The Scientific Status of Projective Techniques

Scott O. Lilienfeld,[1] James M. Wood,[2] and Howard N. Garb[3]

[1]Emory University, Atlanta, Georgia, [2]University of Texas at El Paso, El Paso, Texas, [3]Pittsburgh Veterans Administration Health Care System and University of Pittsburgh, Pittsburgh, Pennsylvania

Abstract—*Although projective techniques continue to be widely used in clinical and forensic settings, their scientific status remains highly controversial. In this monograph, we review the current state of the literature concerning the psychometric properties (norms, reliability, validity, incremental validity, treatment utility) of three major projective instruments: Rorschach Inkblot Test, Thematic Apperception Test (TAT), and human figure drawings. We conclude that there is empirical support for the validity of a small number of indexes derived from the Rorschach and TAT. However, the substantial majority of Rorschach and TAT indexes are not empirically supported. The validity evidence for human figure drawings is even more limited. With a few exceptions, projective indexes have not consistently demonstrated incremental validity above and beyond other psychometric data. In addition, we summarize the results of a new meta-analysis intended to examine the capacity of these three instruments to detect child sexual abuse. Although some projective instruments were better than chance at detecting child sexual abuse, there were virtually no replicated findings across independent investigative teams. This meta-analysis also provides the first clear evidence of substantial file drawer effects in the projectives literature, as the effect sizes from published studies markedly exceeded those from unpublished studies. We conclude with recommendations regarding the (a) construction of projective techniques with adequate validity, (b) forensic and clinical use of projective techniques, and (c) education and training of future psychologists regarding projective techniques.*

Controversy has been no stranger to the field of personality assessment, and no issue in this field has been more controversial than the scientific status of projective techniques. Indeed, the novice reader attempting to make sense of the sprawling and bewilderingly complex literature on projective techniques is immediately confronted with a striking paradox. On the one hand, during the past four decades a litany of personality assessment researchers (e.g., Anastasi, 1982; Gittelman Klein, 1986; Dawes, 1994) have come forth to decry the reliability and validity of most projective techniques (see Lilienfeld, 1999). Jensen's (1965) famous quotation, although 35 years old, still captures the sentiments of many contemporary scientists toward the Rorschach Inkblot Test and numerous other projective techniques: ". . . the rate of scientific progress in clinical psychology might well be measured by the speed and thoroughness with which it gets over the Rorschach" (p. 238). On the other hand, clinicians in the United States and to a lesser extent those abroad continue to use projective techniques with great regularity, and many contend that these techniques are virtually indispensable to their daily practice (Watkins, Campbell, Neiberding, & Hallmark, 1995). The crux of this paradox was incisively summed up by Anastasi (1982), who observed that "Projective techniques present a curious discrepancy between research and practice. When evaluated as psychometric instruments, the large majority make a poor showing. Yet their popularity in clinical use continues unabated" (p. 564).

Indeed, despite the sustained and often withering criticisms directed at projective techniques during the past several decades (Dawes, 1994; Lowenstein, 1987), numerous surveys demonstrate that such techniques continue to enjoy widespread popularity among clinicians. Durand, Blanchard, and Mindell (1988) reported that 49% of the directors of clinical psychology graduate programs and 65% of the directors of clinical psychology internships believed that formal training in projective techniques is important. Watkins et al. (1995) found that 5 projective techniques, including the Rorschach and Thematic Apperception Test (TAT), were among the 10 instruments most frequently used by clinical psychologists. For example, 82% of clinical psychologists reported that they administered the Rorschach at least "occasionally" in their test batteries and 43% reported that they "frequently" or "always" administered it. There is some indication, however, that the popularity of certain projective techniques may be waning. In a recent survey of practicing clinicians, Piotrowski, Belter, and Keller (1998) reported that several projective techniques, including the Rorschach and TAT, have been abandoned by a sizeable minority of users. Some authors (e.g., Piotrowski et al., 1998; Piotrowski & Belter, 1989) have attributed the recent decline in the popularity of projective techniques to the advent of managed care, although at least some of this decline may also stem from the cumulative impact of the criticisms leveled at these techniques during the past several decades. This decline notwithstanding, the Rorschach, TAT, and several other projective techniques remain among the most frequently used assessment devices in clinical practice.

