavioral assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 350–382). New York: Wiley.
- Left, N. J., & Kamphaus, R. W. (1997). Differential validity of the BASC Student Observation System and the BASC Teacher Rating Scales. *Canadian Journal of School Psychology*, 13, 1–14.
- Loeber, R., Green, S., & Lahey, B. B. (1990). Mental health professionals’ perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology*, 2, 136–143.
- Merydith, S. P. (2000). Aggression Intervention Skill Training: Moral reasoning and moral emotions. *NASP Communiqué*, 28, 6–8.
- Merydith, S. P., & Joyce, E. K. (1998, August). *Temporal stability and convergent validity of the BASC Parent and Teacher Rating Scales*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.
- Nelson, B., Martin, R. P., Hodge, S., Havill, V., & Kamphaus, R. (1999). Modeling the prediction of elementary school adjustment from preschool temperament. *Personality and Individual Differences*, 26, 687–700.
- Ramsay, M., Reynolds, C. R., & Kamphaus, R. W. (in press). *Essentials of behavioral assessment*. New York: Wiley.
- Reynolds, C. R. (1997). *Detection of malingering in head injury litigation*. New York: Kluwer Academic Press.
- Reynolds, C. R. (2001, October). *Forensic neuropsychological evaluation*. Workshop presented at the annual convention of the American Board of Forensic Examiners, Nashville, TN.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children*. Circle Pines, MN: American Guidance Service.
- Reynolds, C. R., & Kamphaus, R. W. (2002). *The clinician’s guide to the Behavior Assessment System for Children (BASC)*. New York: Guilford Press.
- Reynolds, C. R., & Richmond, B. O. (1985). *Manual for the Revised Children’s Manifest Anxiety Scale*. Los Angeles, CA: Western Psychological Services.
- Sandoval, J., & Echandia, A. (1994). Behavior Assessment System for Children. *Journal of School Psychology*, 32, 419–425.
- Sattler, J. M. (1998). *Clinical and forensic interviewing of children and families*. San Diego, CA: Author.
- Scahill, L., Schwab-Stone, M., Merikangas, K. R., Leckman, J. F., Zhang, H., & Kasl, S. (1999). Psychosocial

and clinical correlates of ADHD in a community sample of school-age children. *Journal of the American Academy of Child and Adolescent Psychiatry*, 38(8), 976–984.

Thorpe, J. S., Kamphaus, R. W., Rowe, E., & Fleckenstein, L. (2000, August). *Longitudinal effects of child adaptive competencies, externalizing, internalizing behavior problems on behavioral and academic outcomes*. Poster session at the annual meeting of the American Psychological Association, Washington, DC.

Vaughn, M. L., Riccio, C. A., Hynd, G. W., & Hall, J. (1997). Diagnosing ADHD subtypes: Discriminant validity of the Behavior Assessment System for Children (BASC) and the Achenbach parent and teacher rating scales. *Journal of Clinical Child Psychology*, 26, 349–357.

Verhuist, F. C., Koot, H. M., & Van der Ende, J. (1994). Differential predictive value of parents' and teachers' reports of children's problem behaviors: A longitudinal study. *Journal of Abnormal Child Psychology*, 22, 531–546.

17.

# Assessing Social Competence in Children and Adolescents

TIMOTHY A. CAVELL BARBARA T. MEEHAN SAMUEL E. FIALA

This chapter focuses on the task of assessing social competence in children and adolescents, particularly when that task is to be conducted by practitioners. We define social competence as a superordinate construct composed of the following subcomponents: *social adjustment*, *social performance*, and *social skills*. We discuss conceptual issues critical to the task of properly assessing social competence within a developmental and ecological context. We then describe various strategies and specific instruments for conducting developmentally informed and ecologically valid assessments. Finally, we provide an example of how one might apply these strategies in a specific case.

## SOCIAL COMPETENCE:

### A MEGA-CONSTRUCT

Most discussions of social competence begin with an effort to define the term. Numerous definitions have been offered but a single, commonly accepted viewpoint has not emerged. Complicating matters is the fact that a wide range of assessment instruments purport to measure social competence or related constructs. Recognizing this dilemma, Cavell (1990) argued that social competence is a superordinate construct with multiple subcomponents. Based on a review of the many and varied ways that social competence has been measured, Cavell proposed a three-tiered model designed to bring greater order and economy to the way psychologists think about and assess this often elusive, mega-construct. At the top of the hierarchy is *social adjustment*, defined as the extent to which children are currently achieving societally determined, developmentally appropriate goals. Next is *social performance*, which refers to the overall quality of children's responses within relevant, primarily social situations. The third subcomponent is *social skills*—specific abilities that children use to produce a given social response.

433

# 19

## Assessing Social Competence in

# Children and Adolescents

TIMOTHY A. CAVELL

BARBARA T. MEEHAN

SAMUEL E. FIALA

## Social Adjustment

In our society, children are generally expected to be physically and emotionally healthy, to be on grade level academically, to follow conventional rules of social and moral behavior, and to have family members and friends who accept them and care about them. The degree to which children are meeting these expectations is their level of social adjustment. Most children have little difficulty achieving these goals; some even have the good fortune of exceeding these expectations. Other children fail to meet the expectations that significant others and society hold for them. Their social adjustment is compromised and it is these children who are most often referred to practicing psychologists or identified as children at risk.

The construct of social adjustment is distinct from other aspects of social competence in that its focus is on macrolevel variables, many of which are thought to be the products of social functioning but none of which involve social functioning per se (Cavell, 1990). Measures of social adjustment are generally of two types. One type is more or less an inventory of an individual's current status across various domains of life. Commonly appraised areas include medical, emotional, academic, occupational, and legal statuses (Zigler & Trickett, 1978). A second type of social adjustment measure involves the assessment of global traits or molar behaviors that are more closely tied conceptually to social functioning and interpersonal relationships. Examples include measures of reciprocated friendship, global self-worth, peer acceptance, rejected sociometric status, and peer victimization.

## Social Performance

In reviewing the social competence literature, Cavell (1990) discovered that child-focused researchers tended to focus their assessment on global adjustment (e.g., peer acceptance) or on specific skills and behaviors (e.g., encoding skills) thought to be predictive of global adjustment. These skills were typically measured independently of social context or within the context of only one or two situations (e.g., peer group entry

and peer provocation). Little attention was given to children's overall performance in other kinds of situations that may also be important to development and adjustment (cf. Freedman, Rosenthal, Donahoe, Schlundt, & McFall, 1978). Cavell suggested that a more explicit focus on the quality of children's performance across a range of situations is needed to understand fully the relation between discrete social skills and overall social adjustment. This is particularly true when assessing the social competence of adolescents or children whose difficulties are unrelated to peer rejection or overt aggression. The social world of the adolescent, for example, is wider in scope, less accessible to adults, and filled with situations in which the criteria for effective performance are not always so clear-cut (Allen, Weissberg, & Hawkins, 1989; Cavell & Kelley, 1992). Measures of social performance assess a child's typical performance in a representative sample of situations. In discussing this construct, Cavell emphasized the effectiveness of performance; however, the nature and pattern of one's performance as well as the situational characteristics that occasion poor performance are also important variables to consider when assessing social performance.

### **Social Skills**

Responding to the demands of a given social situation requires the execution of a variety of skills and behaviors. Among these are overt behaviors such as making eye contact and using appropriate voice volume. Other skills are social cognitive in nature and include the interpretation of available stimuli and the selection of viable response. The social-information processing model of Crick and Dodge (1994) is a valuable heuristic for those interested in assessing children's social cognitive abilities. According to this model, the on-line processing of situational stimuli entails a number of skills including information encoding, response selection, and response enactment. Also relevant, according to the model of Crick and Dodge, are latent cognitive structures (e.g., schemas, memory stores, and internal models) that interact reciprocally with on-line information processing. These internal representations (e.g., of self and of others) are thought to be particularly important deter-

434 VI. ASSESSMENT OF ADAPTIVE SKILLS/BEHAVIOR  
minants of social performance in situations

when information is processed automatically (as opposed to deliberately and reflectively). Maladaptive social responses have been shown to occur primarily in situations that pull for the automatic processing of social information (Rabiner, Lenhart, & Lochman, 1990). Ideally, the assessment of specific social skills takes place in the context of situations that are recurrently problematic. Narrowing the assessment lens in this way allows for a more efficient search for skill deficiencies and other sources of poor performance.

### **INTERRELATIONS AMONG SOCIAL ADJUSTMENT, SOCIAL PERFORMANCE, AND SOCIAL SKILLS**

The hierarchical nature of this model implies that the quality of children's social performance will depend on the level of their social skills. In other words, children who are well equipped with an array of skills should respond more effectively across situations than children who are less skillful. Similarly, children's social adjustment should vary directly with the quality of their social performance. Less effective performance should lead to poorer adjustment and more effective performance should yield greater adjustment. However, Cavell's (1990) model also recognizes that social performance by itself does not determine children's level of social adjustment and social skills are not the sole determinant of the quality of their social performance. Instead, social performance is seen as a necessary but insufficient determinant of social adjustment and social skills are seen as a necessary but insufficient determinant of social performance. A host of factors can contribute to children's social adjustment, even when measured by a single index (e.g., peer acceptance). For example, age, gender, minority group status, athletic ability, and physical appearance are some of the factors, along with social performance, that can influence a child's peer acceptance. But what factors besides social skills affect how a child responds in a given social situation? Cavell (1990) noted a tendency among some researchers to cast the source of poor social performance entirely in terms of skill deficits. The advantage of this approach is that efforts to improve children's social functioning fit easily within a skills training format in which target skills are explained, modeled, and rehearsed. The disadvantage

is that factors other than skill deficits might be ignored. Selman and colleagues (Selman, Beardslee, Schultz, Krupa, & Podorefsky, 1986; Schultz & Selman, 1989) have argued that a *gap* often exists between children's capabilities and their actual performance in a given situation. Making sense of that gap during assessment and closing that gap through intervention will likely require attending to variables other than social skills. Examples include children's interpersonal goals, the payoffs—both real and perceived—associated with various social responses, and the level of emotional arousal that accompanies and potentially disrupts social performance (Cavell, 1990). Researchers who study adult couples (Fincham & Beach, 1999) and parent-child dyads (Webster-Stratton, 1994) have begun to recognize that skill-deficit assumptions can sometimes limit efforts to improve troubled relationships and hurtful interactions. Likewise, practitioners who assess the nature and cause of children's interpersonal difficulties should recognize that a child's social skills or abilities are only part of the assessment picture and that nonskill factors (e.g., goals, incentives, and interfering emotions) can also contribute to poor performance. Parents and teachers certainly recognize when there is a gap between children's skill and performance (e.g., "We've seen how good he can interact with his classmates"), so practitioners who offer skills training as the only antidote to poor social functioning may be overestimating the social validity of their intervention. Practitioners should also recognize that measures *within* a given level of social competence could have little overlap. This is due primarily to the nature of the constructs. A child's academic success, a common aspect of global indices of adjustment, may bear little relation to other measures of adjustment such as emotional functioning or social status. Similarly, effective social performance in one situation (or kind of social situation) will not guarantee effective performance in a different situation. Situational demands can vary greatly and the skills required for effective performance can differ significantly from one situation to the next (Cavell & Kelley, 1992; Schlundt & McFall, 1987). For example, a child who is assertive and outspoken might do well in situations

that call for peer refusal but could struggle in situations in which the expectation is to defer to teachers' authority. One typically finds greater overlap when measuring specific social skills, especially if the context for assessment is a common situation or set of situations. However, assessments conducted across a range of situations, even within a single skill area, can lead to a reduction in scale or item overlap (Yoon, Hughes, Cavell, & Thompson, 2000).

### **CRITICAL ISSUES IN THE ASSESSMENT OF SOCIAL COMPETENCE**

As is the case when assessing any aspect of a child's personality or behavior, the assessment of social competence is unlikely to yield accurate and productive information if conducted in a simple, cookbook fashion.

We would argue, however, that the assessment of children's social functioning can be especially problematic for practitioners given the scope and complexity of this construct.

The temptation is to latch too soon and too tightly onto instruments that promise to provide accurate information quickly and easily. Therefore, we offer the following six caveats or recommendations to practitioners who assess social competence in children and adolescents.

#### **Stay Abreast of the Research Literature**

Ladd (1999) has recently described research in the area of social competence and children's peer relationships as being in its third "generation." The first generation (pre-1970) was highlighted by studies documenting the predictive power that came from knowing about a child's problematic peer relationships (see Parker & Asher, 1987, for a review and critique of these studies). Second-generation researchers (1970s and 1980s) studied the causes and correlates of poor peer relationships and found evidence that children's social behaviors affected their social adjustment, in particular their acceptance by peers: Competent behaviors lead generally to positive peer relations and coercive, disruptive behaviors often lead to peer rejection and social isolation. Second-generation researchers also found evidence for social-cognitive deficits in children who struggled interpersonally and socially. These two decades also witnessed a number of attempts to directly alter children's social difficulties via skills training interventions. This era ended with looming questions, however, about the effectiveness of such interventions and

about the causal status of poor peer relationships in the development of later maladjustment (Ladd, 1999; Parker & Asher, 1987). According to Ladd (1999), the 1990s marked the beginning of a third generation of research characterized by two broad investigative approaches. One approach was to continue the pursuit of enduring research agendas and the other was to follow more innovative lines of inquiry. Examples of the former are recent studies showing that children rejected by their peers represent a more diverse group than was originally thought (i.e., not simply children who are aggressive). Other studies revealed that children's social reputation among peers—apart from their behavior—can have a significant influence on peer acceptance (e.g., Hymel, Wagner, & Butler, 1990). Research into children's friendships has also expanded to include efforts to explore both the processes (e.g., conflict) and provisions (e.g., intimacy) associated with different kinds of relationships (e.g., friends vs. acquaintances) and the relationship differences of varying groups of children (e.g., young children, adolescents, and delinquent youth). There have also been recent attempts to study those behaviors that affect the maintenance and dissolution of friendships, and not simply the formation of these relationships. Noteworthy is a study by Parker and Seal (1996) that assessed children's friendships over time, enabling the investigators to identify a set of distinct friendship trajectories. These trajectories were labeled rotation, decline, growth, stasis, or friendless depending on the temporal pattern exhibited. Third-generation researchers have also continued their examination of the cognitive underpinnings to social behavior (Crick & Dodge, 1994). Among the issues explored are the overly positive self-views found among aggressive children (e.g., Hughes, Cavell, & Grossman, 1997) and the important distinction between defensive attributions evinced primarily by emotionally reactive children versus the antisocial goals often adopted by instrumentally aggressive children (e.g., Crick & Dodge, 1996). Another enduring agenda is that concerned with the origins of children's social competence, in particular the impact of harsh parenting on children's social cognitions (e.g., Nix et al., 1999). Researchers

who study children's peer relationships have delved more deeply into the relative benefits of peer acceptance versus close friendships (e.g., Parker & Asher, 1993) and the nature and course of bully-victim relationships (Kochenderfer & Ladd, 1996).

Investigations that pursue a more innovative research agenda include studies that examine whether poor peer relations contribute directly to later maladjustment or simply occur incidental to other causal factors such as aggressive behavior (Parker & Asher, 1987). The evidence to date suggests that being actively isolated from one's peers can produce its own negative sequelae. The role of emotional dysregulation and atypical physiology has also been the subject of recent research related to children's social competence. Attention has also been focused on gender, cultural, and ethnic group influences on children's social functioning and friendship patterns. Particularly impressive has been the work of Crick and her colleagues on the topic of relational aggression (Crick, 1996; Crick & Grotpeter, 1995).

Given the prominent role of coercive behavior in the development of poor peer relations, Crick's work has helped to elucidate the range of ways in which aggression can be expressed by children and adolescents. For example, relational aggression is viewed as a form of coercion that involves hurting another's relationships (e.g., spreading rumors) or hurting another *via* relationships (e.g., excluding someone from a group and threatening to end a friendship). Crick and her colleagues have found that girls are more apt to aggress relationally than physically and that relational aggression is predictive of peer acceptance and other adjustment indices. Differences in the trajectories of girls' versus boys' friendships (e.g., Parker & Seal, 1996) and features that characterize the friendships of relationally aggressive girls (Grotpeter & Crick, 1996) have also been the focus of recent study.

This overview of three generations of social competence research, admittedly brief and cursory, is designed not so much to inform as to caution readers about the dynamic and increasingly sophisticated work being done in this field. Practitioners should not limit their choice of assessment strategies and measurement devices to those based solely on findings from earlier, outdated studies.

## Adopt a Developmental Psychopathology Framework

We have suggested that assessing and understanding children's social competence is a complex task. One way to organize the information gathered during the course of assessment is to work from a developmental psychopathology perspective (Cicchetti & Toth, 1992). Current models of human development and child psychopathology have become increasingly sophisticated and detailed. As researchers generate new findings about the development and course of dysfunctional social behavior in children, practitioners face the challenge of incorporating this new knowledge into their assessment behavior. The developmental psychopathology framework can serve as a useful, overarching guide. It addresses the transactional nature of children's interactions with the environment and emphasizes the processes by which these transactions promote or disrupt children's developmental organization across multiple systems (e.g., biological, cognitive, social, and familial). An organizational approach to understanding children's development makes the following assumptions (Mash & Dozois, 1996):

1. The individual child plays an active role in his or her own developmental organization.
2. Self-regulation occurs at multiple levels, and the quality of integration within and among the child's biological, cognitive, emotional, and social systems needs to be considered.
3. There is a dialectical relation between the canalization (or crystallization) of developmental processes and the changes experienced through the life process.
19. Assessing Social Competence 437
4. Developmental outcomes are best predicted through consideration of prior experience and recent adaptations examined in concert.
5. Individual choice and self-organization play important roles in determining the course of development.
6. Transitional turning points or sensitive periods in developmental processes are most susceptible to positive and/or negative self-organizational efforts.

The developmental psychopathology model, though at times unwieldy, can be a useful heuristic for understanding complex developmental issues that are important when assessing and treating children's social difficulties.

For example, highly aggressive behavior in children is often linked to a number of dispositional (e.g., difficult temperament) and environmental (e.g., harsh parenting) variables that can operate in reciprocal and recurring ways. This presents a potential burden to practitioners because cause and consequence are often irretrievably tangled. On the other hand, therapeutic gains could have wide-reaching effects if these recurring transactions shift and become more positive and growth promoting.

### **Gather Both Nomothetic and Idiographic Information**

A complete and accurate assessment of a child's social competence requires two kinds of information. The first is information relative to what is normative for a given group of children (e.g., middle school African American girls). The source of this information is empirical research documenting the tendencies of children in general or on average. Occasionally, this information is provided with the assessment instrument in the form of norm tables. Unfortunately, most measures of children's social functioning either lack normative information from a representative standardization sample or have norms from samples that are limited in size and scope (cf. Gresham & Elliott, 1990).

The second kind of information that practitioners need is *idiographic* data regarding the child in question. Shirk and Russell (1996) suggest a lack of attention to idiographic information and an overreliance on nomothetic data (and vice versa) can mislead the practitioner. They argue that in the nomothetic approach to clinical practice characteristics of the individual case tend to be equated with characteristics of the group as a whole. Consequently, it becomes tempting to move directly from diagnosis to intervention without an adequate assessment of the individual characteristics of the specific case. The logic goes something like this: If aggressive children have been shown to be deficient in problem-solving skills, and a child is identified as aggressive, problem-solving therapy is often judged to be the treatment of choice. (p. 266)

Idiographic questions address the specifics of problem onset, severity, and duration as well as the presence of various risk and protective factors that can function to exacerbate or ameliorate a child's later adjustment. One might ask, for example, what unique set of factors conspired to produce

severe social isolation in a particular child?  
Is the problem long-standing or a recent development?  
What factors are now maintaining the isolation? Is social isolation evident in multiple settings? Is social isolation a source of distress or is the child nonplussed by his or her social detachment. Are there additional adjustment concerns (e.g., learning problems and hyperactivity)? What positive attributes (e.g., intelligence, athleticism, and attractiveness) does the child possess? How healthy and how socially skilled are the child's parents? Do the parents actively promote social participation on the part of their child? Questions of this sort are needed to gain an understanding of the child in question and to plan any intervention that follows.

Armed with both nomothetic and idiographic information, practitioners can better serve the needs of children and the adults who refer those children. Practitioners who are familiar with normative information may know a lot, for example, about the "typical" sociometrically rejected child, but they will not know about a specific peerrejected child. Gaining that knowledge will require additional, idiographic information. On the other hand, normative data about rejected children can help identify the issues to consider when conducting an individualized assessment.

#### 438 VI. ASSESSMENT OF ADAPTIVE SKILLS/BEHAVIOR

##### Recognize the Influence of Age, Gender, Socioeconomic Status, and Cultural/Ethnic Group Membership

The criteria by which children are diagnosed with a psychiatric disorder or deemed eligible for special education services are, for the most part, standard guidelines applied in a fairly uniform fashion regardless of children's demographic background. Evaluating children's interpersonal functioning is less straightforward. It requires sensitivity on the part of the practitioner to beliefs about social behavior that might vary from one demographic subgroup to another. Variables commonly assessed when evaluating a child's social functioning are often loaded with subjective meaning and are thus vulnerable to in-group biases. Peer rejection, intimacy between friends, interpersonal conflict, the use of physical aggression, and assertiveness are all variables that can hold different meanings depending on children's age, gender, socioeconomic status

(SES), and their identification with a particular cultural or ethnic subgroup. Aggressive behavior displayed by an African American boy in a third-grade classroom of predominantly African American students is less likely to be associated with peer rejection than the same behavior exhibited by a female, European American student in a predominantly majority high school (Henington, Hughes, Cavell, & Thompson, 1998; Stormshak, Bierman, Bruschi, Dodge, & Cole, 1999). Because of the lack of empirical research and published norms on the social functioning of children in particular demographic subgroups (e.g., low income, Latina middle school students), practitioners must carefully balance concerns about children persisting in the use of problematic social behavior with the possibility that such behavior is neither unusual nor improper when examined against local norms. Of course, behavior that is normative for a particular subgroup could detract from the development of more adaptive bicultural competence and place a child at risk for later maladjustment. Studies investigating the assessment of adolescent social performance offer a less obvious example of how subgroup norms can flavor the interpretation of social behavior. Gaffney and McFall (1984) developed a measure designed to assess the social functioning of delinquent girls. The instrument contained a range of empirically derived problem situations to which girls were asked to give role-play responses. In developing the scoring criteria for those responses, the investigators asked teenage girls to judge the effectiveness of responses to all situations involving peer-peer interaction. Unexpectedly, scores on these items were the only ones that failed to differentiate between delinquent and nondelinquent girls. One explanation for this finding is that the two groups of girls interact similarly when they are with peers. Alternatively, the scoring criteria for these items may have been inappropriate given the purpose of the measure. Perhaps beliefs held by teenage girls about what is effective peer-to-peer behavior may diverge widely from what parents, teachers, and other adults view as acceptable behavior. And it is adults who are most apt to find fault with the troublesome behavior of delinquent girls. Allen and his colleagues (1989) found that adolescents who

ascribed to adult-oriented values were less accepted by their peers. Considered in light of Gaffney and McFall's findings, it would seem that effective social performance is in the eye of the beholder. Therefore, practitioners must consider whose "eyes" they represent and why.

### **Emphasize Function over Topography**

A behavior's topography is its surface-level appearance, its overt manifestations. Assessment instruments and intervention programs can lead practitioners astray when the topography of social behavior is unduly emphasized. Assessment strategies that begin with extensive lists of social skills or social deficits are examples of this kind of overemphasis. Cross-checking children's behavior against such lists is inadequate as a means of evaluating poor social performance. Jacobson (1992) voiced a similar criticism in his review of current trends in behavioral marital therapy: "neither the intervention strategies nor the targets of intervention stemmed from a functional analysis of couple interaction. Instead, the techniques were pulled from available behavior therapy technology, while the targets of intervention were derived from a 'matching to sample' philosophy" (p. 496). Unfortunately, these same trends are evident in the social skills training literature and the pitfalls are not always evident to practitioners who rely solely on information from glossy marketing brochures.

Practitioners are better served when they recognize that discerning the functional aspects of social behavior is a necessary complement to the identification of behavioral excesses and deficits. Imagine a boy referred for an evaluation because of his poor peer relations and hurtful playground behavior. Based strictly on its topography, this behavior is quickly identified as a problem that needs to be addressed. A more complete assessment, however, would include an effort to understand the functional aspects of this behavior—the variables that occasion it, the consequences that reinforce it, and the factors that would potentially decrease its occurrence. The possibilities are many. Perhaps this boy has a delay in the development of his language that impedes his ability to express himself verbally when a sudden goal arises. Perhaps he is inept at playing common schoolyard games and becomes

frustrated with his own incompetence. Perhaps he becomes emotionally dysregulated when the play activity becomes rather tense or exciting. Perhaps he has witnessed the use of violent behavior in his home. Perhaps he is a bully who uses aggression instrumentally to obtain certain goals (e.g., access to playground equipment). These scenarios are not mutually exclusive and the list is not exhaustive. It should highlight, however, how behaviors with similar topographies can be maintained by very different factors.

### **Consider Lifespan Implications**

This entire volume focuses on children, adolescents, and the psychological problems that arise during their development. A great deal of research in the area of social competence and interpersonal relations has also involved child and adolescent populations (Ladd, 1999). When children suffer socially, emotionally, and academically they need and deserve the attention of competent professionals. Practitioners respond promptly in order to reduce children's distress and the distress that they may be causing others. It is also true, however, that such intervention is driven out of concern for children's subsequent fate as adults. A common assumption is that later adjustment is predicated on significant events that occur and enduring habits that are formed early in life. Applied to children's social functioning, this assumption is well founded but often dormant. Needed is a greater emphasis on constructs, assessment instruments, and treatment programs that adopt an explicit lifespan perspective (e.g., Werner & Crick, 1999).

The ecologies of childhood greatly influence how psychologists assess and treat social maladjustment, poor social performance, and deficits in specific social skills.

This is appropriate and productive but can sometime lead to assessments that are developmentally accurate but clinically sterile.

Parent's homes, teachers' classrooms, and school playgrounds are the arenas within which children operate interpersonally, but these are not the same arenas in which they will operate as adults. And, yet, ecological discontinuity does not equate with developmental discontinuity. Patterns of maladjustment, poor performance, and skill deficits can persist over time and across changing ecologies. Often the outward

forms these patterns take may not be recognized as continuities from earlier developmental periods. Adding to the difficulty in recognizing such continuities is the tendency among researchers to separate the study of child and adolescent social competence from the study of adult social competence (Cavell, 1990). As a result, bodies of research often go unlinked despite clear potential for mutual education and crossfertilization. The benefits of adopting a lifespan perspective are well illustrated in a recent chapter by Geiger and Crick (2001). These authors took on the rather daunting task of reviewing research that addresses childhood precursors to adult personality disorders. Heavily influenced by a developmental psychopathology perspective, Geiger and Crick sought to identify conceptual linkages between childhood vulnerabilities and the symptoms of personality disorders. Geiger and Crick began their task by identifying 440 VI. ASSESSMENT OF ADAPTIVE SKILLS/BEHAVIOR seven distinct themes that captured the range of interpersonal problems associated with personality disorders. These seven themes were

1. Hostile, paranoid world view
2. (a) Intense, unstable, inappropriate emotion or (b) restricted, flat affect
3. (a) Impulsivity or (b) rigidity
4. (a) Overly close relationships or (b) Distant/avoidant relationships
5. (a) Negative sense of self, (b) lack of sense of self, or (c) exaggerated sense of self
6. Peculiar thought processes and behaviors
7. Lack of concern for social norms and needs of others

Geiger and Crick (2001) then considered each theme in light of findings from studies that examined similar constructs in children and adolescents. What emerged was an insightful look at possible precursors to adult personality disorders. By adopting a lifespan perspective, Geiger and Crick also generated a number of recommendations for future research and intervention. For example, children who exhibit a particular *pattern* of interpersonal difficulty in the face of important developmental tasks might benefit from intervention programs designed to prevent the full emergence of personality and other disorders. Recognizing the significance of such patterns requires a lifespan approach in which the goal is not simply

measuring variables but assessing children as whole, developing persons (Bergman & Magnusson, 1997).

## **GOALS, STRATEGIES, AND TOOLS FOR ASSESSING CHILD AND ADOLESCENT SOCIAL COMPETENCE**

In this section we review in more detail the methods for assessing social competence in children and adolescents. In line with Cavell's (1990) three-tiered model, we discuss (in order) the assessment of social adjustment, social performance, and social skills. For each level of social competence we address assessment goals, common assessment strategies, and specific measurement tools. Our review of assessment instruments is selective and based whenever possible on the psychometric quality of the choices available.

### **Social Adjustment**

#### **Assessment Goals**

The primary goal when assessing social adjustment is to obtain a global conceptualization of children's functioning across important domains. To meet this goal, one must consider the extent to which children are achieving developmentally appropriate statuses. Specific statuses that have been deemed important indicators of social adjustment include relational (e.g., having friends and peer acceptance), academic (e.g., grade level and academic achievement), emotional (e.g., not meeting diagnostic criteria for disorders), and legal (e.g., nonadjudicated) statuses (Cavell, 1990). Common reasons for conducting such an assessment would be (1) to evaluate the overall benefits of an intervention program, and (2) to explore the range of vulnerabilities associated with a particular risk factor (e.g., parental neglect). A more specific assessment goal is to focus exclusively on children's interpersonal adjustment, including their level of peer acceptance, the quantity and quality of their friendships, and the degree to which classmates bully them. School psychologists, for example, might wish to identify those students who are being victimized by peers, provide an intervention, and then reassess their adjustment when the intervention is completed.

#### **Assessment Strategies and Tools**

##### **Archival Records**

Although relatively little effort is involved in the collection of archival data, it can provide valuable information regarding children's

adjustment. Academic records can provide data on children's grades, class attendance, standardized tests scores, retention status, and their participation in academic resources (including resource room, bilingual education, tutoring, etc.). Information regarding education-related diagnoses and special class placements can also be extracted from these records. In a similar vein, information obtained from health care professionals (e.g., physicians and psychologists) and from local social service or juvenile justice agencies can add greatly to the assessment of children's current and past levels of adjustment.

#### 19. Assessing Social Competence 441 Self-Report Questionnaires and Interviews

Self-report data can make unique and significant contributions to the understanding of children's level of adjustment (La Greca, 1990). Unlike data collected from other informants, self-report data provide direct access to a child's internal psychological states (e.g., depression, anxiety, and loneliness) and self-perceptions. Of course, self-perceptions are especially prone to bias (Sedikides & Strube, 1997), so obtaining these data should be seen as a strategy to complement others and as a vehicle for gaining insight into children's own view of themselves.

An aspect of social adjustment that is commonly assessed via self-report is children's self-worth or self-esteem. Traditionally, negative views about the self have been considered indicative of poor social adjustment. Supporting this notion are studies that link diminished self-worth to depression in children and adolescents (e.g., Gotlib, Lewinsohn, Seeley, Rohde, & Redner, 1993). However, more recent studies suggest that extremely positive self-views may also be a problem. Several studies have shown that aggressive children tend to report highly inflated and self-serving views of themselves (e.g., Edens, Cavell, & Hughes, 1999; Hughes et al., 1997). Therefore, the assessment of children's global self-worth must consider the possibility that children could be over- or underestimating their view of themselves.

Closely related to the construct of self-worth is self-concept. Unlike the global and one-dimensional nature of self-worth, children's self-concept is typically viewed as multidimensional (Harter, 1988b; Trent, Russell, & Cooney, 1994). Children's evaluations

of their competencies tend to differ across domains, and the number and specificity of these domains increases with age as children move into new arenas of social interaction (Harter, 1988a). For example, physical appearance and level of intimacy in friendships become increasingly important as children transition to adolescence. Therefore, assessment involves measuring and appropriately weighting these domains in a manner that correspond to children's developmental level.

Given that children are considered the most accurate reporters of their own selfworth, it comes as little surprise that self-report measures are used most frequently to obtain this information (La Greca, 1990). Self-report rating scales also offer an efficient and cost-effective method for learning how children perceive their competencies across various domains. Several measures are available that possess good psychometric properties as well as useful norms for both child and adolescent populations. Some of the more widely used measures are the Pier-Harris Self-Concept Scale (Piers, 1984), the Multidimensional Self Concept Scale (MSCS; Bracken, 1992), the Self Description Questionnaire II (Marsh, 1992), the Self-Perception Profile for Children (Harter, 1985), and the Self-Perception Profile for Adolescents (Harter, 1986). With the exception of the MSCS, each of these measures provides a global self-worth score in addition to scale scores for individual domains of functioning (e.g., academic competence, peer acceptance, athletic ability, physical appearance, romantic appeal, job performance, and relationships with parents). Moreover, each measure adequately addresses the increasing number and differentiation among domains of selfconcept that correspond with developmental progression.

Children's self-report can also be used to assess the quality of their interpersonal relationships. Among the available measures are the Friendship Quality Questionnaire (Parker & Asher, 1993), the Dating and Assertion Questionnaire (Levenson & Gottman, 1978), and the Loneliness and Social Dissatisfaction Questionnaire (Cassidy & Asher, 1992). An advantage to using these self-report measures is the ability to collect quickly and easily information on children's subjective experience of their relational status. This information can be particularly

important when there is a discrepancy between children's self-reports and those of others. For example, it would be important to know whether children feel lonely despite being well liked or even popular among their peers.

#### 442 VI. ASSESSMENT OF ADAPTIVE SKILLS/BEHAVIOR

We would also include as measures of social adjustment a number of questionnaires that purport to measure children's social skills. We refer here to self-report scales that ask children to rate the extent to which they perform certain positive (e.g., interpersonal skills, self-control, cooperation, and assertion) and negative (e.g., aggression, impulsivity/recalcitrance, and jealousy) social behaviors without regard to a specific social context. Examples include the Matson Evaluation of Social Skills With Youngsters (MESSY; Matson, Rotatori, & Helsel, 1983), the Social Skills Rating System for Students (SSRS-S; Gresham, & Elliott, 1990), and the Teenage Inventory of Social Skills (TISS; Inderbitzen & Foster, 1992). These instruments are often used to gauge children's level of adjustment at school or with peers.

Self-report questionnaires can also be used to assess children's global emotional and behavioral adjustment. Commonly used measures include the Behavior Assessment System for Children—Self-Report of Personality (BASC-SPR; Reynolds & Kamphaus, 1992), the Youth Self Report (Achenbach & Edelbrock, 1991), the Children's Depression Inventory (Kovacs, 1979), the Reynolds Adolescent Depression Scale (Reynolds, 1986), the Social Phobia and Anxiety Inventory for Children (Beidel, Turner, & Hamlin, 2000), and the Revised Children's Manifest Anxiety Scale (Reynolds & Richmond, 1978). We refer readers to other chapters in this volume that offer more detailed information about these instruments.

Children's perceptions of their adjustment can also be obtained via interviews, both structured and unstructured. A number of formal diagnostic interviews are available for practitioners who want to rule out the presence of a psychiatric disorder. Examples include the Diagnostic Interview Schedule for Children (DISC; Costello, Edelbrock, Dulcan, Kalas, & Klaric, 2000), the Diagnostic Interview for Children and Adolescents (DICA; Reich, 2000), and the Interview

Schedule for Children and Adolescents (Sherill & Kovacs, 2000). Also available are structured interviews designed to evaluate the presence of specific disorders that involve social functioning (e.g., Silverman & Nelles, 1988).

Informal interviews can also be a valuable aid when assessing children's social adjustment. Unstructured interviews allow practitioners to explore in more depth the same global constructs that are commonly measured with paper-and-pencil questionnaires.

Interviews can also be used to clarify confusing questionnaire items or to follow up on ambiguous questionnaire responses. As with self-report measures, the range and complexity of questions will increase with the child's age and corresponding cognitive abilities. Questions can focus on children's achievements and competencies in specific areas (e.g., friendships, academics, and athletics) as well as their perceptions of their roles in relationships with peers and adults. For example, children might be asked (1) to identify areas of success and areas of difficulty, (2) to discuss characteristics they really like about themselves as well as characteristics they would like to change, and (3) to describe how others tend to see them. Of course, children's limited verbal skills and cognitive abilities can make for a difficult interview unless practitioners recognize that they carry primary responsibility for providing a developmentally appropriate format.

Readers are encouraged to apprise themselves of common recommendations when interviewing children (see Hughes & Baker, 1990; Merrell & Gimpel, 1998; Sattler, 1998).

#### Other Report Questionnaires and Interviews

Global measures of children's adjustment are often obtained from others who know them or regularly interact with them. Parents, teachers, and peers are the most commonly tapped resources for adjustment information, but other informants (e.g., siblings, coaches, and school principals) may also play a role. Because these individuals are often familiar with children's adjustment in a specific domain (e.g., home and school), relying on only one source can severely limit the quality of the information obtained. Due to differences in perceptions, setting demands, and children's performances across areas, others' reports may share little overlap. Similarly, disagreements between children's

self-reports and the reports of others are commonplace and do not simply reflect measurement error (Achenbach, McConaughy, & Howell, 1987). For example, a child may disclose feelings of social anxiety on a self-report questionnaire, but parents or teachers (or both) may be unaware of the child's internal distress.

The most widely used tools for assessing others' global judgments are checklists and rating scales. Parent and teacher versions of child self-report instruments are commonly available. Examples of other-report rating scales include the Behavioral Assessment System for Children—Parent and Teacher Rating Systems (Reynolds & Kamphaus, 1992), Achenbach's (1991a) Child Behavior Checklist for parents and the Teacher Report Form (Achenbach, 1991b), and the Conners Parent and Teacher Rating Scales (Conners, 1997). These checklists provide simple means of obtaining significant others' perceptions of children's academic, affective, and behavioral adjustment. Also available is Harter's (1988b) Ratings of a Child's Actual Behavior, a measure that parallels the Self-Perception Profile and that allows for direct comparison to children's report of their academic, behavioral, and social competencies. In addition to the aforementioned measures, teachers are often called on to rank students according to some indicator of adjustment such as peer acceptance or global rating of aggressiveness or social withdrawal.

One of the most valid and useful sources of data regarding children's social adjustment is how they are perceived by their peer group. A common method of obtaining these data involves the use of sociometric questionnaires. These questionnaires ask peers to nominate, rank, or rate children on various dimensions of social status or social behavior. Positive peer nominations generally ask children to identify classmates with whom they would most like to play and negative nominations ask about classmates with whom they would least like to play. Both kinds of information are needed to create indices of social preference and social impact, which are used to form sociometric groups (Coie, Dodge, & Coppotelli, 1982). An alternative to the use of negative nominations is a peer-rating technique in which the lowest scores are treated as a negative

nomination (Asher & Dodge, 1986). This is a more palatable technique because certain children are not singled out for negative nominations. Peer ratings can also be used to assess children on a variety of adjustment-related criteria (e.g., liking, aggression, and quality of teacher–student relationships). Also available are standardized sociometric scales such as the Pupil Evaluation Inventory (Lardon & Jason, 1992) and the Friendship Questionnaire (Bierman & McCauley, 1987). Regardless of which technique is used, peer sociometrics require access to a child’s peers and an ample number of participating peers; otherwise, the validity of the information becomes less reliable as the number of peer informants decreases (Foster, Inderbitzen, & Nangle, 1993). Interviewing those who know children and interact with them can also shed light on issues of social adjustment. Indeed, peer measures are often administered as individual interviews in an effort to ensure confidentiality and reduce any discomfort. Also, structured diagnostic interviews that parallel children’s diagnostic interviews are commonly available. Informal interviews can also be used to gauge parents’ and teachers’ perceptions of children’s adjustment. Checklists and rating scales can be used as springboards for these interviews. For example, one might score and interpret checklists while interviewing parents or teachers, thus facilitating discussion of critical items and patterns of responses. Important topics to cover include whether children have close friends, the length and course of those friendships, and the nature of children’s status among peers. Interviews can be especially helpful when practitioners are assessing an individual child and do not have access to an entire classroom of peers. Children’s siblings can also be a valuable source of information. Siblings who are close in age or grade to the target child often have helpful insights unavailable to adults.

## **Social Performance**

### **Assessment Goals**

Recall that social performance refers to the nature and overall quality of children’s responses to relevant social situations. Because social performance reflects the interface between situational demands and a child’s response to those demands, the source of poor performance can involve as-

pects of the situation, the child, or both. Therefore, the primary goal when assessing social performance is to identify the situations that are associated with poor performance. Ineffective responses to a wide range of situations suggest the presence of skill deficits or other child factors that transcend the demands of a single situation or set of situations. Children exhibiting pan-situational ineffectiveness are also likely to experience more severe social maladjustment unless they possess protective factors that are not performance based (e.g., physical attractiveness) or that are relatively distinct from social performance (e.g., athletic ability). A more circumscribed pattern of social difficulty suggests the presence of situational stimuli that are selectively affecting social performance. It would be a mistake, however, to ignore children's social performance in nonproblem situations. When practitioners have to reconcile differences in the quality of social performance across situations, especially when those situations contain common elements, critical insights can be learned about the factors that may be causing poor performance (e.g., a hypercritical teacher, taunting from several peers). The second goal of assessing social performance is to evaluate the nature and quality of children's actual responses to identified problem situations. Important issues to consider when evaluating children's actual or typical responses include (1) the level or overall quality of children's performance (e.g., how deviant from that which is considered normative or socially acceptable), (2) consistent themes reflected in the pattern of children's ineffective performance (e.g., coercive vs. submissive), and (3) the kinds of task demands and contextual features that are common among or unique to problematic situations. Information about these issues is needed before practitioners can adequately conduct a functional analysis of poor social performance and before they can assess for possible social skill deficits.

#### **Assessment Strategies and Tools Self-Report Questionnaires, Interviews, and Role Plays**

Children and adolescents are often well aware of the situations that are causing them interpersonal problems. Some occur frequently; others are highly public and emotionally charged. Cavell and Kelley (1994) developed the Checklist of Adolescent

Problem Situations (CAPS) as a way to tap into adolescents' knowledge of which situations were potential pitfalls. This 75-item instrument contains seven subscales drawn from a pool of problem situations generated by a sample of seventh, ninth, and 11th graders. Principal components analysis of adolescents' frequency and difficulty ratings resulted in the following factors: parents, siblings, school, make friends, keep friends, work, and problem behavior. Cavell and Kelley found that scores on the CAPS effectively distinguished adolescents whose families had relatively high levels of conflict and parental problem drinking from adolescents whose families had relatively low levels of conflict and parental problem drinking, respectively. The CAPS can be used as a broad screen for problematic situations, or selected subscales can be used to focus on a particular domain of adolescent interpersonal functioning. Careful interviewing can also highlight the situations in which a child's performance is resulting in significant and recurring difficulty. Interviews can be especially helpful when children are less aware of the connection between poor performance and social adjustment problems. Some children may see their behavior as different but not necessarily deviant or deficient. Respectful queries about the nature of these "differences" can often reveal the nature of their poor social performance. More challenging are those children who fail to recognize that their behavior is at all distinct from that of others. This lack of social awareness is often associated with a general ineffectiveness in social performance and with greater social maladjustment. Cavell and Kelley (1994) found indirect support for the adaptive role of recognizing potential social difficulties when they examined gender differences in CAPS scores. Girls rated problem situations as more difficult than did boys, despite evidence from other studies that girls are more attuned relationally and more adept socially. Given the subjective nature of rating a situation's difficulty, Cavell and Kelley used cross-product scores that weighted ratings of situational difficulty by the corresponding frequency rating. In this way, situations rated as difficult but infrequent are given lower scores than situations rated as moderately difficult but frequent.