Our central goal in this monograph is to examine impartially the best available research evidence concerning the scientific

Address correspondence to Scott O. Lilienfeld, Ph.D., Department of Psychology, Room 206, Emory University, Atlanta, GA 30322; email: slilien@emory.edu

status of projective techniques.[1] In contrast to some authors (e.g., Karon, 1978), we do not believe that the question of "Are projective techniques valid?" can be straightforwardly or meaningfully answered. We have assiduously avoided framing the question in this fashion for two reasons.

First, on the basis of the extant literature we will argue that the construct validity (Cronbach & Meehl, 1955) of certain projective indexes is more strongly supported than that of others. As a consequence, blanket assertions regarding the construct validity of all projective techniques appear to be unwarranted. Second, we concur with Messick (1995) that construct validity can be viewed as the extent to which one can draw useful inferences regarding individuals' non-test performance on the basis of their test scores. From this perspective, projective techniques are best regarded not as either valid or invalid, but rather as more or less valid for specific assessment purposes and contexts. Certain human figure drawing indexes, for example, may be moderately valid indicators of artistic ability (Kahill, 1984) or intelligence (Motta, Little, & Tobin, 1993) but largely or entirely invalid indicators of psychopathology. Thus, the overriding question we pose in this monograph is "To what extent are certain projective techniques—and specific indexes derived from them—valid for the purposes to which they are typically put by practitioners?"

It is critical at the outset to distinguish evidence for construct validity from evidence for predictive utility (see also Levy, 1963). An instrument that exhibits construct validity as evidenced by significant differences between pathological and nonpathological groups may nevertheless be virtually useless for real world predictive applications. This is because in many of the studies conducted on psychological instruments, including projective techniques, researchers begin with known groups (e.g., individuals with versus without a history of child sexual abuse) of approximately equal size. This 50-50 split between groups is optimal for predictive purposes from the standpoint of Bayes' theorem (Meehl & Rosen, 1955). Nevertheless, practitioners are most often interested in detecting clinical phenomena whose prevalence in most real world settings is considerably lower than 50 percent (e.g., a history of child sexual abuse, an imminent suicidal plan). As a result, validity estimates derived from investigations of known pathological groups, which are based on "conditioning on the consequence" (i.e., postdicting from group status to the presence or absence of a test indicator), will almost always yield higher estimates of validity than in actual clinical settings, where the practitioner must "condition on the antecedent" (i.e., predict from the presence or absence of a test indicator to group status; see Dawes, 1993). In other words, because clinicians are typically interested in detecting the presence of low base rate phenomena, most research designs used with known pathological groups will overestimate the predictive validity of test indicators. Thus, an index derived from a projective technique may possess construct validity without being useful for predictive purposes in real world settings.

In addition to validity, we examine the extent to which projective techniques satisfy other important psychometric criteria, particularly (a) reliability, viz., consistency of measurement, which itself encompasses test-retest reliability, interrater reliability, and internal consistency, (b) incremental validity, viz., the extent to which an instrument contributes information above and beyond other information (Meehl, 1959; Sechrest, 1963), and (c) treatment utility, viz., the extent to which an instrument contributes to treatment outcome (Hayes, Nelson, & Jarrett, 1987). Reliability is important because validity is limited by the square root of validity (Meehl, 1986). As a consequence, validity cannot be high when reliability is very low. Incremental validity is of considerable pragmatic importance in the evaluation of projective techniques because many of these techniques necessitate extensive training and are time consuming to administer, score, and interpret. If projective techniques do not contribute psychologically useful information above and beyond more easily collected data (e.g., scores on self-report instruments, demographic information) then their routine clinical use is difficult to justify. The question of incremental validity is also significant for theoretical reasons because many proponents of projective techniques claim that these techniques can provide valuable information not assessed by self-report indexes (Dosajh, 1996; Riethmiller & Handler, 1997a; Spangler, 1992). Incremental validity is not a single number, as it can be assessed relative to a variety of forms of information (e.g., scores from questionnaires, demographic data) that the clinician may have on hand. Finally, we agree with Hunsley and Bailey (1999) that the criterion of treatment utility is of paramount importance in the evaluation of all psychological instruments used by practitioners. In the therapeutic context, assessment is virtually always a means to an end, namely improved treatment outcome. If psychological instruments do not ultimately facilitate treatment in some measurable way, they are of doubtful utility in the clinical context, although they may nonetheless be useful for certain research or predictive purposes.