Another approach to identifying problematic situations is to evaluate the quality of children's performance across a broad range of situations. Two issues are important here. First, the situations should be representative of those encountered by the target population. Second, the criteria for judging the effectiveness of performance should be standardized and empirically based. Assessment instruments that meet these criteria are typically referred to as behavioral inventories and most have been developed via the behavior-analytic method of scale construction (Goldfried & D'Zurilla, 1969). This methodology relies on a series of criterion analyses (1) to identify representative situations, (2) to generate a range of possible responses, and (3) to develop reliable and valid scoring criteria. Behavioral inventories offer content-valid ways to assess social performance across a range of situations that often lack clearly defined goals or effectiveness guidelines. The format of behavioral inventories can vary. A commonly used format entails asking children to enact their anticipated response to each situation and their performance is then evaluated against preestablished effectiveness criteria (Dodge, McClaskey, & Feldman, 1985; Freedman et al., 1978; Gaffney & McFall, 1984). An alternative to the role play is a forced-choice, questionnaire format that greatly reduces the time needed for scale administration and response scoring (Cavell & Kelley, 1992; Freedman et al., 1978; Gaffney & McFall, 1984). An example of the forced-choice format is the Measure of Adolescent Social Performance (MASP; Cavell & Kelley, 1992). This paper-and-pencil questionnaire contains 50 multiple-choice items that describe situations previously identified as both frequent and difficult for adolescents regardless of their grade (seventh, ninth, or 11th), sex, ethnicity (European American or African American), and SES (lower, middle, or upper class). Each item also contains four response options that reflect different levels of effectiveness (based on adult judges' ratings of a previously obtained pool of responses). The MASP has been shown to correlate with teacher ratings of behavior (Cavell & Kelley, 1992) and to predict peer rankings of friendly/cooperative behavior (Cavell, Johnson, & Constantin, 1993). Despite empirical support for the validity

of the MASP, the forced-choice format is one that can fall prey to the influence of social desirability and verbal ability, thereby obscuring assessment of adolescents' typical performance (Gaffney & McFall, 1984). These shortcomings are less critical when the goal is simply to screen for potentially problematic situations. However, if the goal is to explore carefully the nature of children's social performance, role-play formats are more suitable (Gaffney & McFall, 1984). Some instruments, such as Freedman and colleagues' (1978) Adolescent Problem Inventory (API), were developed with empirically derived scoring criteria for evaluating the effectiveness of role-played responses. Responses to the API have been shown to differentiate between delinquent and nondelinquent boys and to predict the level of problem behavior within a group of delinquent boys. Level of effectiveness, however, is not the only type of information needed when conducting an evaluation of children's social performance. As noted earlier, it is also important to consider the possibility of consistent response styles or themes and to search for specific task demands or contextual features that reliably predict poor performance. These data allow practitioners to begin to understand why a given child or adolescent is having recurring difficulty. When practitioners discover that poor performance is associated with certain patterns of responding or with specific situational features they are better able to predict future difficulties and to plan effective interventions. Practitioners who search for patterns of social ineffectiveness and for the situational determinants of poor performance are also less likely to leap prematurely to assumptions about specific skill deficits. Children's tacit goals, the level of available incentives, the relationship history of the interactants, and the presence of stimuli that pull for conditioned emotional responses are possible determinants of poor social performance that also deserve consideration. Persistent response themes, despite their ineffectiveness, can also help elucidate the causes of children's social difficulties. Once identified, these themes can be more intensely examined for the possibility of overt and covert (i.e., social cognitive) skill deficiencies (see later). Two patterns of social ineffectiveness that have been studied extensively by researchers

are aggressive strategies and submissive strategies (Michelson & Wood, 1980). These two strategies are typically considered against the more adaptive use of assertive, nonaggressive responses. A different kind of dichotomy is that between autonomy and relatedness. Both are considered adaptive; thus overreliance on strategies that achieve one without the other can be problematic. For example, Kuperminc, Allen, and Arthur (1996) found that adolescents who relied on strategies that reflected autonomous-related reasoning also demonstrated greater social problem-solving skills and academic competence. However, acts of delinquency were more frequent among adolescents whose strategies showed little relatedness striving and lacked an understanding that relationships can support both autonomy and relatedness.

Another useful distinction is that offered by Selman's (Selman et al., 1986; Selman & Schultz, 1988) typology of interpersonal negotiation strategies (INS). He and his colleagues proposed that adolescents' response strategies fall generally into one of the four levels (ordered by increasing level of competency): (1) the *impulsive/egocentric* use of escape or force, (2) the *subjective/unilateral* use of submission or command giving, (3) the *self-reflective/reciprocal* use of communication and turn taking, and (4) the *collaborative* attempt to find mutually satisfying solutions. Cavell and Kelley (1992) relied heavily on Selman's INS hierarchy when selecting response options for the multiple-choice format of the MASP. Rather than select randomly from those responses that judges agreed were at a certain level of effectiveness, Cavell and Kelley chose responses at each level of effectiveness that mapped closely onto Selman's typology. The seven interpersonal themes identified by Geiger and Crick (2001) can also be used as a guide to discerning patterns of poor social performance. Recall that these themes were seen by Geiger and Crick as potential precursors to adult personality disorders.

Self-report questionnaires can also be helpful, especially if the respondents have some awareness about the nature of their ineffective social behavior. As such, many of these measures have been designed for use with adolescent populations. For example, the Adolescent Social Self-Efficacy Scale (SEFF; Connolly, 1989) measures adolescent's

perceptions of their own ability to perform certain skills during peer interactions, and the Adolescent Interpersonal Competence Questionnaire (AICQ; Buhrmester, 1990) assesses self-perceptions of adolescents' interpersonal skills (e.g., self-disclosure, providing emotional support to friends, management of conflicts, negative assertion, and initiation of friendship). These measures suffer from a lack of situational specificity, but the exclusive focus on peer-peer interactions enables practitioners to conduct a cursory screen for patterns of adaptive and maladaptive social performance.

#### Other Report Questionnaires and Interviews

Instruments that survey significant others can be used to ask parents, teachers, or peers to report on target children's typical response to specific situations or to make judgments about the effectiveness of responses to certain situations. Ideally, one would gather these data from multiple sources, each of whom is familiar with the child's behavior in a specific setting. Behavioral inventories lend themselves to being formatted as otherreport questionnaires. For example, King (1993) developed alternative forms of the MASP (Cavell & Kelley, 1992) that can be completed by parents, teachers, and peers. Because items on the MASP describe situations occurring in the family, school, and peer domains, reports from parents, teachers, and peers, respectively, allow for direct comparisons between a child's view of their typical responses and others' perceptions of those responses. Social performance measures developed by Dodge and colleagues (1985) and Buhrmester (1990) also offer versions that can be completed by other informants (e.g., teachers and peers).

Informal interviews are certainly an option when standardized questionnaires are not available. Interviews with parents and teachers often allow for much greater flexibility in how one gathers information about social performance. Significant others can be asked to identify situations in which the child demonstrates performance problems as well as performance successes. Interviews can be used to explore hypotheses about certain patterns of performance or consistency in the kinds of situations that are causing difficulty. Interviews with parents and teachers can also yield insights into the standards of performance to which children

are held, which may be too high or too low.

### **Direct Observation**

Although less commonly used, this method of obtaining information about a child's social performance can be quite informative.

To conduct direct observations in a cost-effective manner would require the prior identification of those situations that are likely to give rise to poor performance. For example, a practitioner may learn that choosing a table at lunchtime and participating in physical education classes are particularly problematic situations for a socially anxious girl. The circumscribed nature of these events allows for relatively easy observation of the girl's actual responses. The opportunity to observe the girl, the setting, the other interactants, and the unfolding events in real time can reveal a wealth of information that might otherwise be unavailable to practitioners. Concerns that observer reactivity would affect these observations can be partly obviated by conducting the observation before beginning a face-to-face evaluation with the child. Direct observation is especially challenging when working with adolescent populations given the frequency of intimate dyadic interactions (Inderbitzen, 1994). One disadvantage to the use of direct observation to assess children's social performance is that quantifying the information obtained often requires familiarity with a structured coding system. Most practitioners are unlikely to go to such lengths.

Therefore, a common recommendation is for practitioners to select at random an additional child to observe so that there is a basis for direct comparison that holds some ecological validity.

### **Social Skills**

#### **Assessment Goals**

The primary goal at this stage of assessment is to identify *specific* skill deficits that are impeding effective performance. A comprehensive assessment of children's social skills addresses overt behaviors, social cognitive processes (encoding information, decision making, response enactment, etc.), and emotion regulation skills. Complicating this assessment task is the fact that children do not perform these behaviors in an easily discernible and predictable sequence; rather, these skills generally occur in a more or less instantaneous and reciprocal fashion. Focusing the assessment lens on a particular skill requires careful manipulation of the situational

context. Behavior is not allowed to “run free,” as would be the case when assessing the quality of children’s overall social performance. Instead, situational stimuli are presented in a way that isolates the target skill and holds constant all other skills (Cavell, 1990). For example, children’s ability to interpret accurately the intentions of another child would be assessed independently of their ability to generate an appropriate response, to anticipate the likely outcomes of a given response, or to enact successfully a selected response. In a sense, practitioners are much like directors of a play, staging first one scene and then another in a way that belies the fluid nature of actual social performance.

#### Assessment Strategies and Tools

##### Self-Report, Interviews, and Role Plays

As mentioned earlier, a number of self-report questionnaires that purport to measure social skills fail to assess skills within a specific situational context. As such, these instruments essentially measure molar behaviors or global traits rather than specific social behaviors or skills (Inderbitzen, 1994). Self-report measures can certainly be used to assess the extent to which children possess certain social skills, but rarely are these types of data obtained using a paper-and-pencil format. Such instruments would have to focus on a narrow subset of situations that have been identified as especially difficult for a target child or a target population (e.g., aggressive children). For example, self-report measures of children’s social cognitive skills often focus on situations in which a peer is provoking the target child in some way or the target child is seeking to gain entry into a peer group activity. These scenarios could be presented to children and the context for each situation modified systematically to isolate the specific behavior or skill in question. However, because skills-based assessment requires a much richer level of detail than that typically produced by paper-and-pencil measures, most instruments of this type are structured individual interviews or role-play procedures. The Social-Cognitive Assessment Profile (SCAP; Hughes, Hart, & Grossman, 1993) is an example of a structured interview designed to assess the social cognitive skills of children prone to using aggressive behavior. Children are presented with eight vignettes

accompanied by line drawings depicting hypothetical situations in which a child appears to be provoked by a peer. Although similar to other tools assessing specific skill areas (e.g., the Means-End Problem-Solving [MEPS] Procedure, Platt & Spivack, 1975; the Normative Beliefs about Aggression and Aggressive Behavior Measure, Huesmann & Guerra, 1997; the Social Goals and Strategies Measure, Renshaw & Asher, 1983), the SCAP provides a more comprehensive measure of children's social cognition by yielding scores on the following dimensions: causal attributions, social goals, aggressive and prosocial solutions generated, and expected consequences of acting aggressively or prosocially (Hughes, Meehan, & Cavell, 2003). In a recent validation study involving confirmatory factor analyses, Hughes and colleagues (2003) found support for these four SIP factors as well as for two situational factors representing peer provocations that were either overtly aggressive or relationally aggressive. Regression analyses predicting teacher and peer ratings of aggression supported the criterion-related validity of the four SCAP scales. Dodge and his colleagues (Dodge, Laird, Lochman, Zelli, & Conduct Problems Prevention Research Group, 2002) reported similar findings in a study that combined scores from a variety of different measures. Both of these studies support Crick and Dodge's (1994) reformulated model of children's social information processing and argue for assessing multiple aspects of children's social cognitions. Important advantages to using the SCAP are that it is easily administered and scored, appears to be valid for racially diverse populations of children, and includes items depicting physically aggressive and relationally aggressive provocations. Perhaps the most common means of assessing children's social skills is the role-play procedure. Standardized lists of role-play scenes can be found in the Social Skills Test for Children (SST-C; Williamson, Moody, & Granberry, 1983) and the Behavioral Test of Interpersonal Competence for Children—Revised (BTICC-R; Hughes et al., 1989). Although traditionally limited to the assessment of overt behaviors, these procedures can be modified so children can be asked to verbalize the social cognitions that accompany exhibited behaviors. Role-play enactments are generally enhanced when situational stimuli are presented live or in a

videotaped format rather than simply as verbal descriptions or two-dimensional line drawings. For example, Dodge's (1980; Dodge & Frame, 1982; Dodge & Newman, 1981) seminal work on the information processing errors of aggressive children relied heavily on role-play procedures in which children responded to audiotaped prompts from child confederates supposedly in the next room.

Practitioners should not feel obligated to stick to generic role-play lists when selecting performance scenes. In fact, role-play assessments may have greater social validity when children and adolescents enact responses to recent, actual events. Kern (1991) found that asking subjects to reenact their responses to actual incidents yielded a more valid assessment of their level of assertiveness than that produced by standard role-play scenarios. If the situation in question is indeed a recurring problem, then focusing on the details of a recent incident should prove to be a helpful exercise.

#### **Other Report Questionnaires and Interviews**

Gathering reports from others regarding a child's social skills is a frequently used strategy. As with many self-report measures, however, others' reports about children's capacity to use a particular skill are often obtained without respect to situational context. Therefore, the utility of these measures for identifying the precise nature of social skill deficits is suspect. Instead, these measures can be used to provide data on how often a child exhibits a specific skill and the perceived importance of that skill (Gresham & Elliott, 1990). The latter is important when base-rate information is unavailable. Ecological validity suffers in the absence of base-rate information because the relative importance of a skill is often reflective of the frequency with which it occurs in the target child's environment. Behaviors may occur frequently and be deemed suitable in one setting (e.g., hand raising in the classroom) but may be infrequent in other settings and thus inappropriate (e.g., hand raising on the playground). Therefore, asking significant others, particularly parents and teachers, to rate the importance of specific behaviors ensures more accurate interpretation of frequency data.

Commonly used rating scales that assess parent or teacher perceptions are the

MESSY (Matson et al., 1983), the Walker–McConnell Scales of Social Competence and School Adjustment (SSCSA; Walker & McConnell, 1988), the Waksman Social Skills Rating Scale (WSSRS; Waksman, 1985), the SSRS-S (Gresham & Elliott, 1990), and the School Social Behavior Scales (SBSS; Merrell, 1993). Information gleaned from peer report measures is typically more global in nature than parents' and teachers' ratings of social skills. For example, the Class Play and Guess Who procedures include only general item content that is not sufficiently detailed to identify specific skill deficits and targets for intervention (La Greca, 1981).

Practitioners may find value in interviewing those who are familiar with a target child's social behavior. This strategy is most productive when practitioners are familiar with the situations that occasion poor social performance and with the pattern of responses that children are most apt to use.

This familiarity would allow practitioners to present to parents, teachers, and others the kinds of staged scenarios that would be reacted to by children themselves. For example, teachers might be asked to give their impressions about the extent to which children would expect that an aggressive response to a provocation would lead to a positive outcome. Information from others, when collected in this focused manner, can help to pinpoint exactly where a child's performance breaks down in a given situation.

#### A CASE EXAMPLE

We end this chapter with the following case example intended to highlight the *process* of conducting an assessment of social competence. Our goal is not to illustrate every possible assessment strategy or tool but to give a more detailed example of how one might apply Cavell's (1990) hierarchical model. Eric was a 15-year-old high school student referred by his pediatrician because of recent conflicts with his mother over schoolwork and school attendance. At the initial meeting it was learned that Eric had been complaining to his mother about feeling physically ill and emotionally distraught at school. Information from his pediatrician and from the initial interview strongly suggested that Eric was capable of attending school and that his mother may have been overreacting to concerns about his health. Encouraged by the therapist to attend

school, Eric did so but continued to express emotional discomfort about his time there. An initial screen revealed that Eric was experiencing adequate adjustment in most areas of his life: He was a good student, he was respectful toward teachers and other adults, he was not involved in any problem behavior (e.g., delinquent acts and substance abuse), and he reported that he had a couple of close, long-standing friends. The one area of concern, according to Eric, was his ability to play sports. This deficiency on his part was a source of continued shame, self-reproach, and depression. At first this attitude was rather confusing because Eric was not a member of any of the school's athletic teams and did not express a particular interest in joining a team. It was soon learned that Eric's desire to be more athletic stemmed from his dissatisfaction with his performance during informal games participated in by a number of his male classmates.

Apparently, Eric would make a tremendous effort to excel at such games only to see his efforts go awry. When this happened, Eric would feel confused, frustrated, and later embarrassed and ashamed.

Having assessed Eric's level of social adjustment and having identified the kind of situations that were causing him difficulty, the next task was to explore further the nature of his responses during these situations.

#### 450 VI. ASSESSMENT OF ADAPTIVE SKILLS/BEHAVIOR

However, to ensure that his difficulties were not more widespread, Eric also completed the MASP (Cavell & Kelley, 1992) to assess the quality of his social performance across a range of adolescent situations. Eric's responses to the MASP were in line with information obtained via interviews with him, his mother, and one of his teachers and failed to support widespread social problems. Therefore, Eric and the therapist focused squarely on the situations in which he was playing sports with his classmates. To learn how he managed himself during these informal games, the therapist conducted a more extensive interview, pushing Eric to describe all that he would do and say in those situations.

The therapist discovered that Eric's usual pattern of responding was to compete intensely but usually with little positive impact on his performance or on his peers' view of him. Eric perceived these events as social failures and a cause for distress and self-loathing.

Further exploration of these feelings

revealed that Eric often felt similarly after his efforts to participate in conversations with a group of his male classmates. In sum, an evaluation of Eric's social performance suggested that he was struggling in situations that involved male group activity, that his style of responding was one of trying to impress his peers, that this was often an ineffective strategy, and that it was usually followed by self-condemnation. Diagnostically, Eric's difficulties were most in line with the criteria for social phobia.

The next assessment task was to identify specific behaviors that were not currently a part of his skills repertoire. One of the more important factors to address was Eric's belief about the most effective way to be accepted by his peers. He had noticed that the more popular students were often successful athletes and were capable of telling jokes and making others laugh. He mistakenly assumed that to be accepted by others, he too would have to "sell" himself in a big way. Role-play assessments of specific incidents from the previous week at school also revealed that Eric lacked some basic conversational skills, in particular the ability to listen in an active, responsive manner. Instead, Eric tried to uphold his part of the conversation with off-topic comments that he hoped would be seen as amusing. Novel to him was the skill of tracking others' comments with nods, paraphrases, and other cues that showed he was interested in and listening to what they had to say.

Intervention with Eric followed closely the results of the assessment. Active listening skills were modeled by the therapist and rehearsed by Eric in session and at school. Eric's distorted beliefs about the strategies that lead to peer acceptance were reconsidered in light of available research findings and Eric's own experiences with using his newfound listening skills.

## REFERENCES

- Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington: University of Vermont.
- Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 profile*. Burlington: University of Vermont.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.
- Allen, J. P., Weissberg, R. P., & Hawkins, J. (1989). The relation between values and social competence in early adolescence. *Developmental Psychology*,

25, 458–464.

Asher, S. R., & Dodge, K. A. (1986). Identifying children who are rejected by their peers. *Developmental Psychology*, 22, 444–449.

Beidel, D. C., Turner, S. M., & Hamlin, K. (2000). The Social Phobia and Anxiety Inventory for Children: External and discriminative validity. *Behavior Therapy*, 31, 75–87.

Bergman, L., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, 9, 291–320.

Bierman, K. L., & McCauley, E. (1987). Children's descriptions of their peer interactions: Useful information for clinical child assessment. *Journal of Clinical Child Psychology*, 16, 9–18.

Bracken, B. A. (1992). *Multidimensional Self-Concept Scale*. Austin, TX: Pro-Ed.

Buhrmester, D. (1990). Intimacy of friendship, interpersonal competence, and adjustment during preadolescence and adolescence. *Child Development*, 61, 1101–1111.

Cassidy, J., & Asher, S. R. (1992). Loneliness and peer relations in young children. *Child Development*, 63, 350–365.

Cavell, T. A. (1990). Social adjustment, social performance, and social skills: A tri-component model of social competence. *Journal of Clinical Child Psychology*, 19, 111–122.

#### 19. Assessing Social Competence 451

Cavell, T. A., Johnson, P. A., & Constantin, L. P. (1993, March). *Perceived attachment as a predictor of adolescent social competence within newly formed groups*. Paper presented at the biennial meeting of the Society for Research in Child Development, New Orleans, LA.

Cavell, T. A., & Kelley, M. L. (1992). The measure of adolescent social performance: Development and initial validation. *Journal of Clinical Child Psychology*, 21, 107–114.

Cavell, T. A., & Kelley, M. L. (1994). The Checklist of Adolescent Problem Situations. *Journal of Clinical Child Psychology*, 23, 226–238.

Cicchetti, D., & Toth, S. L. (1992). The role of developmental theory in prevention and intervention. *Development and Psychopathology*, 4, 489–493.

Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of social status: A cross-age perspective. *Developmental Psychology*, 18, 557–570.

Conners, C. K. (1997). *Conner's Rating Scales-Revised*. Toronto: Multi-Health Systems.

Connolly, J. (1989). Social self-efficacy in adolescence: Relations with self-concept, social adjustment, and mental health. *Canadian Journal of Behavioural Science*, 21, 258–269.

Costello, A. J., Edelbrock, C., Dulcan, M. K., Kalas, R., & Klaric, S. H. (2000). *Diagnostic Interview Schedule for Children (DISC)*. Pittsburgh, PA: Western Psychiatric Institute and Clinic, School of Medicine, University of Pittsburgh.

Crick, N. R. (1996). The role of overt aggression, relational aggression, and prosocial behavior in the prediction of children's future social adjustment. *Child Development*, 67, 2317–2327.

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, 115, 74–101.

Crick, N. R., & Dodge, K. A. (1996). Social information-processing mechanisms on reactive and proactive

aggression. *Child Development*, 67, 993–1002.

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development*, 66, 710–722.

Dodge, K. A. (1980). Social cognition and children's aggressive behavior. *Child Development*, 51, 162–170.

Dodge, K. A., & Frame, C. L. (1982). Social cognitive biases and deficits in aggressive boys. *Child Development*, 53, 620–635.

Dodge, K. A., Laird, R., Lochman, J. E., Zelli, A., & Conduct Problems Prevention Research Group. (2002). Multidimensional latent-construct analysis of children's social information processing patterns: Correlations with aggressive behavior problems. *Psychological Assessment*, 14, 60–73.

Dodge, K. A., McClaskey, C. L., & Feldman, E. (1985). Situational approach to assessment of social competence in children. *Journal of Consulting and Clinical Psychology*, 53, 344–353.

Dodge, K. A., & Newman, J. P. (1981). Biased decision-making processes in aggressive boys. *Journal of Abnormal Psychology*, 90, 375–379.

Edens, J. F., Cavell, T. A., & Hughes, J. N. (1999). The self-systems of aggressive children: A cluster-analytic investigation. *Journal of Child Psychology and Psychiatry*, 40, 441–454.

Fincham, F. D., & Beach, S. R. H. (1999). Conflict in marriage: Implications for working with couples. *Annual Review of Psychology*, 50, 47–77.

Foster, S. L., Inderbitzen, H. M., & Nangle, D. W. (1993). Assessing acceptance and social skills with peers in childhood. *Behavior Modification*, 17, 255–286.

Freedman, B. J., Rosenthal, L., Donahoe, C. P., Schlundt, D. G., & McFall, R. M. (1978). A socialbehavioral analysis of skill deficits in delinquent and nondelinquent adolescent boys. *Journal of Consulting and Clinical Psychology*, 46, 1448–1462.

Gaffney, L. R., & McFall, R. M. (1984). A comparison of social skills in delinquent and nondelinquent adolescent girls using a behavioral role-playing inventory. *Journal of Consulting and Clinical Psychology*, 49, 959–967.

Geiger, T. C., & Crick, N. R. (2001). A developmental psychopathology perspective on vulnerability to personality disorders. In R. E. Ingram & J. M. Price (Eds.), *Vulnerability to psychopathology: Risk across the lifespan* (pp. 57–102). New York: Guilford Press.

Goldfried, M. R., & D'Zurilla, T. J. (1969). A behavioral-analytic model for assessing competence. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 1, pp. 51–196). New York: Academic Press.

Gotlib, I. H., Lewinsohn, P. M., Seeley, J. R., Rohde, P., & Redner, J. E. (1993). Negative cognitions and attributional style in depressed adolescents: An examination of stability and specificity. *Journal of Abnormal Psychology*, 102, 607–615.

Gresham, F. M., & Elliott, S. N. (1990). *The Social Skills Rating System*. Circle Pines, MN: American Guidance Service.

Grotpeter, J. K., & Crick, N. R. (1996). Relational aggression, overt aggression, and friendship. *Child Development*, 67, 2328–2338.

Harter, S. M. (1985). *Manual for the Self-Perception Profile for Children*. Denver, CO: University of Denver Press.

Harter, S. M. (1986). *Manual: Self-Perception Profile*

*for Adolescents*. Denver, CO: University of Denver Press.

Harter, S. M. (1988a). Developmental and dynamic changes in the nature of the self-concept: Implications for child psychotherapy

## Issues in Child Custody Evaluation and Testimony

PATRICE H. BUTTERFIELD

Louis Nizer (1968), the well-known divorce lawyer, states in his book *My Life in Court*: “All litigations evoke intense feelings of animosity, revenge, and retribution. Some of them may be fought ruthlessly. But none of them, even in their most aggravated form, can equal the sheer, unadulterated venom of a matrimonial contest” (p. 153). Gardner (1976) adds: “And of all the titans of matrimonial litigation, the most vicious and venomous by far is custody litigation” (p. 381).

It may be that psychologists and other mental health professionals are requested to become involved in matrimonial problems most frequently when the issues include child custody. But psychologists, as a rule, are reluctant to enter into therapy situations when they know that litigation is involved and a court appearance is likely. Indeed, psychologists refuse so often under these circumstances that clients sometimes misrepresent their reasons for seeking therapy. Once so involved, psychologists can be subpoenaed and then must testify.

The experience of being on the witness stand can be humiliating, frightening, and confidence destroying if the psychologist is unprepared. In spite of the many pitfalls, however, psychologists should not shy away from the opportunity to use their skills, especially as those skills relate to expertise in the personality and behavior assessment of children and their parents. A moral and an ethical responsibility exists to assess individuals as accurately and completely as possible to determine appropriate treatment and recommendations. In cases involving legal disputes over child custody, the psychologist's expertise in assessment can provide a valuable service to help determine placements in the best interests of the child(ren) in question.

The inexact nature of the field of psychology makes psychologists feel vulnerable to attack, especially when subjective test instruments are used to help formulate an opinion or diagnosis. This chapter begins by providing a brief history of the bases for custodial decisions, the reasons why evaluations

may be requested, and criteria for custodial decisions. Models of evaluations and the evaluation report itself are then described and discussed. In addition, this chapter provides suggestions on how to pre-

# 22

## Issues in Child Custody Evaluation and Testimony

PATRICE H. BUTTERFIELD

pare to be an expert witness and some guidelines for dealing with specific problems on the witness stand. This information should help to alleviate the stress and the ambivalence reported by many psychologists in accepting cases that expose them to the legal system, especially cases associated with the highly emotionally charged issues in child custody disputes.

### HISTORY OF CUSTODIAL DECISIONS

The vast majority of child custody arrangements are made through stipulated agreements between parents. In a large majority of these cases, the mother gets custody without the father contesting it. The small percentage of cases tried by a judge or jury tend to be the most difficult and contentious ones. The number of these cases has increased in recent years in part due to the increase in the attempts by fathers to obtain custody or more frequent visitation. Psychologists need to understand the criteria used to determine custody so they can help parents and lawyers determine what custody and visitation arrangements are reasonable to expect or demand if custody is contested.

### Historical Antecedents Guide and Shape Judges in Their Determinations

Prior to 1920, children were considered the property of the father, and if parents divorced the father typically retained custody. The general reasoning was that the father was best for the child because he could provide physical protection, nurturing, financial maintenance, and education (Weiss, 1979) However, as the United States moved from an agrarian to an industrial society during the late 1800s and early 1900s, fathers entered the labor force often in factories, leaving mothers as the primary child care providers (Clingempeel & Reppucci, 1982) Mother love became an important aspect

of a child's development because of the emerging attitude that one purpose of the family was to protect the child from an impersonal world (Weiss, 1979) Freud's influence in the early 1900s was also an important factor. Emotional bonding between mother and child was recognized increasingly as a strong influence in the young child's development.

### **Tender Years Doctrine**

The "tender years" doctrine related to the assumption that a young child's best interests are served by the mother. In the United States, this belief began to be voiced as early as 1830 (Foster & Freed, 1978) but did not become common until the 1920s. The age defined as "tender" varied. Some courts set the range from birth to 6 years of age; others extended it to age 9 or even age 12 (Hodges, 1986). The upper age to which the tender years doctrine applied seemed to increase over time (Weiss, 1979)

Under the tender years doctrine, in a disputed custody case, the father had to prove the mother unfit rather than simply prove that he was the better parent. It was not until the 1960s and 1970s that several states repealed the tender years doctrine because (1) it clearly discriminated against fathers; (2) there was a growing recognition that fathers could be very adequate parents even for young children (Orthner & Lewis, 1979; Santrock & Warshak, 1979); and (3) increasingly, divorced mothers became employed, and therefore not able to remain at home with their children at all times (Hodges, 1986).

### **Best Interests of the Child**

A child custody opinion based on "the best interests of the child" was written in the late 1920s, and rapidly replaced the tender years doctrine to become accepted throughout the country (Hodges, 1986). Today this criterion as a legal basis for child custody decisions has been adopted by most states.

The doctrine of the best interests of the child gives a judge a great deal of discretionary power. Issues of which parent is "at fault" ostensibly were removed from decision criteria by the advent of "no-fault" divorce. However, in spite of the Uniform Marriage and Divorce Act (National Conference of Commissioners on Uniform State Laws, 1971) adopted by many states, the view that one parent is at fault in the dissolution of a marriage (e.g., has had an affair)

and should be punished is still common.

#### 494 VII. ADVANCED TOPICS

The act specifically enjoins that the conduct of a parent that does not directly influence the child should not be considered part of the custody decision. Nonetheless, judges still seem to be influenced by evidence of immorality (Weiss, 1979). Giving custody to such a parent may be perceived as rewarding immoral behavior (Hodges, 1986) and therefore not in the best interests of the child.

Cohabitation as a living arrangement (chosen by many since the 1970s) also seems to offend morality and may affect judges' decisions in custody battles. This is true in spite of another uniform code, the Uniform Custody Code (National Conference of Commissioners on Uniform State Laws, 1971), which should protect adults from bias about this factor. The Illinois Supreme Court, for example, ruled that such living conditions have the potential to do future harm to the characters of children and that their mental and emotional health may be affected adversely (Hodges, 1986). These issues of morality do not take into consideration the quality of the relationships among the adults and the children in question. It is the quality of relationships that psychologists and other mental health professionals often find themselves attempting to evaluate.

#### Current Perspectives

It should be noted that the best interests of the child doctrine seldom is used as the sole criterion in custody decisions (Hodges, 1986). For example, Derdeyn (1975) reported that the interests of the biological parents are given greater weight than the child's. Another study (Felner et al., 1985) reported that in one state only 15% of the lawyers and only half of the judges included the best interests of the child as one of the five most critical factors in deciding custody. Many other factors seem to take precedence in the mind of the court when it comes to custody issues. Some of these include the persistence of the notions that the sex of the child and parent should match and that the parent with the better economic resources should have custody.

Today, two different types of custodial arrangements seem to be the most common alternatives to the mother having sole legal responsibility and residential custody. These

are joint custody and father custody. A less common but growing number of children end up in grandparent custody.

### Joint Custody

The general public often confuses legal responsibility with physical location of the child. It is possible for parents to have joint responsibility for raising children, even if the children physically reside with one parent. In the past 5 to 10 years, joint custody has been the custody decision of choice when neither parent is considered inappropriate in terms of quality of parenting. This approach seems to work best with parents who can communicate fairly well and generally agree on parenting style (Reppucci, 1984).

Although "joint custody" technically means shared legal responsibility for child welfare, in practical terms it also usually means more equal distribution of the time the children spend with the parents. Some children switch homes every 6 months (with visitation with the other parent during this time), and some children switch every day or every few days. Four primary patterns of shared physical custody have been described in the literature by Atwell, Moore, Nielsen, and Levite (1984): (1) long-term (summer-winter; school year-vacations; alternating full years); (2) short term (alternate months, alternate weeks, split week, every other day, and split day); (3) "bird's nest" (parents move in and out of children's home); and (4) free access (children go back and forth from parent to parent at will). All these combinations work best if the parents both reside within the same school district. Problems noted by Ramos (1979) include the possible perception by the children of transience and instability, the increased costs of maintaining two households, and changes in children's and parents' needs over time. The issue of transience for children received a great deal of attention from the media in the late 1980s. Lack of consistency by parents appears to be exacerbated by the maintenance of two entirely separate sets of living conditions. Irving, Benjamin, and Trocme (1984) voiced concerns about joint physical custody for very young children, for emotionally unstable children, for parents who might use their children as a weapon, or for parents for whom joint custody is court-ordered rather than a desired

custody outcome.

If mediation is not required by the original custody decision for solution of conflicts, the parties must return to court if visitation schedules and arrangements do not work well. Psychological evaluation of all parties may be necessary either again or for the first time. The reader is referred to an excellent source by Blau (1984) who recommends highly structured and detailed visitation schedules and environmental arrangements for children under the age of 5. These recommendations can be extremely useful for the psychologist to present in court if it becomes necessary to mandate specific visitation for young children of warring parents.

#### Father Custody

Fathers' parenting styles have become the subject of several studies, as the father as the sole physical custodian (with either joint or sole legal responsibility) becomes a more acceptable and common outcome of custody disputes. Lewis, Feiring, and Weintraub (1981) note that fathers' patterns of parenting do differ from mothers'. As a group, fathers in intact families spend less one-on-one time with children, are more concerned than mothers with sex-role development, spend less time in caregiving activities, and spend more time in play. Lewis and colleagues concluded that although fathers' styles may differ, they appear to be as sensitive and as concerned with child rearing as mothers.

The most extensive direct-measure research (as opposed to case study research) on father custody has been done by Warshak and Santrock (Santrock & Warshak, 1979; Warshak & Santrock, 1983a, 1983b). Results suggest a clear relationship between type of custody and adjustment of the child. Children living with the same-sex parent were better adjusted than those living with the opposite-sex parent. For mother custody, there was little difference in these children's adjustment when compared with that of children in intact homes, but the boys in these studies seemed to fare especially well with fathers who had custody.

It should be noted that the children in this study were between 6 and 11 years of age, and that fathers and mothers may be differentially beneficial to children of each gender over time. This is a research question that requires further attention especially as the nature of the American family has changed

so dramatically over the last two decades. Overall, however, there seems to be no evidence suggesting that fathers who want custody and obtain it are in general inferior to mothers who obtain custody. For psychologists and mental health professionals who will be asked to assist in making recommendations based on research in child custody cases, this is an especially important field to monitor.

#### **Grandparent Custody/Visitation**

Under common law, parents historically have controlled who may interact with their children. Parents may limit or forbid interactions with neighbors, schoolmates, or relatives with whom they disagree on any number of issues (religious, moral, social). In the last 25 years, however, every state has enacted a statute that permits grandparents to petition the court for visitation rights over the objections of the parents. Again, judges are required to consider the best interests of the child in their decisions.

According to Hafemeister and Jackson (2000), supporters of grandparent visitation emphasize that grandparents play important roles in the lives of their grandchildren as mentors, role models, and family historians. Indeed, grandparents often have taken on full responsibility of parenting in cases in which parents are unable or unwilling to do so (due to drug/alcohol abuse or some other serious circumstances). Opponents of court-mandated grandparent visitation argue that it can undermine parental authority and interfere with parents' ability to shape the values and education of their children. Clearly, the conflict caused by such a situation adds more upset to the existing distress of the divorce for children and parents alike. Hafemeister and Jackson recommend an alternative to courts dealing with grandparent visitation. They suggest family mediation (which may include counseling) both at the time of the divorce and again at a later time if necessary

#### **496 VII. ADVANCED TOPICS**

as relationships evolve. Initial mediation sets up a model for conflict resolution and can help families avoid court involvement completely.

#### **REASONS FOR CHILD AND FAMILY EVALUATION**

There are a number of reasons related to child custody for which a mental health practitioner may be asked to evaluate individuals

and provide testimony. Each requires the ability to formulate and articulate expert opinions based upon sound psychological foundations. The psychologist must know how to select appropriate instruments for evaluation, which requires a knowledge of the validity, reliability, and general strengths and weaknesses of a wide variety of assessment techniques. (This aspect of child custody evaluation is discussed in more detail later in this chapter.) The requirements of appropriate instrument selection must be kept in mind as each of the reasons for evaluation is enumerated below.

#### **Parental Fitness**

The question of the mental health of adults caring for children may not be asked solely in relation to custody issues arising from divorce. Parental fitness issues may be raised by social agencies, as in the case of mentally handicapped or mentally ill parents of a physically or mentally handicapped child. Although mental handicaps or controlled (or recurrent) mental illness in itself does not constitute a reason for removal of a child into foster or institutional care, a professional psychologist may need to perform intellectual, personality, and behavioral assessments to help determine custody that is in the best interests of the child.

#### **Custody Changes**

Just as in initial custody proceedings, parents sometimes request changes in custody based on what is termed "change of circumstances." Financial, geographical, or health (physical and/or mental) considerations; wishes of the child (especially after age 12 to 14); and many other factors may precipitate a request for legal changes in custody by either parent. The issues are much the same as in initial custody battles and often are just as emotion charged. After all, if the parties were in agreement, custody changes would probably simply be made without any need for court involvement.

One especially difficult situation which has become more commonplace in the late 1990s involves the relocation of one parent due to a job or marital change. With both parents working, many find that to obtain, advance, or keep a job may entail making a geographic move. If the ex-spouses each remarry, the professional/personal goals of up to four adults complicate the lives of all. Some states (e.g., Missouri) give one expartner the right to stop the custodial parent

from moving anywhere without permission, effectively holding the family hostage.

While the trend today is to make it easier for ex-spouses to go their separate ways, moving still involves petitioning the courts or lobbying state legislators to get new laws passed to remove this veto power from the noncustodial parent.

Other states (e.g., Illinois) make it even harder for the custodial parent to move.

The parent who wants to move must prove it is in the best interest of the child to go. In some cases this has resulted in a change of custody, effectively giving judges the power to say "If you want to go, go; but the kids stay here" with the other parent. The stress of this new legal involvement can severely damage relationships between children and the dissenting parent ("The moving van wars," 2000).

Some parents are now addressing this sensitive issue during the initial divorce proceeding.

Points such as visitation schedule and child support modifications can be specified ahead of time. For example, if a relocation interferes with the agreed upon visitation schedule, visitation may be altered so that children visit the noncustodial parent for longer than the standard or previous periods (all summer rather than 4 weeks) and for three out of the four typical school vacation periods (Christmas, Thanksgiving, and Easter). In addition, child support amounts may be altered if one parent is required to pay travel expenses. Psychologists will be asked to make recommendations on this thorny issue as the economic realities of a mobile work force impacts children of divorce (Levine, 1999).

### **Parental Alienation Syndrome**

Visitation schedules can continue to be a source of conflict after custody is determined for reasons other than a geographical move. One parent may be discouraging visitation with the noncustodial parent in subtle and sometimes unconscious ways. This interference with scheduled visitation has come to be known as Parental Alienation Syndrome, or PAS. This is not to be confused with a child's (especially a teenager's) desire to stay home with the custodial parent in order to attend a school dance or basketball game. PAS is clearly detailed by Gardner (1998), Darnell (1998), and Clawar and Rivkin (1991). The syndrome

can be mild, moderate, or severe, but the psychological drama of the good parent” and “bad parent” becomes evident in that the child will support sometimes wild accusations made against the noncustodial (“bad”) parent by the custodial (“good”) parent in order to avoid visitation. In severe cases, the alienating parent and child develop a psychopathological relationship of shared delusions surrounding alleged atrocities committed by the other parent. Severe trauma can result from such “indoctrination” or “brainwashing” (Clawar & Rivin, 1991). In mild to moderate cases, once the specific syndrome is diagnosed, legal and mental health professionals must work together to enforce visitation with the noncustodial parent. Interestingly, these visitations usually go smoothly once the child is with the noncustodial parent and this phenomena is a key diagnostic feature of PAS. Sanctions are outlined by the court if visitation and counseling requirements are not followed, including, in severe cases, change in custody to the maligned parent. PAS is a new concept to many judges, and psychologists who do forensic work should be well informed about this syndrome as it becomes better understood and documented.

#### **Child Abuse/Neglect**

If abuse or neglect is charged by one parent to another, court intervention may be required. Indeed, if abuse is suspected and reported, social agencies automatically initiate procedures that may result in legal intervention. In such cases, evaluation of all parties and adjustments in visitation and/or custody may well be required. Children have been removed from parents who are chronic alcoholics (nonresponsive to treatment), sexually promiscuous, or mentally incompetent if the parents have not arranged for appropriate alternative care when they are indisposed or otherwise unavailable.

#### **Juvenile Misbehavior**

Juvenile court services often employ mental health workers and psychologists in cases in which children under the age of 17 (16 in some states or for some offenses) are involved in violations of the law. Juveniles and their families may be evaluated for psychopathology, parental fitness, and likelihood of responsiveness to treatment. The testimony of psychologists weighs heavily in determining disposition of such cases, including the possibility of foster care.

## **Adoption**

Following the establishment of the fact that the parental rights of the biological parents have been severed, a determination must be made as to whether a proposed adoptive family can meet the physical and emotional needs of a specific child or children. Many of the factors to be taken into account are similar to those in criteria for custody—the age and health of the adoptive child, the age and health of the prospective parents, religious preferences, financial stability, emotional climate, and marital status (Shapiro, 1984).

## **Marital Reconciliation/Mediation**

In cases in which separated or divorced individuals reconcile, child custody issues may be handled out of court but often will require the assistance of a therapist. However, issues (which perhaps contributed to the initial marital conflict) concerning children, their care, handling, discipline, school problems, and the like will all require the same kind of careful, complete evaluation used in custody conflicts. The largest difference is that communication with both parents is direct, rather than through an attorney or in a courtroom. Reconciliation evaluation obviously also requires a careful assessment of the parents' relationships with each other. The prospect of future court appearances should always be considered.

Regardless of the purpose of the requested evaluation, be it court-ordered or family initiated, the issues of careful test selection, administration, scoring, and interpretation are the same. Clinical interviews may be challenged in court and must be clearly differentiated from simply "talking to the patient."

Subjective test instruments must be qualified as such and supported by behavioral observations. Questionnaires should be selected for their ability to obtain specific types of information about the individual. Standardized tests must be administered strictly according to manual directions. All the psychologist's skills in personality and behavior assessment must be used in order to assist in the making of a clear, concise, and sound decision that can affect the lives of the individuals for a long time to come.

## **CRITERIA FOR CUSTODY EVALUATIONS**

Charnas (1981) noted that there is a lack of guidelines or criteria in the law and social science literature for defining "the best interests

of the child” in the child custody decision-making process. This makes the psychologist’s job a more difficult one. Hodges (1986) explains that in the absence of criteria, an assessment of parent–child bonding, or of which parent is the “psychological parent,” has been used as a guide for child placement. The term “psychological parent” refers to the parent who better fulfills the child’s psychological and physical needs (Goldstein, Freud, & Solnit, 1973). Judges and lawyers are uncomfortable using the bond as a criterion because they prefer to use objective rather than subjective evidence. This is one major reason why psychologists and their test instruments come under such close scrutiny in the courtroom.