A PRIMER OF PROJECTIVE TECHNIQUES AND THEIR RATIONALE

The appropriate definition of projective techniques is less clear-cut than many authors have assumed. In contrast to structured ("objective") personality tests, projective techniques typically present respondents with an ambiguous stimulus, such as an inkblot, and ask them to disambiguate this stimulus. In other cases, projective techniques require participants to

[1]Due to space constraints, we have elected to focus only on the most central issues pertinent to the scientific status of projective techniques. Readers may obtain a more complete version of this manuscript from the first author upon request. This more comprehensive version also contains sections on the history of projective techniques, examiner and situational influences on projective techniques, the susceptibility of projective techniques to response sets (e.g., malingering, impression management), and reasons for the continued popularity of projective techniques.

generate a response (e.g., a drawing) following open-ended instructions (e.g., "Draw a person in any way you wish"). In addition, most projective techniques permit respondents considerable flexibility in the nature and sometimes even number of their responses. Although some authors (e.g., Schweighofer & Coles, 1994) define projective techniques as instruments that permit an extremely large (sometimes infinite) number of scoreable responses, this definition is overly inclusive. For example, according to this definition the vocabulary subtests of many standard intelligence tests would be classified as projective techniques, because the questions on these subtests (e.g., "What does 'justice' mean?") can in principle be answered in an infinite number of ways. We therefore view projective techniques as differing from structured tests on both the stimulus and response end. The stimuli used in such techniques tend to be more ambiguous than in structured tests, and the nature and number of their response options more varied.

As Meehl (1945) noted, however, most projective and structured personality instruments are best conceptualized as falling on a continuum. For example, many structured personality test items (e.g., "I often have headaches") entail a certain degree of ambiguity, because they consist of stems containing referents (e.g., the term "often") that can be interpreted in various ways. The extent to which such item ambiguity is a source of valid trait variance (Meehl, 1945) as opposed to measurement error (Jackson, 1971), however, remains a point of contention. Conversely, some traditional projective techniques place constraints on the variety and quantity of responses. For example, the once popular but long discredited (Borstellmann & Klopfer, 1953) Szondi test (Szondi, 1947) asks respondents to examine a set of photographs of patients suffering from different psychological disorders (e.g., paranoia, mania) and to select the photograph they most prefer, the assumption being that individuals tend to identify with the psychopathological condition to which they are most prone.

The rationale underlying most projective techniques is the projective hypothesis (Frank, 1948; see also Sundberg, 1977). According to this hypothesis, respondents project aspects of their personalities in the process of disambiguating unstructured test stimuli. As a consequence, the projective technique interpreter can ostensibly "work in reverse" by examining respondents' answers to these stimuli for insights regarding their personality dispositions. The concept of projection originated with Freud (1911), who viewed it as a defense mechanism by which individuals unconsciously attribute their negative personality traits and impulses to others. Nevertheless, the Freudian concept of projection ("classical projection") has not fared well in laboratory studies (Holmes, 1978), most of which offer relatively little evidence that the attribution of negative characteristics onto other individuals either reduces anxiety or protects individuals from the conscious awareness of these characteristics in themselves.

The negative experimental evidence regarding the existence of classical projection does not, however, necessarily vitiate the raison d'etre underlying most projective techniques. Instead, most of these techniques can be thought of as drawing on "generalized" or "assimilative" projection, namely, the relatively uncontroversial tendency for individuals' personality characteristics, needs, and life experiences to influence their interpretation ("apperception") of ambiguous stimuli (Sundberg, 1977). The principal advantages of most projective techniques relative to structured personality tests are typically hypothesized to be their capacity to (a) bypass or circumvent the conscious defenses of respondents and (b) allow clinicians to gain privileged access to important psychological information (e.g., conflicts, impulses) of which respondents are not consciously aware (Dosajh, 1996). As a consequence, proponents of projective techniques have often maintained that these techniques provide incremental validity in the assessment of personality and psychopathology above and beyond structured measures (e.g., Finn, 1996; Spangler, 1992; Riethmiller & Handler, 1997a; Weiner, 1999).

Before discussing subtypes of projective techniques, a word regarding terminology is in order. In this monograph we have elected to use the terms projective "techniques" or "instruments" rather than "projective tests" because most of these techniques as used in daily clinical practice do not fulfill the traditional criteria for psychological tests (see also Veiel & Coles, 1982). Specifically, with some important exceptions that we will discuss, most of these techniques as commonly used by practitioners do not include (a) standardized stimuli and testing instructions, (b) systematic algorithms for scoring responses to these stimuli, and (c) well calibrated norms for comparing responses with those of other individuals (see also Hunsley, Lee, & Wood, in press). As we will see, the absence of these features, particularly (a) and (b), renders the literature on certain projective techniques difficult to interpret, because some investigators have used markedly different stimuli, scoring methods, or both, across studies (e.g., see Keiser & Prather, 1990).

Following Lindzey's (1959) taxonomy, we subdivide projective techniques into five broad and partly overlapping categories (see also Aiken, 1996). Association techniques include inkblot or word association techniques. Construction techniques include human figure drawing methods and story creation methods, such as the TAT. Completion techniques include sentence completion tests and the Rosenzweig Picture Frustration Study. Arrangement or selection techniques include the Szondi Test and the Lüscher Color Test. Finally, expression techniques include projective doll play, puppetry, and handwriting analysis. The five major types of projective techniques in Lindzey's (1959) taxonomy, along with brief descriptions of two instruments illustrating each type of technique, are presented in Table 1.

FOCUS OF THE PRESENT MONOGRAPH

In this monograph, we examine the scientific status of three major projective techniques: (1) the Rorschach Inkblot Test,

Table 1. *The Five Major Subtypes of Projective Techniques and Two Examples of Each Subtype*

Subtype	Examples	Description
Association	*Rorschach Inkblot Test* (Rorschach, 1921)	Respondents are shown 10 symmetrical inkblots, 5 in black-and-white and 5 in color, and are asked to say what each inkblot looks like to them.
	Hand Test (e.g., Wagner, 1962)	Respondents are shown various pictures of moving hands, and are asked to guess what each hand "might be doing."
Construction	*Draw-A-Person Test* (Machover, 1949)	Respondents are asked to draw a person on a blank sheet of paper, and are then asked to draw another person of the opposite sex from the first person.
	Thematic Apperception Test (Murray & Morgan, 1938)	Respondents are shown pictures of ambiguous social situations and are asked to tell a story concerning the characters in each picture.
Completion	*Washington University Sentence Completion Test* (Loevinger, 1976)	Respondents are presented with various incomplete sentence stems (e.g., "If my mother . . .") and are asked to complete each stem.
	Rosenzweig Picture Frustration Study (Rosenzweig, Fleming, & Clark, 1947)	Respondents are shown cartoons of various frustrating situations (e.g., being accidentally splashed with water by a passing car) and are asked how they would respond verbally to each situation.
Arrangement/Selection	*Szondi Test* (Szondi, 1947)	Respondents are shown photographs of individuals with different psychiatric disorders, and are asked which patients they most and least prefer.
	Lüscher Color Test (Luscher & Scott, 1969)	Respondents are asked to rank order different colored cards in order of preference.
Expression	*Projective puppet play* (e.g., Woltmann, 1960)	Children are asked to play the roles of other individuals (e.g., mother, father) or themselves using puppets.
	Handwriting analysis (see Beyerstein & Beyerstein, 1992, for a review)	Individuals are asked to provide spontaneous samples of their handwriting.