Although the order of the lists that follow may vary among judges, lawyers, and mental health practitioners, criteria for custody seem to include the same major issues.

Chasin and Grunebaum (1981) summarize these points well. They recommend favoring the parent with the following characteristics:

1. Is more likely to foster visitation and shows the more objective attitude toward the other parent.
2. Will maintain the greater continuity of child contact with relatives, friends, neighborhood, schools, etc.
3. Has the better child-rearing skills.
4. Shows the greater humanity, consistency, and flexibility in handling the child(ren).
5. Is the one to whom the child is more deeply attached.

Other criteria for custody decisions, provided by Benedek (1972), Woody (1977), and Fine (1980), include the following:

1. The reasonable preference of the child (if the child is of sufficient age; 12 or 14, depending on the state)
2. Support for the parent who will provide geographic stability.
3. Support for the parent with capacity to provide the child with food, clothing, and medical care.
4. Support for the parent willing to provide continued religious training.
5. Avoidance of separation of siblings.

Awad (1978) notes that mental fitness (implying degree of adjustment) of the parent is not necessarily relevant to child custody evaluations and that diagnosis in and of itself may not be a useful criterion. The issue for the psychologist is to determine

whether any existing mental incapacity interferes with parenting abilities and affects the child directly. For example, a diagnosis of schizophrenia in a parent is, in and of itself, not sufficient reason to decide against this individual in a custody case.

Using any of the aforementioned criteria, one is struck immediately with the awareness of the obvious difficulties associated with evaluation in these areas. McDermott, Tseng, Char, and Fukunaga (1978) identified four major problem areas in conducting a custody evaluation. First, they noted that it was difficult to obtain evidence on the natural parent-child relationship or natural home environment. This is probably because a custody evaluation home visit by a psychologist invariably creates high anxiety or a forced tolerance in the home that would not exist otherwise. Second, McDermott and colleagues remind the reader that research indicates that women look less well adjusted at the time of the separation decision, whereas men look worse some time after the separation. Discovering children's real preference for either parent is also difficult, as loyalty conflicts often exist. Third, data in other areas often conflict or are inconclusive. Psychologists may have difficulty predicting the future effects of certain factors on children's later development or over time (e.g., custody award of infants to fathers, cross-sex parent-child combinations, alcoholism recurrence, and religious fervor). The last major problem noted by this study was the difficulty in choosing between "poisons." The more available parent may not have the better skills.

#### **CUSTODY EVALUATION PROCEDURES**

"Custody evaluation" refers to the process of gathering information, interpreting data, and forming and communicating recommendations concerning child custody. As

stated previously, there are no set national standards for such evaluations. Given the complexity of family relationships, the problem of predicting future stability in the midst of upset over a divorce, and the problem of changing developmental needs over time, a judge needs the wisdom of Solomon. The psychologist needs the courage of Samson to explore this minefield and avoid the pitfalls of bias to arrive at a decision in the best interests of the child.

**Precustody Evaluation Negotiations**

Hodges (1986) recommends that the first responsibility of the mental health professional is the welfare of the child. Therefore, he cautions that the professional should avoid representing one parent against the other in a child custody dispute, regardless of who is paying. Indeed, sometimes a psychologist can request to serve as a “guardian ad litem,” thereby avoiding representation of either parent and focusing only on the needs of the child(ren). Another good option is for the psychologist to ask that the evaluation be court ordered. Regardless, it should be made clear in advance (in writing) that the evaluation report will be provided to the court and to both attorneys. Parents should sign a permission-for-evaluation form (if not under court order) and should be informed about the evaluation process, the limits of confidentiality, and the procedure for billing.

Hodges also recommends obtaining fees in advance and placing them in an escrow account to avoid the difficulty of collecting fees from the “losing party” after the court decision is made.

#### **Agency versus Private Practitioner Evaluations**

Custody evaluations conducted by social service agencies, probation departments, or the court itself are probably shorter than those done by private practitioners. Watson (1979) describes the Denver District Court procedure as including the following steps which are still recommended today around the country:

1. An interview with each parent, including information about education, number of marriages, number of children, and military service;
2. Employment and police check;
3. Evaluation of the physical and emotional health of each parent, including physician and therapist reports;
4. Interviews with neighbors and other witnesses if appropriate;
5. Background information from school principal, teachers, social workers, nurses, or other school personnel; and
6. Interviews of children at school by an agency social worker.

A private evaluation may include much of the foregoing but would probably take place in an office. In addition, a comprehensive private evaluation would include the following, according to Blau (1984):

1. Several interviews with each parent

alone. Parents need to be reminded that nothing in the interview is confidential.

2. Several interviews with each child alone.

Gardner (1976) recommends observing preschool children two or three times, and older children three or more times.

#### 500 VII. ADVANCED TOPICS

Children also need to be informed that the interview is not confidential.

3. Interviewing the parent and children together.

4. Interviewing teachers, babysitters, and other significant people in the children's lives. Grandparents may be seen, as well as any new person to whom the children will be exposed who is involved with either parent. Chasin and Grunebaum (1981) suggest also seeking out housekeepers, friends, neighbors, physicians, and psychotherapists.

5. A home visit. This allows determination of the safety of the home setting and information about the sensitivity of the parents to the children's needs.

In addition, Chasin and Grunebaum (1981) suggest telling parents that the practitioner is willing to read any material parents would like to submit and to talk to anyone the parents feel has information that would help the court make a decision about custody.

The type of private evaluation described previously may take up to 40 staff hours in a two-child-family situation, plus staff meetings and report-writing time. The time obviously increases with the complexity of the case. Added to actual court appearance time, such an evaluation can cost anywhere from \$3,000 to as much as \$8,000. Because parents are already paying substantial legal fees, private evaluations are not possible for many families. Many times this large out-of-pocket expense results in the briefer evaluation procedures described at the beginning of this section or in a noncontested divorce/custody proceeding.

#### Test Selection

Psychological testing is most commonly done as part of an evaluation of the children and parents only, not grandparents or others unless a clear need to do so is indicated.

Projective techniques may be used to assess bonding. For example, the Children's Apperception Test, a sentence completion test, and/or the Kinetic Family Drawing procedure may be used with children, while the Thematic Apperception Test may be used

with adults. The Minnesota Multiphasic Personality Inventory or the Rorschach Ink Blot Test may alert the psychologist to some borderline pathology missed in an interview. Other tests may be used at the evaluator's discretion. However, no custody evaluation should be based on the results of psychological testing only. Projective test results, in particular, are meaningful only in light of other information obtained in interviews, questionnaires, and direct behavioral observation. Hodges (1986) warns psychologists never to answer questions in court about the interpretation of a single response, because conclusions reported in writing should be based on the evaluation of patterns of responses.

Blau (1984) recommends close inspection of school adjustment and the child's status, special abilities, needs, and problems. Such an examination should include a thorough evaluation of (1) intellectual status and potential, (2) neuropsychological factors, (3) academic achievement, and (4) personality and personal adjustment. Specific tests and procedures obviously depend on the age and background of the child.

#### Interview Techniques

The reader is referred to helpful reviews by Hodges (1986) and Deyoub and Douthit (1996) for numerous lists of potential questions to ask in interviews with parents, children, and parent-child dyads or triads.

Hodges outlines play situations for interviewing young children. He cautions psychologists to make careful notes of all observations, interactions, and test information. Memory has been shown to be highly unreliable, especially when there is (as is almost always the case) a long delay before reaching court. Another excellent, detailed description of the interview process is offered by Gardner (1976), who is well known for his work in psychotherapy with children of divorce.

If videotaping is used, sufficient time must be allowed for individuals to desensitize to the camera during the interviews.

Audiotaping is viewed as less intrusive and interfering and can be more easily ignored by participants. Obviously, audio and videotaping can be useful memory aids but are time-consuming to review. Taping

22. Issues in Child Custody [Evaluation and Treatment] 501  
does tend to free the evaluator from extensive note taking and thus may be seen by

practitioners as preferable because of this feature.

## THE CUSTODY EVALUATION REPORT

### Writing the Report

The custody evaluation report is a major instrument used by opposing attorneys to formulate court tactics. Because this report must be available to all parties before the court date, and is so closely scrutinized, it is imperative that this written document be accurate, concise, and clear. The report should be written by the major evaluator if a team approach has been employed. It should begin with a statement of the purpose of the evaluation (e.g., "to determine custody placement in the child's best interests") and a summary of the data, techniques, and instruments used in the evaluation.

Writing should be direct, clear, simple, and free from professional jargon. If possible, both parents' strengths should be cited, but the report should state the facts, the opinions formulated from the facts, and the conclusions, avoiding biases. Awad (1978) recommends that for the situations in which there is no clear-cut recommendation, it is useful to report the situation as precisely as possible and to list the advantages and disadvantages of each possible recommendation.

Gardner (1976) reminds the reader that a diagnosis is not necessary in a report. The purpose of the report is to recommend custody placement, and a diagnostic label may serve to distract and confuse the court, thereby clouding the main issue. Gardner also warns against making statements (e.g., quoting a parent) as if they were proven facts. The court views this as hearsay and therefore not admissible. For example, instead of saying, "Mr. X drinks one case of beer daily," the psychologist should say, "If the court accepts as true Mrs. X's allegation that her husband consumes, on the average, one case of beer daily, then I would consider this an argument against his receiving custody of the children." It is important to remember that although the custody evaluation report is a critical factor, the court has the ultimate responsibility for deciding placement.

Clawar (1984) has proposed criteria for evaluating the scientific respectability of a custody report. These criteria serve as useful guides to the custody evaluator. The following points should be kept in mind and included in a written custody evaluation report:

1. A full history of the family and situation(s) with all appropriate data is required.
2. Credentials of the writer (expertise) should be in line with the testimony needed.
3. The amount of time involved in the evaluation should be stated.
4. Sources should be cited, including referral source.
5. The report should focus on patterns or themes rather than isolated incidents or responses.
6. Conclusions supported by other professionals or research should be noted.
7. Appropriate tests should be used for the questions asked.
8. Clarity of the report is crucial.
9. Language of the report should be neutral, not adversarial. Technical terms should be avoided.
10. The context in which observations were made (office, home) should be described, and the writer should state whether the conclusions can be generalized.
11. The limitations of the report should be indicated.
12. Conclusions should be drawn from the material used in the evaluation.

### **Feedback to Parents**

After the evaluation is completed and depending on the adversarial level of contesting parents, the custody evaluator may take one of two stances regarding direct feedback to parents and their attorneys. The evaluator may be asked to meet with the opposing attorneys to explore constructive alternatives to litigation. However, if this outcome is improbable, then Gardner (1976) recommends that professional mediators handle this and that the evaluator maintain a strictly neutral position. The written report should be sent to the court, which will

### **502 VII. ADVANCED TOPICS**

give a copy to both attorneys. Any meetings should then be avoided. The place to air differences of opinion with regard to the psychologist's report should be in court.

### **PSYCHOLOGICAL TESTS IN THE COURTROOM**

Psychologists' expertise in the area of testing is what makes them unique among professionals and in demand for testimony in many types of legal disputes. The development, standardization, and use of psychological tests have had much to do historically with the emergence of psychology as a

scientific field. Since 1954, the American Psychological Association has formally recognized the importance of scientific standards in the development and application of test instruments. In 1974, the American Psychological Association published the original *Standards for Educational and Psychological Tests*; these were published in revised form in 1985 and again in 1999. The standards specify in detail the qualifications of test users, methods of choosing tests for specific purposes, administration and scoring standards to be followed, and acceptable procedures and limitations for the interpretation of psychological tests.

This last point is particularly important, especially during cross-examination in child custody cases. Psychologists who testify about the meaning of psychological tests should realize, warns Blau (1984), that their opinions will be subjected to minute analysis by competent opposing counsel (Poythress, 1980; Ziskin, 1981). Reports and depositions will probably be reviewed by other psychologists specifically for the purpose of identifying errors, distortions, improprieties, and inaccuracies. The psychologist who testifies in court must adhere to the highest standards in order to serve his or her client properly, avoid embarrassment, and protect the good name of psychology. Readers are encouraged to become familiar with the most recent American Psychological Association Standards for these reasons. Similarly, the *Ethical Principles of Psychology* (American Psychological Association, 1992) and the *Standards for Providers of Psychological Services* (American Psychological Association, 1990) should be well known and practiced by any licensed psychologist, especially before going on the witness stand. Studies have indicated (e.g., Ash & Guyer, 1984) that judges are influenced heavily by the testimony of expert witnesses (more so than juries are), and this is why opposing counsel will try so hard to discredit a psychologist on the witness stand. It is important for a psychologist to avoid having his or her entire testimony lose credence because of ignorance of the standards upon which professional clinical psychology is founded.

#### **PREPARING AS AN EXPERT WITNESS**

There are many details about which to be concerned when preparing for testimony in court. However, there are a few important

points to consider before even meeting a proposed client. If an attorney contacts a psychologist to request expert testimony in a child custody case, as mentioned earlier in this chapter, the psychologist should obtain assurance in writing that he or she is appearing as an advocate for the child's best interests and are not in support of either parent. Moreover, in the initial interview with the attorney (often conducted by telephone for 15–30 minutes), the psychologist should obtain the pertinent facts of the case sufficiently clearly to determine that his or her unique and specific professional skills can be applied appropriately to the case. In addition, some attorneys are unaware of the time needed by an expert witness to conduct a thorough examination (as described earlier), to prepare reports, and to research the literature. Therefore, an anticipated court date is a critical piece of information to have before accepting a case. Availability of records and willingness of family members to participate in the evaluation are also helpful to know about before committing time and energy to such an undertaking. Once a case is accepted and the evaluation is complete, much time may pass before actual court testimony is required. Detailed records of dates, time, facts, and documentation must be cataloged carefully for easy retrieval when needed. No detail is too small to record. Blau (1984), in his book *The Psychologist as Expert Witness*, has many useful forms to aid in data organization, and the reader is referred to this excellent source. Deyoub and Douthit (1996) also offer an excellent guide to preparing for court in their book *A Practical Guide to Forensic Psychology*. They suggest that a psychologist should not be intimidated by a court appearance. While the report and testimony of the professional will be used to make custody decisions, such conclusions are based on expert analysis and sound principles. Neither parent has “purchased” the opinion of the psychologist. As long as the report is well written and presented, the psychologist need not fear testifying even under cross-examination. Prior to the court appearance, the attorney should be briefed by the psychologist as well as vice versa, in order to assist in the planning of the kinds of questions that should and will be asked in court. The attorney

can help the psychologist understand the best way to present his or her data; the psychologist can assist the lawyer in formulating questions that will elicit the opinion the psychologist wishes to express. In addition, the expert needs to anticipate in advance what some of the challenges to his or her opinion might be, and should attempt to deal with any weaknesses frankly during *direct* examination. This avoids the appearance of defensiveness upon cross-examination if such weaknesses should be pinpointed by the opposing counsel. Many problems can be avoided if psychologists train the attorneys who will examine them (Denton, 1987).

### Psychologists on the Witness Stand

Psychologists often are ignorant of courtroom dynamics and the rules of courtroom performance. These rules cover who can be spoken to, by whom, when, and how. They also structure what can and what cannot be said in court. In an adversarial situation, these requirements lead attorneys to manipulate the rules to present evidence in support of their case, to discredit the evidence of the other side, and to prevent the other attorney from doing the same (Hodges, 1986)

Attorneys have different styles just as psychologists do (Girdner, 1985) Some are principled bargainers who attempt to negotiate a fair agreement. A few are soft bargainers and like to avoid conflict and risk.

Others are hard bargainers who have winning as their primary goal and rarely negotiate except in an aggressive manner (Hodges, 1986). Regardless of the opposing attorney's style in cross-examining an expert witness, the psychologist should keep several points in mind:

1. The expert witness should be prepared to state professional qualifications clearly and in a well-articulated voice. These include educational level, clinical experience, experience as an expert witness, professional organizations to which the professional belongs or in which he or she holds office, membership on any local committees (especially when they may have to do with the case, such as a council on child abuse), and any other relevant information. The opposing counsel may challenge the credentials, but the judge will determine the witness's status as an expert.

A licensed psychologist in clinical practice generally will be considered qualified to serve as an expert, or it is unlikely that he or she would be on the witness stand.

2. Professional dress is a requirement to avoid offending judges or suggesting lack of respect for the court, as well as to protect professional credibility.

3. The witness should speak slowly and clearly, avoiding jargon. This is an area of criticism of attorneys as well as expert witnesses. One should avoid continually defining words, or worse, ignoring the need for their definition.

4. Joking or wisecracking should be avoided, even if the attorney does this. The witness should be relaxed and unintimidated but should show that he or she takes the role of expert witness seriously (Gardner, 1976).

5. The witness should address the judge as “Your Honor” and learn the names of all attorneys, so that responses to questions may be prefaced with the correct name.

6. The witness may restate an attorney’s questions, changing small words he or she may have used to get the witness to contradict an earlier statement. The question may then be answered as re-

504 VII. ADVANCED TOPICS  
stated (e.g., “If you mean . . . then I would have the following opinion . . .”).

7. The witness should feel free to take notes and/or books on the stand and should not be afraid to use them. However, he or she should be familiar with the location of the information to avoid delays and fumbling. Also, anything taken on the stand may be entered as an exhibit, so only those items should be taken that the witness would not be uncomfortable having the court see.

8. If an attorney asks a multiple-part question and wants a “yes” or “no” answer, the witness should not be afraid to say that the question cannot be answered with a simple “yes” or “no.”

9. At the end of testimony, if an expert witness feels that an important piece of information has not been revealed, he or she may request of the judge an opportunity to present evidence felt to be vital to the court’s decision. In family court, this may well be permitted.

10. If asked about his or her fee, the witness should not be apologetic. Expert witnesses are paid for their time, not their support. Fees for court should be charged at the same hourly rate as fees for office visits.

It is a wise professional who reads as much as possible concerning courtroom etiquette and procedure before going on the witness stand. A valid evaluation and a valuable opinion could well be lost in attorney manipulation. Being prepared helps to ensure that such damage is kept to a minimum in the best interests of the child in a custody dispute. Several additional valuable courtroom tips are offered by Brodsky (1977), Gardner (1976), Hodges (1986), MacHovec (1985), Melton, Petrila, Poythress, and Slobogin (1997), and Deyoub and Douthit (1996).

### **Testifying under Protest**

Any therapist who has had any association with a case can be subpoenaed to testify. Sometimes a therapist is involved because a parent mentions to his or her attorney that he or she or a child was in therapy with the practitioner. Child therapists are particularly loath to testify because of confidentiality, which is controlled by the parent, and which can be damaged sufficiently to destroy benefits achieved by therapy. The child psychologist can request that the judge keep the child's welfare protected by avoiding testimony of information obtained from therapy sessions with the child. If that request is denied, the therapist can request that such evidence be given in chambers so that the parents will not be informed about the child's concerns.

### **SUMMARY**

As psychologists become more aware of the opportunities to serve as expert witnesses, and are more willing to do so, a broadening involvement between psychology and the law can be expected. The increasing numbers of psychologists who have served as expert witnesses or in a consultation role with attorneys suggests that such interaction indeed has grown over the last two decades (Deyoub & Douthit, 1996; Kerr & Bray, 1982; Saks & Hastie, 1978; Sales, 1981). Judges appointed or elected to serve judicial roles in family matters now enter their jobs with broader backgrounds in psychology from their undergraduate training and in law school. It is probable that psychologists

will be called on in even greater numbers in the future to meet the ever-growing demands placed on the legal system, especially as they relate to family matters and children. In the past, courts have understood little about family dynamics or child development; as a result, experts could be qualified from a wide range of backgrounds in the behavioral sciences. However, with the growing sophistication of attorneys and judges in family practice, experts will find it harder to qualify without true ability to field the sophisticated cross-examination coming from more knowledgeable attorneys. Attorneys are aware of the existence of the relevant body of literature and expect the experts also to be knowledgeable. Increased availability of information over the Internet allows all parties access to information previously seen as the purview of professional psychologists only.

There exists a great need for training and continuing education programs to improve psychologists' preparation to meet the challenges most certainly to be presented by family court and child custody issues (as well its other forensic work). With the continually increasing divorce rates in the United States, and the increase in the numbers of cases in which fathers succeed in winning custody, the likelihood of increased need for expert witnesses is great. Too few psychologists feel prepared to meet the demands placed on them by today's legal system. We must meet these demands to continue to be advocates for the best interests of children. Internship experiences and coursework need to be directed at this important and growing aspect of psychological service, lest we be considered negligent in meeting our moral and ethical standards of training and competence.

## REFERENCES

- American Psychological Association. (1990). *Standards for providers of psychological services*. Washington, DC: Author.
- American Psychological Association. (1992). *Ethical principles of psychologists*. Washington, DC: Author.
- American Psychological Association. (1999). *Standards for educational and psychological tests* (rev. ed.). Washington, DC: Author.
- Ash, P., & Guyer, M. (1984). Court implementation of mental health professionals' recommendations in contested child custody and visitation cases. *Bulletin of the American Academy of Psychiatry and the Law*, 12, 137-147.
- Atwell, A. E., Moore, U. S., Nielsen, E., & Levite, Z. (1984). Effects of joint custody on children. *Bulletin of the American Academy of Psychiatry and the*

*Law*, 12, 149–157.

Awad, G. A. (1978). Basic principles in custody assessments. *Canadian Psychiatric Association Journal*, 23, 441–447.

Benedek, E. P. (1972). Child custody laws: Their psychiatric implications. *American Journal of Psychiatry*, 129, 326–328.

Blau, T. H. (1984). *The psychologist as expert witness*. New York: Wiley.

Brodsky, S. (1977). A mental health professional on the witness stand: A survival guide. In B. D. Sales (Ed.), *Psychology in the legal process* (pp. 269–276). New York: Spectrum.

Charnas, J. F. (1981). Practice trends in divorce related to child custody. *Journal of Divorce*, 4(4), 57–67.

Chasin, R., & Grunebaum, H. (1981). A model for evaluation in child custody disputes. *American Journal of Family Therapy*, 9, 43–49.

Clawar, S. S. (1984). How to determine whether a family report is scientific. *Conciliation Courts Review*, 22(2), 71–76.

Clawar, S. S., & Riviin, B. V. (1991). *Children held hostage: Dealing with programmed and brainwashed children*. Chicago: American Bar Association.

Clingempeel, W. G., & Reppucci, N. D. (1982). Joint custody after divorce: Major issues and goals for research. *Psychological Bulletin*, 91, 102–127.

Darnell, D. (1998). *Divorce casualty: Protecting your children from parental alienation*. Dallas, TX: Taylor.

Denton, L. (1987, December). Expert testimony: Law and science. *APA Monitor*, p. 24.

Derdeyn, A. P. (1975). Child custody consultation. *American Journal of Orthopsychiatry*, 45, 791–801.

Deyoub, P. L., & Douthit, G. V. K. (1996). *A practical guide to forensic psychology*. Northvale, NJ: Jason Aronson.

Felner, R. D., Terre, L., Goldfarb, A., Farber, S. S., Primavera, J., Bishop, T. A., & Abner, M. S. (1985). Party status of children during marital dissolution: Child preference and legal representation in custody decisions. *Journal of Clinical Child Psychology*, 14, 42–48.

Fine, S. (1980). Children in divorce, custody, and access situations: The contributions of the mental health professional. *Journal of Child Psychology and Psychiatry*, 21, 353–361.

Foster, H. H., & Freed, D. J. (1978). Life with father: 1978. *Family Law Quarterly*, 11, 321–342.

Gardner, R. A. (1976). *Psychotherapy with children of divorce*. New York: Jason Aronson.

Gardner, R. A. (1998). *The parental alienation syndrome: A guide for mental health and legal professionals* (2nd ed.). Cresskill, NJ: Creative Therapeutics.

Girdner, L. K. (1985). Strategies of conflict: Custody litigation in the United States. *Journal of Divorce*, 9(1), 1–15.

Goldstein, J., Freud, A., & Solnit, A. (1973). *Beyond the best interests of the child*. New York: Free Press.

Hafemeister, T. L., & Jackson, S. (2000, February). Grandparent visitation: Who should decide? *Monitor on Psychology*, p. 81.

Hodges, W. F. (1986). *Interventions for children of divorce: Custody, access, and psychotherapy*. New York: Wiley.

Irving, H. H., Benjamin, M., & Trocme, N. (1984). Shared parenting: An empirical analysis using a large data base. *Family*

## Chapter 14

# ***SENTENCE COMPLETION METHODS***

Sentence completion methods consist of words or phrases, usually referred to as *item stems*, that people are asked to extend into complete sentences. How individuals go about this task tends to reflect the kind of person they are, and their sentence completions often provide indirect clues to their underlying attitudes, affects, and concerns. Sentence completion methods function primarily as a performance-based measure of personality, although they may at times contain self-reports that identify personality characteristics fairly directly. A sentence completion response of “I FEEL depressed” is much the same as endorsing “I feel depressed” on a self-report inventory; unless there is reason to suspect dishonesty, both responses provide direct evidence that a person is in fact feeling depressed.

Sentence completions can also function in the same manner as self-reports when both give rise to slightly indirect inferences. The sentence completion “I WISH that I had never been born” would be taken by most clinicians as suggesting some depression, as would endorsement of the self-report statement, “I wish that I had never been born.” In both instances, the evidence for depression is not as direct as a person’s saying or endorsing “I feel depressed,” but in both, the implications for depression are fairly compelling.

When sentence completions shift from self-descriptions to comments on other people or events, however, they become less direct in their implications than self-report endorsements. Consider a man who responded, “I WISH my father had been a different kind of person when I was growing up.” This sentence completion is not a descriptive self-report, except as it may indicate this man does in fact wish that his father had been a different kind of person—which by itself is not a particularly useful piece of information. What is useful about the response is its indirect suggestion of possible interpersonal issues (e.g., a strained paternal relationship), certain underlying affects or attitudes (e.g., I hate my father; my father was a bad man), or tendencies to externalize blame (e.g., it’s not my fault that I’m the way I am, it’s because of how my father treated me).

As for commenting on an event, consider a woman who responded, “I WISH that I had waited longer to get married.” Like the information that the man in the previous example wishes his father had been a different kind of person, knowing that this woman regrets not having waited longer to get married is not by itself very useful. Indirectly, however, her response provides clues to possible underlying concerns that could be important for her to recognize and a therapist to discuss with her. Does she regret her choice of a spouse and believe that she would have “married better” had she not married so young or too hastily? Or is she happily married but sorry about having become a wife and mother at an early age at the expense of a promising career or opportunities to travel?

Some sentence completion tests have been constructed with standard sets of items for use in general personality assessment; others comprise items selected to examine such specific

### **518 Performance-Based Measures**

characteristics as egocentricity (Exner, 1973), underachievement (Riedel, Grossman, & Burger 1971), moral attitudes (Musgrave, 1984), marital satisfaction (Inselberg, 1964), managerial motivation (Miner, 1978, 1984), and depression (Barton, Morley, Bloxham, Kitson, & Platts, 2005). Sherry, Dahlen, and Holaday (2004) list 47 sentence completion tests that have been described in the literature, and an unknown but probably large number of unpublished item sets have been used by personality assessors for various ad hoc purposes. This chapter focuses on the nature, history, administration, scoring, interpretation, applications, and psychometric foundations of two standardized sentence completion measures for general personality assessment: the Rotter Incomplete Sentences Blank (RISB), which is by far the most widely used sentence completion method in clinical practice, and the Washington University Sentence Completion Test (WUSCT), which is the most widely studied and best validated sentence completion method currently available.

### **NATURE OF SENTENCE COMPLETION METHODS**

Like other performance-based personality assessment instruments, sentence completion tests have the potential to generate three types of data: structural, thematic, and behavioral.

The structural data in sentence completions consist of such objective response characteristics as the reaction times to individual items, the total time required to complete the test, the length of the responses, and the frequency of personal pronouns in the responses. Historically, however, little attention has been paid in the literature to the interpretive significance or possible valid correlates of such structural features of sentence completion tests. Instructions to record this structural information are not included in standard sentence completion test manuals, and it is not common practice for clinicians to take systematic note of them. As for behavioral data, the manner in which people work on the sentence completion task (e.g., whether carefully or carelessly, energetically or wearily, seemingly relaxed or obviously edgy) and how they interact with the examiner (e.g., pleasant or grumpy, deferential or resistive) usually provide clues to their general nature, their current frame of mind, and their interpersonal attitudes. Survey results indicate, however, that sentence completion tests are much more commonly administered in written form than orally, which limits opportunities for examinees to display interpersonal attitudes (Holaday, Smith, & Sherry, 2000).

Written sentence completions, like the written TAT and figure-drawing stories described in Chapters 12 and 13, provide information about a person's handwriting, grasp of grammar and spelling, and language usage. Differences between how people use language when they write and when they respond orally, as during an interview or on other tests, can be a useful source of behavioral data. Is a person's written language formal or informal in style, coherent or convoluted, prissy or profane, bland or dramatic, or in some other respects suggestive of certain personality characteristics? If the person's written and oral language differ in such respects, what implications might these differences in self-expressive style have for how the individual adapts to the requirements of various situations?

The most extensive and informative data source in sentence completions is the thematic imagery they contain. The content of individuals' associations to the item stems typically provides a rich array of clues to their underlying feelings, attitudes, and concerns. In addition to eliciting general information in these respects, sentence completion stems can

#### **Sentence Completion Methods 519**

be phrased to evoke persons' attitudes toward specific individuals (e.g., "MEN "; "A MOTHER ") and their perspectives on particular events and circumstances in their lives (e.g., "IN SCHOOL "; "I FAILED "). Unless indicated otherwise, these examples and the other item stems used for illustrative purposes in this chapter appear in the RISB. The ease with which item stems can be formulated to serve certain purposes or pursue particular lines of inquiry has contributed to the previously noted abundance of general purpose and specific purpose sentence completion tests, and some additional information in this regard is included in the history section of this chapter. Historical considerations aside, the remaining discussion in this section on the nature of sentence completion methods focuses on the two most noteworthy sentence completion measures: the one that is most frequently used in clinical practice (RISB), and the one that has been most thoroughly researched and validated (WUSCT).

The RISB was originally published in 1950 and currently is in its second edition (Rotter, Lah, & Rafferty, 1992). It is a 40-item test printed on the front and back of a one-page,

8<sub>12</sub> × 11 inch form, with 15 items on the front and 25 items on the back. Eight of the 40 item stems consist of a single word (e.g., "PEOPLE "; "SPORTS "), 21 have two words (e.g., "I SUFFER "; "MOST WOMEN "), six have three words (e.g., "THE ONLY TROUBLE "; "WHAT PAINS ME "), and the remaining five are four or five words in length (e.g., "I AM BEST WHEN "; "WHEN I WAS A CHILD "). These examples are taken from the Adult Form of the test. There are two other 40-item forms of the RISB, a College Form for use in college and university settings and a High School Form intended for secondary school students. The College and High School Forms are printed in the same format as the Adult Form but differ slightly from it in their content.

The WUSCT has been published in several versions beginning in 1970, the most recent of which is Form 81 (Hy & Loevinger, 1996; Loevinger, 1998). The WUSCT consists of 36 item stems printed on two pages with 18 items per page. In common with the RISB

stems, the WUSCT stems touch on aspects of how people perceive and respond to such matters as authority (e.g., “RULES ARE ”), frustration (e.g., “IF I CAN’T GET WHAT I WANT ”), relationships (e.g., “MY MOTHER AND I ”), and responsibilities (e.g., “RAISING A FAMILY ”). As a difference from the RISB, however, six of the WUSCT items have alternate forms for men and women (e.g., “SOMETIMES HE [SHE] WISHED THAT ”; “THE WORST THING ABOUT BEING A WOMAN [MAN] ”).

There is also a Youth version of the WUSCT, the SCT-Y, which contains 32 items and is intended for use with young people aged 8 to 18 (Westenberg, Treffers, & Drewes, 1998). Of the 32 Y-SCT items, 21 are identical with WUSCT items, and the other 11 are new or revised items considered appropriate for younger people. Additional information on the adaptation of other sentence completion methods for personality assessment of children of elementary school age is available in contributions by Haak (1990) and Hart (1986).

## **HISTORY**

Sentence completion methods of assessing psychological characteristics first appeared in the context of efforts to construct formal measures of intelligence in children. Herman Ebbinghaus, who was a pioneering figure in the experimental study of human memory, was also interested in how intellectual capacity and reasoning ability develop in young people. In **520 Performance-Based Measures**

pursuit of this interest, he devised what appears to have been the earliest sentence completion task (Ebbinghaus, 1897). Binet and Simon (1905) later included a version of Ebbinghaus’s sentence completion task in their intelligence scale, which eventually became the Stanford-Binet Intelligence Scale and in which sentence completion items were retained through many revisions (see Roid, 2003). As noted by Lah (1989b), a wide variety of sentence completion tests continue to be used in assessing achievement, intelligence, and language skills in young people.

The notion that sentence completion methods could contribute to assessing personality characteristics as well as intellectual functioning originated with some observations by the Swiss psychoanalyst Carl Jung about the possible personal meanings of word associations. Jung (1916) popularized the notion that much could be learned about the inner life of people by eliciting their associations with various words. “Say the first word that comes into your mind,” the person would be instructed, and alternative hypotheses could then be generated concerning the possible meaning of such associations as “MOTHER—good,” “FATHER—scary,” “SEX—dirty,” and “WORK—proud.”

The association method was formalized in the United States by Grace Kent and Aaron Rosanoff, who constructed a Free Association Test consisting of 100 ordinary, everyday words like “table,” “dark,” “music,” “sickness,” and “man.” Kent and Rosanoff (1910) published frequency tables for the occurrence of various kinds of content in Free Association Test responses of 1,000 “normal” adults, 279 “insane” adults, and large groups of “normal” and “defective” children aged 4 to 15. Many other word association lists were published or used informally in subsequent years, probably the best known of which was a 60-item measure developed by David Rapaport, Merton Gill, and Roy Schafer for a seminal research study of the clinical diagnostic application of psychological tests conducted at the Menninger Foundation in the early 1940s (Rapaport, Gill, & Schafer, 1946/1968, chap. 8).

Rapaport et al. departed in two respects from Kent and Rosanoff’s method. First, instead of limiting their list to ordinary and relatively neutral words, they included words with aggressive or sexual content (e.g., “fight,” “gun,” “penis,” “vagina,” “masturbation,” “intercourse”). Second, instead of emphasizing the content of responses, they focused on the interpretive significance of structural features of a word association protocol, including reaction times, tendencies to use synonyms (“STREET—road”) or antonyms (“GOOD—bad”), preference for single word or multiword associations, and such strange ways of responding as repeating stimulus words (e.g., “MOUTH—mouth”) or giving nonsensical clang associations to them (“BEEF—weef”).

Reviews by Daston (1968) and Rotter (1951) include the full list of words used in these and other word association tests, together with summaries of the early clinical and research applications of these tests. The word association technique gained some popularity

in clinical practice and remains visible in parlor games and fictionalized re-creations of psychological evaluations. Over time, however, professional assessors concluded that single-word responses to one-word stimuli do not fully tap the potential of an association method to identify an individual's personality characteristics. Gradually the notion emerged that association tasks could be enriched by replacing the word-word format with brief phrases extended into full sentences, and by the late 1920s, contemporary sentence completion methods had begun to evolve. As for word association tests, survey data indicate that they are no longer among the 50 most widely used psychological tests (Hogan, 2005).

#### **Sentence Completion Methods 521**

Three persons notable for their early work in developing sentence completion formats for assessing personality are Arthur Payne, Alexander Tandler, and Amanda Rohde. Payne (1928, 1930) constructed what appears to have been the first formal list of sentence completion items, which he designed for use in vocational counseling as a source of information about career-related personal traits. Tandler (1930) was interested primarily in emotionality and compiled a 20-item list of sentence stems each beginning with a first-person pronoun and intended to sample an affective state (e.g., "I LOVE "; "I GET ANGRY WHEN "). He described his measure as a test of "emotional insight" and recommended its use in clinical evaluations of emotional responsiveness.

Whereas the Payne and Tandler sentence completion tests were designed with specific purposes in mind, Rohde (1946, 1957) began working in the late 1930s to develop the sentence completion method into an instrument for general personality assessment. Originally published in 1940, the Rohde Sentence Completion Test was the first carefully constructed and validated measure of its kind, and its items covered a broad range of personal issues and experiences. The expressed purpose of the instrument was to "reveal latent needs, sentiments, feelings, and attitudes which subjects would be unwilling or unable to recognize or to express in direct communication" (Rohde, 1946, p. 170). Rohde was clear in her emphasis on the performance-based rather than the self-report aspects of sentence completion assessment.

Rohde's Sentence Completion Test served as a model for numerous similar instruments that were developed in the 1940s and 1950s. Three of these sentence completion tests from a half-century ago helped to shape the format of the method and still find some occasional use. The Sacks Sentence Completion Test consists of 60 item stems, most of which include a first-person pronoun (I, me, my) and were designed to elicit information about a person's family relations, interpersonal relations, sexual perspectives, and self-attitudes (Sacks & Levy, 1950).

The Forer Structured Sentence Completion Test comprises 100 relatively long and specifically focused structured items, such as "WHEN I FEEL THAT OTHERS DON'T LIKE ME, I ." The relatively structured nature of the Forer test, which he supplemented with an elaborate system for categorizing responses, was expected to facilitate treatment planning by identifying a broad spectrum of attitudes and behavioral tendencies (Forer, 1950, 1993).

By contrast, the 73-item Miale-Holsopple Sentence Completion test consists mainly of brief, open-ended stems (e.g., "A WOMAN'S BODY "; "BEHIND ONE'S BACK ") that are designed to maximize subjectivity in the response process. Interpretation of the Miale-Holsopple test emphasizes examiner impressions, rather than categorizations, with the aim of achieving global descriptions of people, especially their unconscious attitudes and concerns (Holsopple & Miale, 1954). Before turning to the history of the RISB and the WUCST, information about contemporary usage of sentence completion methods is relevant to note.

#### **Usage of Sentence Completion Methods**

Asked about their use of sentence completion tests in the survey by Holaday et al. (2000), members of the Society for Personality Assessment identified five such measures by name: the aforementioned Sacks, Forer, and Miale-Holsopple sentence completion tests, which 1% or 2% of the respondents reported using with adults or adolescents; the RISB, which

#### **522 Performance-Based Measures**

was reported as being used by 28% of the respondents in their evaluations of adults and 19% in their evaluations of adolescents; and the WUSCT, reportedly used by 2% of respondents

with adults and 1% with adolescents. Because of its previously mentioned extensive research base, the WUSCT was selected along with the relatively frequently used RISB for detailed presentation in this chapter.

As for the overall usage of sentence completion methods, 53% of clinical psychologists responding to a survey by Camara, Nathan, and Puente (2000) reported using some form of sentence completion test in their practice. This 53% usage frequency made sentence completions the fourth most commonly used personality assessment method instrument among these respondents, following the Minnesota Multiphasic Personality Inventory (MMPI), the Rorschach Inkblot Method (RIM), and the Thematic Apperception Test (TAT). Other surveys have indicated particularly wide use of sentence completion methods in evaluations of young people. In one sample of 162 child and adolescent clinical psychologists working in diverse settings, sentence completion methods were reported as being the most frequently used personality test and were exceeded in frequency of use among all tests only by the Wechsler Intelligence Scale for Children and the Child Behavior Checklist (Cashel, 2002). In another study of testing practices, 362 psychologists working with adolescents in a variety of settings reported using only Wechsler scales and the RIM more frequently than sentence completion methods (Archer & Newsom, 2000).

Sentence completion methods have held some appeal for forensic as well as clinical psychologists, especially in child custody cases. Among forensic psychologists surveyed by Ackerman and Ackerman (1997) and Quinnell and Bow (2001), 29% to 30% reported using sentence completion methods in their evaluations of children in custody cases, and 22% to 26% reported using them in evaluating the adults in such cases. These usage frequencies made sentence completions the second most frequently used performance-based personality assessment measure in these child evaluations, following the TAT/Children's Apperception Test, and the third most frequently used in the adult evaluations, following the RIM and the TAT. In a survey by Archer, Buffington-Vollum, Stredny, and Handel (2006), 27% of the responding psychologists reported using sentence completion methods in forensic evaluations of adults, a frequency exceeded among performance-based measures only by the RIM, at a 36% frequency of use.

Finally worth mentioning with respect to interest in sentence completion methods are survey data obtained by Clemence and Handler (2001) from 382 directors of psychology internship training programs. Among training directors from child care facilities, community mental health centers, and private psychiatric and general hospitals, from 59% to 78% preferred their interns to have had some coursework in sentence completion methods prior to entering their program, or at least a good working knowledge of these methods.

### **Rotter Incomplete Sentences Blank (RISB)**

The history of the RISB begins with Julian Rotter, a clinical psychologist and prominent social learning theorist who spent most of his career as a faculty member and clinical training director at the Ohio State University (1946–1963) and at the University of Connecticut (1963–1987). Working in an army convalescent hospital in 1945, Rotter saw a need for a brief screening measure that could be administered to large groups of people, scored reliably to provide a quantitative index of adjustment level, and used to evaluate fitness to

return to duty. He constructed a sentence completion task with items from various other sentence completion measures that had been used by army psychologists during World War II. This effort resulted in a 40-item Incomplete Sentences Test that Rotter, working in collaboration with Benjamin Willerman, administered to 200 army hospital patients. Rotter and Willerman (1947) subsequently published a description of this new measure, together with guidelines for administering and coding it and preliminary evidence of its reliability and validity.

Shortly thereafter, Rotter collaborated with Janet Rafferty to modify the army test for use with college students. This was done by retaining 32 item stems from the Incomplete Sentences Test used with the army personnel and revising the other eight stems to increase their suitability for a college population (e.g., "OVERSEAS" and "COMBAT" were replaced with "IN HIGH SCHOOL" and "BOYS"). The revised measure was renamed as the RISB and described by Rotter and his colleagues as a means of screening

for maladjustment and obtaining diagnostic information for use in treatment planning (Rotter & Rafferty, 1950; Rotter, Rafferty, & Schachtitz, 1949). As noted, adult and high school forms for the RISB were constructed soon after the college form and published simultaneously with it. Construction of the RISB adult and high school forms involved some minor changes in wording to make them appropriate for these age groups. For example, an item on the college form that reads, "IN HIGH SCHOOL" reads as "IN SCHOOL" on the adult form and as "IN THE LOWER GRADES" on the high school form.

Although Rotter and Rafferty (1950) described the RISB as useful in treatment planning as well as in determining adjustment level, they emphasized that the test is intended primarily for screening purposes and not for providing comprehensive assessments of personality characteristics or their "deep layers" (p. 6). Nevertheless, the minimal instructions and open-ended nature of its item stems create considerable potential for the test to reveal aspects of a person's inner life.

### **Washington University Sentence Completion Test (WUSCT)**

The WUSCT was designed for research rather than applied purposes and was constructed to operationalize elements of a particular theoretical formulation, not to obtain general information about psychological adjustment or personality functioning. Its history begins with Jane Loevinger, who was educated first as a measurement psychologist at the University of Minnesota and later as a developmental psychologist at the University of California at Berkeley, where she received her PhD in 1944 and worked for a time as a research assistant to Erik Erikson. Like Erikson, Loevinger formulated a stage theory of development, which she based on conceptions of "ego level."

Ego level in Loevinger's formulation is a "master trait" that encompasses aspects of cognitive complexity, moral perspective, behavioral control, and interpersonal relatedness, and she viewed this trait as accounting for most of the ways in which people differ from each other. With respect to development, Loevinger described eight stages of maturation in ego formation, with each stage of personality growth characterized by distinctive ways of thinking, feeling, and interacting with others. Maturation in her theory consists of developmental advances in a person's characteristic manner of perceiving and responding to the social world (Loevinger, 1976).