(2) the TAT, and (3) human figure drawings. In addition, we will briefly review the evidence for the validity of one other projective technique, the Washington University Sentence Completion Test (Loevinger, 1998). A number of other projective techniques, such as the projective interpretation of handwriting (graphology; Beyerstein & Beyerstein, 1992), the Rosenzweig Picture Frustration Study (Rosenzweig et al., 1947), the Blacky Test (Blum, 1950; see also Bornstein, 1999), the use of anatomically detailed dolls in child abuse assessment (Alridge, 1998; Koocher, Goodman, White, Friedrick et al., 1995; Wolfner, Faust, & Dawes, 1993), the use of the Bender-Gestalt neuropsychological test for projective purposes (Naglieri, 1992), and the interpretation of early childhood memories (Bruhn, 1992), have been reviewed elsewhere and will not be discussed here. We recognize that readers with particular theoretical or psychometric preferences may quarrel with our principal focus on these three instruments. In limiting the primary scope of our inquiry to the Rorschach, TAT, and human figure drawings, we do not intend to imply that other projective techniques are without promise or potential merit. With the possible exception of the Washington University Sentence Completion Test and Rosenzweig Picture Frustration Study (Lilienfeld, 1999), however, we believe that none of these techniques is sufficiently well validated to justify its routine use in clinical practice. Ironically, neither the Washington University Sentence Completion Test nor the Rosenzweig Picture Frustration study is commonly used by practitioners (Holiday, Smith, & Sherry, 2000; Watkins et al., 1995).

We have elected to focus on the Rorschach, TAT, and human figure drawings for two major reasons. First, these three instruments, as well as cognate versions of them, are among the most frequently used projective techniques in clinical practice (Watkins et al., 1995). For example, a 1991 survey of clinical psychology graduate programs ranked these three instruments as the most highly emphasized of all projective techniques (Piotrowski & Zalewski, 1993). A later survey of clinical psychology internship directors ranked these instruments as the three projective techniques most often administered by interns (Piotrowski & Belter, 1999; see Durand & Blanchard, 1988, for similar results). The Rorschach remains especially popular. A fairly recent estimate placed the number of Rorschachs administered each year at 6 million (Sutherland, 1992). Second, these three instruments are among the most extensively researched of all projective techniques and therefore permit the most comprehensive evaluation at the present time. Because the psychometric concerns traditionally raised regarding these three instruments are applicable *a fortiori* to less well researched projective techniques, many of the conclusions we

draw concerning their scientific status will likely be applicable to other projective techniques.

As noted earlier, our review of the scientific status of these three instruments focuses primarily on zero-order validity (i.e., the correlations of these instruments with external indicators), although we also examine the evidence for their reliability, incremental validity, and treatment utility. We emphasize zero-order validity for two major reasons. First, such validity is a prerequisite for the clinical utility of projective techniques. Second, the absence of zero-order validity renders moot any examination of either incremental validity or treatment utility. In evaluating the zero-order validity of these instruments, we adopt with minor modifications the three criteria outlined by Wood et al. (1996b, p. 15) for the Rorschach. Specifically, following Wood et al., we propose that the indexes derived from projective techniques should exhibit (a) a consistent relation to one or more specific psychological symptoms, psychological disorders, real-world behaviors, or personality trait measures in (b) several methodologically rigorous validation studies that have been (c) performed by independent researchers or research groups. The lattermost criterion is important because it minimizes the possibility that replications are a consequence of systematic errors (e.g., artifacts stemming from flawed methods of administration or scoring) that may be inadvertently produced by researchers originating from the same laboratory. In this monograph, projective technique indexes that satisfy these three criteria will be provisionally regarded as "empirically supported." As noted earlier, however, we urge readers to bear in mind that even empirically supported indexes may be essentially useless for predictive purposes, especially when the clinician is interested in detecting low base rate phenomena (Dawes, 1993).