#### **524 Performance-Based Measures**

Loevinger devoted much of a long career, which culminated in her appointment as Stuckenberg Professor of Human Values and Moral Development at Washington University in St. Louis, to examining the validity and implications of her theoretical formulations. Advancing her research agenda required an assessment instrument for identifying level of ego development, and she devised for this purpose a sentence completion measure that became known as the WUSCT. This instrument was first published in 1970, underwent some minor revisions over time, and appeared in its current form in 1996 (Hy & Loevinger, 1996; Loevinger, 1985, 1998; Loevinger & Wessler, 1970; Loevinger, Wessler, & Redmore, 1970). Loevinger's ego development stage theory has not become particularly well known or widely cited, compared with the more influential stage theories of Erikson, Piaget, and Kohlberg. Nevertheless, the WUSCT has gained respect as a useful measure of psychosocial maturity and, as elaborated by Westenberg, Hauser, and Cohn (2004), has been used in a substantial body of personality research.

### **ADMINISTRATION**

The RISB is printed on both sides of a 1-page form, as previously noted, and the following instructions appear at the top of the front side:

Complete these sentences to express *your real feelings*. Try to do every one. Be sure to make a complete sentence.

With these instructions in mind, examiners can introduce the test in the following way, varying their choice of words according to which if any measures have previously been administered during their evaluation:

Next we're going to do [a test] [another test] that you will work on by yourself. This test consists of a list of words or phrases, and you're asked to complete each of them to make a sentence that expresses your feelings. As you will see, these instructions are printed on the

front of the test form, and you should also notice that there are items on both the front and the back of the form.

Seated comfortably in a well-lighted room and with a flat writing surface in front of them, examinees should then be handed the RISB form and a pencil with an eraser. People who invest themselves in the RISB task, especially if they are struggling with uncertainty in their lives, often decide to change what they have written as they proceed through the items, and writing with a pencil and eraser rather than a pen facilitates their doing so.

In keeping with customary procedure in administering relatively unstructured performance-based personality assessment instruments, additional guidance to examinees should be kept to a minimum. Occasional factual questions can be answered directly, as in “Is there a time limit?” (“No”) or “Do I have to answer every one?” (“It will be helpful if you do them all”). Other requests for structure should for the most part be turned back to the person, as in “How long should the sentences be?” (“It’s up to you”) or “Should I put in people’s names?” (“However you like”). Depending on the pace at which people work and the amount of detail they include in their responses, administration of the RISB has **Sentence Completion Methods 525**

generally been observed to require from 10 to 40 minutes, with most people completing the task in 20 to 25 minutes (Lah, 1989b, 2001; Sherry et al., 2004).

As an alternative to the written form of the RISB, the sentence stems can be read aloud to people and their oral responses recorded by the examiner. Rotter et al. (1992) discourage proceeding this way, however, because taking oral rather than written responses can confound the obtained data by infusing the administration with an interpersonal component.

This consideration is worth keeping in mind, especially when personal encounter with the examiner is likely to evoke a guardedness that leads people to limit the length of their responses or sanitize their content. Relatively silent or inarticulate people who dislike talking, have difficulty expressing themselves orally, or do not feel comfortable interacting with an examiner may be able and willing to say more about themselves when they are writing sentence completions than when they are asked to vocalize them.

The following is an example of the potential advantage of a written, as opposed to an oral, RISB protocol. A 16-year-old high school boy being evaluated was pleasant, socially appropriate, and seemingly relaxed during an initial clinical interview, but he was also tight-lipped and had little that he wanted to say. His answers to usual kinds of interview questions were bland and psychologically uninformative, and his Rorschach responses and oral TAT stories were sparse and mundane, providing some indications of his personality characteristics but few clues to any underlying concerns. On the RISB, however, he wrote the following revealing completions:

I WANT TO KNOW why my parents and I don’t get along.

AT HOME my parents and I argue a lot.

WHAT ANNOYS ME most is when my parents bother me.

MY GREATEST FEAR is my parents not wanting me at the house.

WHEN I WAS YOUNGER I had a great relationship with my parents.

I NEED my parents to understand me.

I HATE when I get in fights with my parents.

I WISH that I had a better relationship with my parents.

I love my parents even though they think otherwise.

These sentence completions by a boy who could not unburden himself in direct conversation with a mental health professional, and who might otherwise have given the impression of being either untroubled or psychologically insensitive, provided compelling indications of perplexity and regrets in the face of dysfunctional family relationships and of deeply felt ambivalent feelings toward his parents.

Paralleling Rotter et al.’s (1992) recommendation for written rather than oral administration of the RISB, standard procedure for the WUCST also calls for a written format.

Westenberg, Hauser, et al. (2004) have noted in this regard, “The written procedure fits best with the purpose of the test: to reveal the respondent’s frame of reference without distortion that might arise from the presence of the administrator” (p. 596). Westenberg et al. express concern in particular that examinees’ oral interaction with the examiner can lead them to

respond in a socially desirable rather than a personally revealing manner.

Interestingly, research comparing oral and written administration has found no significant difference between them in the types of responses or ego-level scores they evoke, on either **526 Performance-Based Measures**

the WUSCT or the SCT-Y (McCammon, 1981; Westenberg, van Strien, & Drewes, 2001).

Opting for written rather than oral administration of these measures may not prevent an uninformative protocol from a guarded examinee. Nevertheless, an oral administration may forgo the opportunity to obtain an especially revealing written protocol from a reticent and interpersonally aversive person, as in the case of the adolescent boy whose RISB was just discussed.

In the administration of the WUSCT, the instructions printed on the test form merely state, "Complete the following sentences." Hy and Loevinger (1996, p. 26) suggest accompanying distribution of the test with the following introduction:

Now I would like you to fill out this sentence completion form. You see that these are incomplete sentences. Please finish each one. Notice that there are two pages; please make sure that you have finished both.

As in administering the RISB, examiners giving the WUSCT should respond to requests for additional guidance by indicating in a noncommittal way that there are no right or wrong answers and that people can write out their sentences however they wish. In addition, Hy and Loevinger (1996) stress that an examiner should always be present to monitor both individual and group test administrations. To maintain the security of test measures and obtain dependable data, this recommendation holds for all personality and mental tests that can be self-administered. Allowing people to take test forms out of the examining room and complete them without supervision opens the door to indiscriminant dissemination of their item content and to responses based on unknown sources of influence (as in a woman asking her husband, "How do you think I should answer this one?" or a man consulting a personality assessment textbook for clues to creating some impression). Firsthand monitoring of test administration to safeguard the integrity of test results is widely endorsed among assessment psychologists and is reflected in Standards 5.6 and 5.7 of the *Standards for Educational and Psychological Testing* developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999).

## **SCORING**

In keeping with his original interest in developing an objective quantitative index of adjustment, Rotter devised a scoring system for rating each of the RISB items on a 7-point scale from 0 (most positive adjustment) to 6 (most indication of conflict). Introduced in preliminary form in the original Rotter and Willerman (1947) publication, this scoring system remains in place and is elaborated with detailed guidelines and specific examples in the most recent RISB manual (Rotter et al., 1992).

Positive adjustment is scored from 0 to 2, depending on the extent to which sentence completions express such adaptive perspectives as an optimistic outlook (e.g., "THE FUTURE looks bright for me"), receptive attitudes toward people (e.g., "THE HAPPIEST TIME is when I'm with my friends"), and an upbeat affective state (e.g., "DANCING is something I enjoy doing"). Conflict is scored from 4 to 6, depending on the intensity with which negative attitudes, interpersonal strains, and various kinds of symptoms or concerns are expressed (e.g., "WHAT ANNOYS ME is how stupid people can be sometimes");

### **Sentence Completion Methods 527**

"I REGRET being by myself so much of the time"; "OTHER PEOPLE don't appreciate me"; "I CAN'T stand heights"; "MY NERVES are frazzled"). Neutral responses that have no personal reference or consist only of descriptions or catch phrases are given a score of 3 (e.g., "AT BEDTIME I watch TV and fall asleep"; "MEN are different from women"; "THIS PLACE is okay").

After adjustment/conflict ratings have been assigned for the 40 responses, they are totaled to provide an overall adjustment index, which can range from 0 to 240. Should any items be left blank, the total score is prorated to maintain comparability with scores for a full 40-response protocol. This scoring procedure generally requires 15 to 35 minutes,

depending on the examiner's familiarity with the scoring criteria (Sherry et al., 2004), and it makes the RISB unusual among personality assessment instruments in paying as much attention to indications of good adjustment as to evidence of psychological difficulties. Despite this advantage of measuring both how well and how poorly people seem to be feeling and functioning, the RISB scoring system is seldom used in clinical practice. In the previously mentioned survey by Holaday et al. (2000), most of the clinicians who reported using sentence completion measures indicated that they rarely if ever score these measures as prescribed in the test manual.

Like the RISB, the WUSCT is scored in two phases, first at the item level and then for determination of an overall score, which is called the Total Protocol Rating (TPR). Each of the 36 sentence completions (or 32 on the SCT-Y) is given an independent rating from E2 to E9 according to which of Loevinger's eight ego levels it appears to reflect. There is no E1 in Loevinger's system because it would refer to a stage in early infancy during which ego development is not accessible "by our methods of study" (Hy & Loevinger, 1996, p. 4). There are four alternative procedures for determining the TPR, three of which are straightforward: summing the item scores that range from 2 to 9 for each item, calculating the mean value of these item scores, or taking the mode of the item ratings as representative of the ego level at which the person most frequently operates. Hy and Loevinger (1996, chap. 5) express reservations about all three of these approaches because none of them takes account of outliers in the distribution of item scores. Outliers in this context are extremely high or low item scores that distort the meaning of a person's total or average score and thereby generate an inaccurate or misleading indication of the developmental level at which the person generally functions.

To circumvent this problem, Hy and Loevinger (1996) developed a set of rules for considering both item scores and the distribution of these scores in determining the developmental level that a person has achieved. These rules for determining the TPR, which can be found in Table 5.1 on page 39 in their text, are not difficult to implement. However, the criteria for assigning developmental level scores to individual item responses are complex. Hy and Loevinger (1996) provide a detailed set of scoring guidelines, case examples, and practice exercises that examiners will find helpful in mastering this scoring system. In their opinion, becoming proficient at scoring the WUSCT correctly requires a minimum of 2 hours daily work with the practice exercises for a period of 2 to 3 weeks.

## **INTERPRETATION**

In common with the TAT and the figure-drawing methods discussed in Chapters 12 and 13, sentence completion methods can be objectively scored, and they provide structural and

### **528 Performance-Based Measures**

behavioral data that enrich personality evaluations. As is true of the TAT and figure-drawing methods, however, sentence completion methods are rarely scored in clinical practice, and the richest and most useful information they provide usually comes from inspection of their content. Accordingly, this section of the chapter first presents some guidelines for interpreting the thematic imagery in sentence completion responses and then considers the interpretive significance of the RISB and WUSCT total scores and of certain structural features of the data in them.

### **Content Features**

The inspection method for interpreting the content of sentence completion responses proceeds according to the same principles delineated in Chapters 11, 12, and 13 for interpreting thematic imagery. The completions should be read and pondered for their possible personal meanings, and the hypotheses that are generated in this way should then be evaluated for how compelling and plausible they are in the individual case.

In generating alternative hypotheses, examiners should rely on their understanding of personality development and dynamics, their familiarity with how different types of persons commonly respond on sentence completion measures, and their own associations to what certain responses might signify. As testimony to the particular value of the RISB and WUSCT, this phase of the interpretive process is generally less challenging than the interpretation of Rorschach and TAT imagery, because it requires a lesser level of inference and produces fewer alternative possibilities.

To reprise illustrations from Chapters 11 and 12, a Card I Rorschach response of “A woman with her arms up” could reflect persons’ attitudes toward or perceptions of their mother, their wife, or women in general, and having her “arms up” could represent waving hello, waving goodbye, preparing to deliver a blow, or reacting to something with fear or surprise. The man leaning against a lamppost in Card 20 of the TAT who “feels good about himself, knows what he wants to do, and thinks he can do it” could reflect a positive self-image or some wishful thinking about how the person would like things to be when they are actually quite otherwise.

Unlike the Rorschach task, which calls for reporting what inkblots “might be,” and the TAT, in which people are instructed to “use your imagination,” the RISB asks people to “express your real feelings,” and its stimulus stems include such specifications as “A MOTHER ” and “MOST WOMEN .” Hence, the possible interpretive significance of sentence completion responses is often fairly obvious and relatively unambiguous. A 51-year-old man who wrote “A MOTHER is someone very special” and “MY FATHER was a good man” could with reasonable confidence be inferred to have enjoyed good parent-child relationships, unless there were good reasons to question his veracity. To illustrate a somewhat less direct but still compelling interpretive hypothesis, the previously mentioned 16-year-old boy troubled by dysfunctional family relationships wrote, “MY FATHER travels a lot.” This statement strongly suggested that growing up with a frequently absent father had contributed to adjustment difficulties he was having, which on further investigation turned out to be the case.

Whether sentence completion interpretations are compelling depends in part on how clear they are and how easy it is to recognize their basis. Given the clarity and readily apparent basis of the illustrative interpretations in the previous paragraph, and in view of how little **Sentence Completion Methods 529**

inference is required to arrive at them, they could be considered compelling. Along with being clear, sentence completion interpretations become compelling to the extent that the evidence for them is repetitive or dramatic. The more frequently an attitude or perspective is suggested by the completions in a protocol, the more likely it is to characterize the individual. The previously quoted 51-year-old man who expressed positive feelings toward his parents also gave compelling repetitive evidence of pleasure in family living, with the following responses: “I LIKE being a father and a family man”; “THE BEST thing that every happened to me is my family”; “MARRIAGE is very important to me”).

Significantly, however, this same contented family man, who had a 3-year history of persistent depression, also gave repetitive indications of feeling (a) that his better days were in other respects behind him (“I REGRET the closure of my business”; “MY GREATEST FEAR is not gaining control of my life again”); (b) that he had not become the person he had hoped to be (“I SECRETLY wish I was as good as my dad”; “I FAILED to be the best I could be”); and (c) that he had little to look forward to (“I HATE growing old”; “I AM VERY worn out”).

As for drama, a single sentence completion response can at times have compelling interpretive significance by virtue of the intensity with which it is expressed. To illustrate this possibility, consider first the response “IN SCHOOL I didn’t do very well.” This is a mildly stated and largely ambiguous sentence completion. It suggests some adjustment difficulty that could be important to investigate further, but by itself it gives little indication of just how badly the person did in school and no indication of whether the person’s difficulties were primarily academic or primarily social, or possibly athletic or artistic, in a school in which these talents were highly valued.

By contrast, suppose a person gives an RISB protocol that is generally bland and uninformative with respect to interpersonal relationships (e.g., “PEOPLE come in all shapes and sizes”; “MARRIAGE is something you need a license for”), but contains the response, “IN SCHOOL things were rotten and I hated all of the kids—I couldn’t stand them and they couldn’t stand me.” Such a dramatically intense response is sufficient by itself to warrant hypothesizing that this person experienced serious social problems as a student and may even have been developing an avoidant or antisocial personality disorder. Having this particular dramatic breakthrough in the context of guardedness suggested by an otherwise

bland protocol makes its interpretive implications even more compelling. Whether sentence completion interpretations are plausible as well as compelling in the individual case goes beyond their being clear, repetitive, or dramatic and depends on their confirmation by external data sources. Three types of external data in particular can help establish the plausibility of these interpretations: (1) findings from other tests suggesting personality characteristics or life experiences similar to those that have been hypothesized on the basis of sentence completion responses, (2) interview or case history information that is consistent with inferences drawn from a person's sentence completions, and (3) collateral reports or direct observations of a person that match the implications of the sentence completion data for certain behavior patterns. The more these external data sources confirm sentence completion interpretations, the more confident examiners can be that these interpretations accurately describe an individual's personality characteristics. At times, on the other hand, apparent divergence among data sources may also provide valuable information. In one case, a woman gave the sentence completion response, "I FEEL very good about my life," which would seem to indicate that she was generally

### **530 Performance-Based Measures**

in good spirits. According to family members, however, she had been struggling with episodes of moodiness and trying hard to "put on a happy face." This collateral evidence suggested that "I FEEL very good about myself" represented a defensive effort to ward off depression and that this woman's reliance on denial might make her susceptible to such hypomanic characteristics as unwarranted optimism or exuberance. Instances of this kind illustrate how sentence completion responses can generate interpretive hypotheses with important treatment implications when they diverge from collateral reports as well as when they closely parallel them. The potential information value of divergent findings, and their implications for conducting psychological examinations with a multifaceted test battery, are elaborated in Chapter 2 (see pp. 24–26).

### **Total Scores**

As indicated in describing the scoring of RISB and WUSCT protocols, both tests generate a total score. The RISB total score, ranging from 0 to 240, is intended to serve as an adjustment index. Based on the standardization data they obtained from college students, Rotter and Rafferty (1950) originally set 135 as the score for distinguishing people who should be categorized as "adjusted" (total score less than 135) from those who should be categorized as "maladjusted" (total score of 135 or more). This same cutting score was subsequently used for administrations of the adult and high school forms of the test as well, even though no additional normative data were obtained from adult or adolescent samples. In three new samples of college students subsequently tested in 1965, 1970, and 1975, Lah and Rotter (1981) found that these groups did not differ significantly among themselves in their mean RISB total scores but showed significantly higher mean scores than the 1950 standardization sample. On the basis of two further studies, Lah (1989a) subsequently recommended increasing the adjusted/maladjusted cutting score from 135 to 145. In the first of these studies, he obtained mean RISB scores of 128.0 and 130.8 from groups of 64 female and 52 male college students, respectively.

In the second study, he examined the impact of alternative cutting scores ranging from 125 to 160 among 120 students being seen in the college campus mental health service and 120 control participants. An unacceptably high frequency (35%) of the control participants had scores of 135 or more, and 145 proved to be the most efficient cutting score for discriminating between the control participants (of whom 82.4% scored below 145) and the mental health service group (85.0% of whom received scores of 145 or more). In a similar vein, Ames and Riggio (1995) found a 40% frequency of maladjustment in a sample of 136 college students, as defined by the original RISB total score of 135 or more, and a 55% frequency of scores above 135 among 368 high school students.

These findings demonstrated that the original cutting score for adjustment problems was no longer appropriate for high school or college students, and Rotter et al. (1992) increased this cutting score to 145 in their second edition of the RISB. Like all cutting scores used in clinical practice, 145 should be regarded as the midpoint in a range, and not as an absolute index for inferring maladjustment or rendering opinions about a person's

need for psychotherapy or fitness to return to work. In assessing psychopathology, there will inevitably be some false positive findings at scores above 145 and some false negative results at scores below 145. A score somewhat above 145 may not necessarily warrant referral for mental health services, depending on other considerations; and a score somewhat below 145

#### **Sentence Completion Methods 531**

may not reflect sufficient peace of mind and emotional resiliency for a person to function effectively in a high-stress situation. Additionally, the Ames and Riggio (1995) finding of higher scores among high school than college students argues for particular caution in inferring maladjustment from RISB total scores on the High School form.

Turning to the WUSCT, the TPR on this measure serves to classify people as having on the average attained a particular developmental level of ego functioning. How people are classified on the basis of the WUSCT does not mean that they always and in every respect—cognitively, emotionally, and interpersonally—behave at their identified level of psychosocial maturity. It means only that they are likely to conduct themselves at this level most of the time and in most respects. The following descriptions of the eight levels of ego development in Loewinger's model, as elaborated by Hy and Loewinger (1996, chap. 1) and Westenberg, Hauser, et al. (2004), illustrate the types of personality characteristics commonly associated with each level.

#### *Impulsive Stage (E2)*

Individuals at the impulsive stage of development are self-centered, pleasure-seeking persons with propensities for impulsive aggressive and sexual behavior. They expect other people to cater to their needs and desires, and they have little tolerance for frustration. They tend to be oppositional and defiant and to regard other people as either all good or all bad, depending on whether these other people are meeting their needs. They typically see rules as arbitrary or unfair, and only external constraints may suffice to keep their impulses in check. This stage of development typifies young children, and persons who have not advanced beyond this stage by the time they enter school are likely to be diagnosed as having an impulse disorder.

#### *Self-Protective Stage (E3)*

The self-protective stage is characterized by a focus on being in control and finding ways to maximize one's own pleasure. Self-protective persons can control their impulses when they sense that doing so will benefit them, and they also begin to appreciate and follow rules if abiding by those rules will be to their advantage. Interpersonally, they tend to mistrust other people and to manipulate and exploit them to their own ends. They are creatures of the moment, lacking long-term goals and ideals. If their actions get them into trouble, they externalize blame by holding other people or unavoidable circumstances responsible. Older children and adults who remain at this stage of development are likely to become hostile, opportunistic individuals who often show psychopathic personality features.

#### *Conformist Stage (E4)*

This level of ego development involves a transition from the earlier egocentric selfprotective stage to a group-centered conformist stage. Conformist individuals endorse social norms and are committed to gaining social approval. They pay close attention to what other people think and expect, they believe that the same rules and requirements should apply to everyone, and they regard individual differences in attitudes and codes of conduct as inappropriate and undesirable. People at this stage of development tend to describe interpersonal interactions mainly in terms of what people do, not how they feel, and to describe themselves and people close to them by referring to their socially acceptable behavior.

#### **532 Performance-Based Measures**

#### *Self-Aware Stage (E-5)*

In the self-aware stage of ego development, people begin to recognize being different from other people and having their own private feelings, ideas, and opinions. This increasing selfawareness and examination of one's distinctive inner life promotes a growing appreciation of differences among people. Individuals at this stage begin to distinguish between public and private aspects of themselves and to appreciate the importance of being true to one's self. People who mature to this level also begin to recognize that morality is not absolute

and that exceptions to the rules may at times be acceptable.