In certain cases, we evaluate the validity of a projective index not only by means of statistical significance but also with measures of effect size (e.g., *d* and *r*), which provide standard metrics for gauging the magnitude of an effect. For example, the *d* statistic describes the number of standard deviations that separate the means of two groups. According to Cohen (1982), $d = .2$ represents a small effect size, $d = .5$ represents a medium effect size, and $d = .8$ represents a large effect size. For most studies in which we report the *d* statistic, we have calculated this statistic from the means and standard deviations reported in the original article. Corresponding values for the correlation coefficient *r* are $r = .10$ (small), $r = .24$ (medium) and $r = .37$ (large; see Lipsey, 1990).

We also examine the extent to which published evaluations of projective techniques may be distorted by the "file drawer problem" (publication bias), i.e., the selective tendency of negative findings to remain unpublished (Rosenthal, 1979). Given the massive volume of research conducted on many projective techniques, it is possible that a substantial number of findings unfavorable to these techniques have not appeared in print. If so, the published literature on these techniques could paint an unduly positive picture of their validity. Despite the potential importance of the file drawer problem, it has received virtually no empirical attention in the literature on projective techniques (Parker, Hanson, & Hunsley, 1988). By comparing the magnitude of effects reported in published and unpublished studies of projective techniques in a large and important body of research—the detection of child sexual abuse—we hope to obtain a preliminary estimate of the magnitude of the file drawer problem in the projectives literature.

RORSCHACH INKBLOT TEST

No projective technique has aroused more controversy than the Rorschach Inkblot Test. As Hunsley and Bailey observed (1999, p. 266), the Rorschach "has the dubious distinction of being, simultaneously, the most cherished and the most reviled of all psychological assessment instruments."

Developed by Swiss psychiatrist Hermann Rorschach in the 1920s, this association technique (Lindzey, 1959) consists of 10 inkblots (five black and white, five containing color) that are each printed on a separate card. In the standard procedure, the client is handed the cards one at a time and asked to say what each blot resembles. This part of the procedure lasts about 45 minutes and an additional 1.5 to 2 hours are typically spent in scoring and interpreting the responses (Ball, Archer, & Imhoff, 1994). The respondent's statements can be scored for more than 100 characteristics, including those in the three major categories of (a) content (e.g., Did the client report seeing sexual content in the blots? Or human figures? Or food?), (b) location (e.g., Did the client report seeing the whole blot as one picture or just one particular area of the blot?), and (c) determinants (e.g., Did the client report seeing something that involved color? Or movement? Or shading?). Introduced into the United States in the late 1920s and 1930s, the Rorschach became a common target of scientific criticism in the 1950's and 1960's. Critics argued that the Rorschach lacked standardized administration procedures and adequate norms, and that evidence for its reliability and validity was weak or non-existent (Eysenck, 1959; Jensen, 1965; see summary by Dawes, 1994).

In the face of such criticisms, most psychologists might have gradually abandoned the Rorschach. However, the appearance of *The Rorschach: A Comprehensive System (TRACS)* (Exner, 1974) in the 1970s dramatically revived its fortunes. This book, along with its subsequent extensions and revisions (Exner, 1986, 1991, 1993; Exner & Weiner, 1995), seemed at last to establish the Rorschach on a firm scientific foundation. John Exner's Comprehensive System (CS) for the Rorschach provided detailed rules for administration and scoring, and an impressive set of norms for both children and adults. Exner did not view the Rorschach primarily as a projective technique. Instead, like Hermann Rorschach (see Rabin, 1968), Exner emphasized the perceptual nature of the client's response to the inkblots and the importance of location and determinants for test intepretation (Aronow, Reznikoff, & Moreland, 1995; Exner, 1989). Various editions of *TRACS* reported strikingly posi-