#### *Conscientious Stage (E-6)*

The conscientious stage is marked by pursuit of self-evaluated standards and preoccupation with aspirations, accomplishments, ideals, and morality. Individuals who reach this level of psychosocial maturity tend to be self-critical and self-motivated; they are intent on defining their own goals and working toward these goals in their own way. Concurrently, they also become increasingly likely to take responsibility for their actions and for whatever choices they make in their lives.

#### *Individualistic Stage (E7)*

At the individualistic stage of development, people can be expected to show good comprehension of psychological development and causation and a clear sense of personal identity. They are likely to have gained an appreciation of the complexity of interpersonal relationships and of the multiple roles that people play in their lives (e.g., as a daughter, mother, spouse, and professional). No longer trying to change themselves and other people to fit an ideal image (which is a pursuit characteristic of the conscientious stage), people at the individualistic level recognize and take note of contradictory emotions, motivations, and inner conflicts in themselves and others. Whereas relationships with people have ordinarily become deeper and more intense with maturation from the conformist to the conscientious stage, interpersonal attachments are likely to be viewed at the individualistic stage as a potential obstacle to achievement striving and a source of excessive responsibility for others.

#### *Autonomous Stage (E-8)*

Should they reach the autonomous stage of development, people tend to become less concerned than before with issues of achievement and morality and less preoccupied with evaluating their own actions and the actions of other people. Instead, this stage is characterized by belief in the need for people to find their own way, wherever their roads may take them, and to make their own mistakes and reap their own rewards. Individuals at this stage may also display a bemused appreciation of life's paradoxes, contradictions, and inconsistencies, without a sense of resignation to them.

#### *Integrated Stage (E-9)*

The integrated stage of ego development as conceived by Loevinger is more of an ideal to which people should aspire than a reality they are apt to attain. She likens this stage to Maslow's (1954) concept of self-actualizing people, defined as those rare individuals who have fully realized their human potential. Loevinger estimates that fewer than 1% of adults

#### **Sentence Completion Methods 533**

reach this developmental milestone. Because of the infrequency with which persons at the integrated stage have been encountered in their research samples, Hy and Loevinger (1996) conclude that available data "do not suffice to describe fully this theoretical high point" (Hy & Loevinger, 1996, p. 6). For practical purposes in classifying people on the basis of their WUCST scores, Loevinger accordingly recommends collapsing the integrated stage of development with the autonomous stage.

For examiners who administer the WUSCT and determine its TPR, these descriptions provide some information about features of personality that are likely to characterize individuals who fall clearly into one of these eight ego stage categories. Although indicating a person's level of psychosocial maturity, the WUSCT PTR does not serve as an adjustment index. Unlike the RISB total score, which provides a numerical index along a continuum from relatively poor to relatively good adjustment, the PTR level is largely independent of adjustment. Persons at higher stages of ego development who are acutely self-aware may, as a consequence of inner conflicts and uncertainty, be more upset and less psychologically comfortable than persons at lower stages who rarely examine their own behavior or entertain self-doubts. In younger people, limited psychosocial maturity may be more adaptive than a high level of maturity that is unusual among persons of their age. Whether a person functions and feels better as a conformist or an individualist is likely to vary with what is expected of them in their particular life circumstances and by the people with whom they interact.

These and other considerations in inferring adjustment level from stage of ego development are elaborated by Hy and Loevinger (1996), who conclude that "ego maturity and

adjustment must be described independently in order to ascertain the relation between them” (p. 7).

### **Structural Features**

The interpretive significance of reaction time, word count, and other structural features of sentence completions has seldom received much attention from assessment psychologists, either in clinical practice or in research. Nevertheless, some observations by Kleiger (2004) are notable for describing sentence completion characteristics that can help identify disordered thinking. Drawing on guidelines proposed by Rapaport et al. (1946/1968) for inferring thought disorder from word associations, Kleiger described how sentence completions can sometimes point to thinking disorder by being expressed in peculiar ways or showing too little or too much distance from the content of the item stems.

Exemplifying peculiar expressions, the sentence completions of persons whose thinking is disordered are sometimes marked by strange word usage, as in “I NEED more physicalness for sexuality.” At other times, deviant verbalization may take the form of a rambling and illogical discourse on some matter of concern, as in “MY GREATEST WORRY is about the glaciers melting, which is going to cause a lot of war and disease, especially if people don’t eat better.” As Kleiger noted, the more incoherent, disorganized, or irrational people are in giving sentence completion responses, the more likely it is that their thought processes are similarly muddled, dissociated, and illogical.

As for their distance from the item stems, sentence completions with too little distance take the form of concrete responses that stick closely to the content or sound of the stems at the expense of appropriate attention to their meaning. Such loss of distance can result in responses that simply repeat all or part of a stem (e.g., “A MOTHER is a mother”;

#### **534 Performance-Based Measures**

“I REGRET what I regret”) or consist of a clang association to it, as in “PEOPLE steeple.” Sentence completions involving too much distance take item stems as a springboard for nonsensical associations that have little or no apparent connection to the stem content. Such disconnected association was illustrated by a man who wrote, “THE ONLY TROUBLE is fear itself.” Perhaps this man was associating “trouble” with the Great Depression of the 1930s and connecting it with Franklin Delano Roosevelt’s exhortation at the time: “The only thing we have to fear is fear itself.” The omission of such explanatory links makes overly distant completions difficult to follow and a likely manifestation of disordered thinking.

### **APPLICATIONS**

On the one hand, sentence completion methods bring to a personality assessment battery an easily administered performance-based test that can be given to groups as well as individuals and can be completed by most people in 30 minutes or less. Barring a language or reading difficulty, people taking the test usually have no difficulty grasping the meaning of sentence completion stems, and they are not ordinarily asked by the examiner to explain, elaborate, or account for their responses. For this reason, sentence completion methods tend to be less anxiety provoking than the inkblot, picture-story, and figure-drawing methods discussed in Chapters 11 through 13. As a written measure involving little interaction with the examiner, moreover, sentence completions are less likely than oral measures to cause distress to people who are by nature untalkative or uncomfortable in social situations. Like the adolescent boy first mentioned on page 525, such people may be less guarded and more revealing about themselves in what they write than in what they say.

On the other hand, the relative lack of ambiguity in many sentence completion stems means that people are more aware of how they are presenting themselves than they are when describing inkblots, pictures, or drawings they have made. The clarity and directness of self-report of stems like “I FEEL,” along with the obvious implications of completing such a stem with “happy” or “sad,” make it easy for people to paint a deceptively positive picture of themselves or to exaggerate the extent of their distress or difficulties. Responding in a deceptive mode, a person whose life has been a showcase of failures to cope adequately might write, “I FEEL that I have managed my life very well,” and a demonstrably wellfunctioning and psychologically stable person intent on malingering disorder might write, “I FEEL that I’m falling apart.”

As reviewed by Schretlen (1997), there is some research evidence that presumed indices

of malingering on sentence completion tests correlate modestly with MMPI validity scales. Additionally, some success has been reported in using features of sentence completion responses to identify misrepresentation among persons applying for disability status and defensiveness in a selection program for special assignments in the military (Lanyon, Almer, & Maxwell, 2002; Picano, Roland, Rollins, & Williams, 2002). However, no formal coding system for assessing impression management with sentence completion measures has been developed sufficiently to warrant its general use in clinical practice.

### **RISB**

Turning now to specific measures, the RISB was developed to serve as a brief screening measure of adjustment, primarily for use in identifying persons in need of mental health

#### **Sentence Completion Methods 535**

intervention, and it serves this purpose well. With an RISB cutting score of 145, Lah (1989a) could distinguish with above 80% accuracy between college students receiving campus mental health services and a nonpatient comparison group. In a separate sample of 116 college students, Lah (1989a) additionally found significant correlations between the RISB scores and sociometric ratings of several aspects of adjustment, including general happiness, sense of humor about oneself, and self-acceptance. In studies comparing delinquent adolescent boys in residential treatment with randomly selected high school boys, Fuller, Parmelee, and Carroll (1982) found significant differences between the groups in their RISB total score and developed a maladjustment score that correctly identified 80% of the delinquents and 75% of the high school boys.

Along with helping to identify persons with adjustment difficulties for which treatment may be indicated, sentence completion data can sometimes suggest a specific type of disorder. With respect to psychopathic personality features, Endres (2004) reported moderate positive correlations between scores on the Psychopathy Checklist-Revised (PCL-R) and certain cognitive and linguistic characteristics of sentence completions, including use of coarse language and a prominent focus on exerting power. Although not unique to psychopathy, these sentence completion characteristics were found by Endres (2004) to be highly sensitive to it.

In addition to suggesting patterns of possible personality disorder, sentence completions can give direct expression to manifestations of symptomatic disorders. The illustration of “I FEEL depressed” at the beginning of this chapter is a case in point. Similar examples would be “I FEEL afraid of heights” as an indicator of phobic disorder; “AT BEDTIME worrying about having a nightmare keeps me awake” as a clue to posttraumatic stress disorder; and “SOMETIMES my moods go from happy to sad all at once” as suggestive of bipolar or cyclothymic disorder.

Despite its potential for serving these diagnostic purposes, the RISB is rarely relied on for them. Most of the other commonly used assessment measures discussed in this *Handbook* provide more dependable indices of overall adjustment and differential diagnosis than the RISB. With respect to screening for adjustment difficulties, the previously noted infrequency with which practitioners calculate scores for sentence completion tests indicates how seldom the RISB is used in this way. As for differential diagnosis, there are no systematic guidelines in the literature for discriminating among types of disorder on the basis of sentence completions. It is not in these respects that RISB data prove particularly useful, then, but rather in providing clues to persons’ underlying conflicts and concerns, their feelings about themselves and other people, and their attitudes toward various life situations and experiences. Information of this kind creates valuable applications of the RISB in clinical practice, forensic consultation, and personality research.

In clinical practice, RISB clues to a troubled person’s conflicts, concerns, feelings, and attitudes can contribute to planning psychotherapy and evaluating treatment progress and outcome. In the planning phase, sentence completion reflections of a person’s inner life help to identify issues to address in the course of therapy. The extent to which subsequent retesting indicates that these issues have been resolved helps to monitor progress in treatment and determine whether therapy has proved beneficial.

In the case of the adolescent boy who revealed more of himself on the RISB than on other tests in the battery, his initial sentence completions identified as treatment targets

some resolution of his ambivalent feelings toward his parents and an easing of his conflicts with them. Suppose on retesting during the course of therapy he were to write, instead

#### **536 Performance-Based Measures**

of “AT HOME my parents and I argue a lot,” “AT HOME things are going pretty well.” Suppose as other notable changes in his sentence completions he were to write, instead of “I NEED my parents to understand me,” “I NEED to find a girlfriend,” and instead of “I WISH that I had a better relationship with my parents,” “I WISH that I could get my grades up.” Then his sentence completions would strongly suggest positive change in treatment, with reduced family conflict and increased attention to normative age-related concerns.

In forensic consultations, the utility strength of the RISB in reflecting aspects of a person’s inner life has only limited applications. The nature of a person’s underlying conflicts, concerns, feelings, and attitudes has little bearing on the psycholegal issues with which courts are concerned in criminal and personal injury cases. Moreover, the limited availability of research findings that validate thematic interpretations of sentence completion responses may in some courtrooms prevent RISB-based testimony from being accepted into evidence. Still, as suggested by the previously mentioned frequency with which forensic psychologists have reported using sentence completion methods in custody cases (29% in evaluations of children, 22% in evaluations of parents), RISB data can at times be put to effective use in family law cases.

In particular, information concerning how parents feel about their children and about the responsibilities of parenthood is often sought and valued by family court judges. How children feel about their parents and about their home life, together with indications of family conflict, are also commonly of interest in family law hearings and mediations. As illustrated in this chapter, RISB responses can provide clues to such feelings and attitudes. There may also be instances in which sentence completions have face validity that the court might consider relevant. The adolescent boy who wrote “MY FATHER travels a lot” also gave the sentence completion, “A MOTHER helps guide their children in life.” If this boy’s parents were contesting custody or visitation rights, these contrasting responses might speak loudly to the court about who was the more important and available parent in his life.

In research applications, the RISB has proved useful not only in identifying personality characteristics, but also in providing an overall quantitative index of adjustment level. Whereas most personality assessment instruments measure adequate adjustment by the absence of indicators of maladjustment, the RISB item scores range from 0 (most positive) to 6 (most conflict). Hence the RISB total score provides a bipolar measurement with normality in the middle, superior adjustment at the low end, and severe psychopathology at the high end of the scale. This sensitivity of the RISB to both particularly good and markedly poor adjustment facilitates selection of adjusted and maladjusted samples for research purposes. Early research using the RISB for this purpose is summarized by Lah (1989a, 1989b), and interesting results have also been reported in more recent investigations in which the RISB total score was used as a criterion measure of adjustment. In a study of children of divorce, for example, Kouser and Najam (1992) used the RISB to demonstrate more maladjustment among adolescent girls whose parents were divorced than among girls from intact families.

#### **WUSCT**

The WUSCT was developed primarily for research purposes to measure level of ego development, and it has served this purpose often and well. Westenberg, Hauser, et al.

#### **Sentence Completion Methods 537**

(2004) report that more than 300 empirical studies employing the WUCST had been published by the early 1990s. As an example of its research utility, the WUSCT has proved helpful in identifying developmental aspects of the personality dimensions measured by the NEO PI-R, which is discussed in Chapter 10, especially in relation to cognitive and interpersonal levels of functioning (Hogansen & Lanning, 2001; Kurtz & Tiegreen, 2005). In other research, the WUCS has been used as a measure of psychosocial maturity in studies examining emotional complexity (Kang & Shaver, 2004), mothers’ feelings about their maternal role (Luthar, Doyle, Suchman, & Mayes, 2001), attitudes toward athletics

and academic achievement in high school students (Bursik & Martin, 2006; Takenouchi, Taguchi, & Okuda, 2004), developmental changes in childhood fears and vulnerability to social anxiety disorder (Westenberg, Drewes, Goedhart, Siebelink, & Treffors, 2004), gender and cross-cultural differences in levels of ego development (Bakken & Huber, 2005; Truluck & Courtenay, 2002), and self-regard in adolescent psychiatric inpatients (Evans, Brody, & Noam, 2001).

Research applications aside, the previously noted infrequent use of the WUSCT in clinical practice seems regrettable. In its format and content, the WUSCT is just as suitable as the RISB for generating thematic data that provide clues to a person's inner life. Moreover, placement at one of eight levels of psychosocial maturity coded with the WUSCT has implications for differential diagnosis and treatment planning. As noted, the level of ego development measured by the WUSCT is not necessarily related to the presence or absence of psychopathology (see Hy & Loevinger, 1996, chap. 1). Ego level, however, does have implications for the type of disorder people are likely to develop if they become disturbed. As elaborated by Westenberg, Hauser, et al. (2004), there is evidence to suggest that behavioral and externalizing problems occur mostly among persons who are functioning at the Impulsive and Self-Protective levels of ego development and less so among persons who attain the Conformist and higher levels of psychosocial maturity. Emotional and internalizing problems like anxiety and depression can characterize people at all ego levels but tend to become most noticeable among persons functioning at the Conformist or higher levels of ego development.

In treatment planning, WUCST findings can assist in selecting a treatment approach appropriate to a patient's needs and capacities. Individuals at lower levels of ego development may be relatively likely to think about psychotherapy in concrete terms as a service that their therapist will provide for them and be responsible for making it helpful. Persons at higher ego levels, by contrast, may be more likely than psychosocially immature individuals to regard psychotherapy as a participatory process of learning about themselves in which they will bear primary responsibility for progress.

Stakert and Bursik (2006) reported some relevant findings in a study of 100 adult patients with serious psychological disturbance who were being treated in a community mental health center. Those patients who showed relatively high levels of ego development as measured by the WUSCT had more complex therapeutic goals and were more committed to those goals than patients with lower levels of ego development. Patients functioning at higher ego levels were also more likely than those at lower levels to include improved social relationships and increased self-understanding among their treatment goals, whereas a treatment focus limited to symptom relief and rehabilitation was more prevalent among patients at lower ego levels. To the extent that these distinctions apply in the individual case, the WUSCT Total Protocol Rating of persons seeking psychotherapy is likely to

**538 Performance-Based Measures**  
have implications for whether their treatment needs will be better met by a supportive and directive approach or by an expressive and relatively unstructured form of psychotherapy.

### **PSYCHOMETRIC FOUNDATIONS**

In their psychometric foundations, as in their frequency of use for clinical purposes, the RISB and WUCST stand in sharp contrast. The RISB has been used frequently in research studies as a criterion measure of adjustment, as noted in discussing its applications, but its psychometric properties have rarely been examined directly. The WUSCT, despite its infrequent clinical use and in addition to its prominent research role as a measure of ego development, has been studied extensively to determine its reliability and validity. Both the overall adjustment score on the RISB and the total score on the WUSCT have shown good psychometric properties, although the evidence in this regard is much stronger for the WUCST than for the RISB. Loevinger's categorization of levels of ego development has also shown predictive and construct validity, but the clues to aspects of a person's inner life that can be inferred from the thematic information in sentence completions have not been the subject of any systematic validity studies.

### **RISB**

Although limited, the available data have consistently demonstrated solid reliability for the

RISB total score. With respect to interrater reliability, Rotter and Rafferty (1950) reported in their original test manual that examiners had achieved correlations for agreement of  $r = .96$  in coding protocols of females and  $r = .91$  for males. Lah (1989a) in his two previously mentioned research samples obtained coefficients for interrater agreement of  $r = .94$  and  $r = .90$ . As summarized by Sherry et al. (2004), other studies over the years have shown similarly high levels of scoring reliability, with observed correlations for agreement ranging from .72 to .99 and falling mostly in the .90s. Split-half reliabilities for the RISB for its different forms have ranged from .83 to .86 for females and from .74 to .84 for males in various studies, and retesting over 1- to 2-week intervals has yielded reliability coefficients averaging in the low .80s (Sherry et al., 2004).

As for the validity of the RISB, the studies mentioned in discussing clinical applications of this test bear witness to the validity of its total score for identifying maladjustment, and the many research studies in which it has been used effectively as a criterion measure of adjustment level provide further indirect evidence of its validity. Regrettably, the validity of the RISB for drawing inferences either about observable personality characteristics or underlying thoughts and feelings has seldom been examined, and there is no body of accumulated empirical evidence in this area. Sherry et al. (2004) wrote in their 2004 review of the literature that they had not found any validity studies of the RISB published since Lah's (1989a) report.

In the absence of this validation, examiners need always to treat inferences drawn from the thematic content of sentence completions as clues to possible features of an individual's personality, and not as certain fact. Even so, as elaborated in this and the three preceding chapters, thematically based hypotheses can prove useful in clinical practice and merit careful consideration. As the distinguished psychologist Paul Meehl (1945) once wrote **Sentence Completion Methods 539** about personality assessment, "The final test of the adequacy of any technique is its utility in clinical work" (p. 302).

## **WUSCT**

As reviewed by Westenberg, Hauser, et al. (2004) and Westen, Feit, and Zittel (1999), the WUSCT has shown excellent reliability in numerous studies. Obtained correlations for interrater agreement in scoring responses for their psychosocial maturity level ranges from  $r = .80$  to  $r = .90$ , perfect agreement on item scoring averages about 85%, and disagreement by more than one level in categorizing stage of development is often less than 10%. The split-half reliability of the Total Protocol Rating on WUSCT and the SCT-Y is about .80, and most studies report a Cronbach alpha for internal consistency of .90 or higher.

An impressive body of research has also validated both Loevinger's theoretical formulations and her assessment method. With respect to her stage theory of ego development, empirical data have confirmed (a) that ego development exists on a continuum of stages; (b) that individuals pass through an invariant sequence of these stages as they mature, from the initial E2 (Impulsivity) up to the highest level of maturity they can attain; and (c) that ego level of development is a personality typology in which there are individual differences among people at every chronological age. As for Loevinger's assessment method, the WUSCT and SCT-Y have been extensively validated as measures of ego development. They have shown expected age changes in longitudinal studies of children and adolescents, they have predicted delinquent behavior and academic achievement in adolescents, and they are associated with behavior patterns reflecting such personality characteristics as altruism, responsibility, and conformity (see Bursik & Martin, 2006; Loevinger, 1998, chap. 5; Noam, Young, & Jilnina, 2006; Stakert & Bursik, 2006; Westen et al., 1999; Westenberg, Hauser, et al., 2004).

These abundant validating data have earned considerable praise in the literature for the psychometric soundness of the WUSCT. Garb, Lilienfeld, Wood, and Nezworski (2002) have described the WUSCT as an extensively validated measure of personality that can be "used to evaluate a range of traits, including impulse control, moral development, cognitive style, interpersonal styles, and conscious preoccupations" p. 463). Manners and Durkin (2001) concluded in a critical review of ego development theory and its measurement, "There is substantial empirical support for the conceptual soundness of ego development

theory and the WUSCT” (p. 541).

## REFERENCES

Ackerman, M. J., & Ackerman, M. C. (1997). Custody evaluations in practice: A survey of experienced professionals (revisited). *Professional Psychology, 28*, 137–145.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ames, P. C., & Riggio, R. E. (1995). Use of the Rotter Incomplete Sentences Blank with adolescent populations: Implications for determining maladjustment. *Journal of Personality Assessment, 64*, 159–167

### 540 Performance-Based Measures

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test usage among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94.

Archer, R. P., & Newsom, C. R. (2000). Psychological test usage with adolescent clients: Survey update. *Assessment, 7*, 227–235.

Bakken, L., & Huber, T. (2005). Ego development at the crossroads: Identity and intimacy among Black men and White women in cross-racial relationships. *Journal of Adult Development, 12*, 63–73.

Barton, S., Morley, S., Bloxham, G., Kitson, C., & Platts, S. (2005). Sentence completion test for depression (SCD): An idiographic measure of depressive thinking. *British Journal of Clinical Psychology, 44*, 29–46.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux [New methods for the diagnosis of abnormal intellectual level]. *L'Annee Psychologique, 11*, 193–244.

Bursik