tive findings from hundreds of reliability and validity studies by Exner's Rorschach Workshops, although the large majority of these studies were unpublished and were not described in detail. The achievements of the CS elicited widespread praise. For example, the Board of Professional Affairs (1998, p. 392) of the American Psychological Association commended Exner for his "resurrection" of the test. Surveys in the 1990s indicated that the Rorschach was widely used in clinical and forensic settings and that the CS was the most commonly used Rorschach scoring system (Ackerman & Ackerman, 1997; Lees-Haley, 1992; Pinkerman, Haynes & Keiser, 1993; Piotrowski, 1999).

The present review focuses on the CS for two major reasons. First, although other Rorschach approaches are still used clinically, the scientific evidence to support them is generally weak. The same fundamental criticisms that were made in the 1960's concerning inadequate norms, poor or undemonstrated reliability, and limited evidence of validity still apply with equal force to virtually every non-CS approach in use today (Dawes, 1994; Gann, 1995; McCann, 1998).

The present review focuses on the CS for a second reason. Despite its popularity, the CS is currently engulfed in a scientific controversy that is at least as heated and widespread as the Rorschach controversy of the 1950s and 1960s. Numerous articles concerning the scientific status of the Rorschach CS have appeared in recent years (e.g., Garb, 1999; Meyer, 1997), and in 1999 and 2000 three peer-reviewed journals (*Psychological Assessment, Assessment, Journal of Clinical Psychology*) devoted Special Sections to debates concerning the psychometric properties of the CS. The points in contention include such fundamental issues as accuracy and cultural generalizability of the CS norms, scoring reliability, validity, clinical utility, and accessibility of supporting research (Acklin, 1999; Archer, 1999; Garb, 1999; Garb, Wood, Nezworski, Grove, & Stejskal, in press; Hunsley & Bailey, 1999, in press; Stricker & Gold, 1999; Viglione, 1999; Weiner, 1996, 1999, 2000; Wood & Lilienfeld, 1999; Wood, Lilienfeld, Garb, & Nezworski, 2000a, 2000b). The present discussion summarizes the central issues in the debate and reviews the most relevant publications on the topic.

Adequacy of the CS Norms

TRACS (Exner, 1991, 1993) provided extensive normative information for non-patient American adults and children, as well as statistical tables for several clinical reference groups (e.g., patients with schizophrenia). Rorschach proponents have frequently pointed to these norms as a major strength of the CS. For example, Weiner (1998, p. 27) asserted that "the size and diversity of these normative and reference samples provide more standardized information than is available for most psychological assessment measures and establishes the Rorschach Inkblot Method as adequately normed for a U.S. population."

Despite such positive appraisals, we and others have criticized the CS norms on the grounds that they are out of date and based on rather small samples compared with norms for well-established psychological instruments, such as the Wechsler intelligence tests and Minnesota Multiphasic Personality Inventory-2 (MMPI-2; e.g., Wood & Lilienfeld, 1999). More important, substantial evidence has recently emerged that the CS norms are unrepresentative of the U.S. population and tend to make normal adults and children appear maladjusted. In a study of 123 nonpatient adults from California, Shaffer, Erdberg, and Haroian (1999) found substantial discrepancies from the CS norms for many important Rorschach variables. For example, about one in six of the Shaffer et al. non-patient participants scored in the pathological range (≥ 4) on the Schizophrenia Index (SCZI). More than one-fourth of the nonpatients (29%) gave at least one Reflection response, a supposedly rare Rorschach indicator of narcissism (Exner, 1991). Substantial discrepancies were also reported for numerous other Rorschach indicators of emotional functioning and psychopathology. Nearly all the discrepancies had the effect of making the nonpatient group appear maladjusted compared with the normative data.

As a follow-up to the findings of Shaffer et al. (1999), Wood, Nezworski, Garb, and Lilienfeld (in press) recently aggregated data from 32 other published and unpublished Rorschach studies of nonpatient adults. The results reported by Wood et al. (2000) are similar to those reported by Shaffer et al.; apparently normal adults residing the community appear to be strikingly pathological when compared with the CS norms. Wood et al. concluded that (a) the norms for many CS variables are in error for both adults and children, (b) these discrepancies have the effect of "overpathologizing" normal individuals, and (c) the use of the CS norms in clinical or forensic settings may harm clients and be contrary to the ethical principles of psychologists and current professional standards for test usage.

No plausible explanation has been offered for why the CS norms might be so seriously in error. When critics of the CS have attempted to obtain copies of the unpublished manuscripts describing Exner's Rorschach Workshops studies (on which the norms are largely based), they have been told that the studies are not available in a form that can be released (for details, see Exner, 1996; Nezworski & Wood, 1995; Wood, Nezworski, & Stejskal, 1996a; Garb, Wood, et al., in press). Although the Rorschach Workshops studies form the primary empirical basis for the CS, they apparently cannot be examined by the scientific community for clues regarding problems with the CS norms.

Cultural Generalizability of the CS

Although Rorschach proponents often suggest that the Rorschach is well suited for use with American minorities or non-Americans (e.g., Butcher, Nezami, & Exner, 1998; Viglione,

1999), research evidence does not offer much support for this claim (Garb. Wood, et al., in press; Wood & Lilienfeld, 1999). Studies show that Blacks, Hispanics, Native Americans, and non-American groups often score differently on important variables comprising the CS and other Rorschach systems. Most criticisms have focused on the lack of appropriate normative data. For example, Krall et al. (1983) found that inner-city black children differed from then-current CS norms on 5 of 10 Rorschach variables. Glass, Bieber, and Tkachuk (1996) compared Alaskan native and non-native prisoners and concluded: "There were clear differences between native and non-native inmates on both the MCMI II and the Rorschach" (p. 583, Abstract). This study revealed that the Alaskan Native Americans differed significantly from the CS norms on two-thirds of Rorschach scores. Furthermore, Boscan (1999/2000) found that Rorschach scores of 101 Mexican college students differed significantly in many respects from the CS norms. Similar discrepancies have been reported for CS scores in Central and South American countries as well as in several European countries (see Dana, 2000, for a review of recent research). The interpretation of such comparative studies is complicated because, as discussed earlier in the present article, the CS norms themselves are questionable and do not accurately reflect the performance of normal North American adults and non-adults.

Extant studies therefore suggest that use of the CS with American minorities and non-Americans can be highly problematic. In addition, there is little, if any, research on the differential validity of Rorschach indexes across different racial and cultural groups. Such research is necessary to rule out the possibility of racial and cultural bias. As Dana (1993, p. 160) concluded, "The Rorschach and the Exner Comprehensive versions are not recommended for routine cross-cultural applications."

Scoring Reliability of the CS

Irving Weiner (1998, p. 55), a Rorschach proponent and former editor of the *Journal of Personality Assessment,* asserted that the scientific status of the CS rests on "three pillars": (1) a representative normative database, (2) objective and reliable scoring, and (3) standardized administration. We have already discussed the CS norms. In the present section we discuss problems with another of these three pillars: the scoring of CS variables.

For many years, psychologists accepted claims by Exner (1993, p. 23; see also Groth-Marnat, 1997, p. 397) that the scoring reliability of CS variables is uniformly above a minimum acceptable threshold of .85. However, recent studies of CS scoring reliability indicate that only about half of CS variables attain a reliability of .85 or higher according to the modern approach of calculating reliability using intraclass correlations or Kappa coefficients (Acklin, McDowell, Verschell, & Chan, 2000; Gronnerod, 1999; Nakata, 1999; see also Meyer, 1997a, 1997b, Shaffer et al., 1999, Wood, Nezworski, & Stejskal, 1997).[2]

For example, in a study with strong methodology, Acklin et al. (2000) computed intraclass correlation coefficients for approximately 95 CS scores in both a clinical ($n = 20$) and a non-clinical ($n = 20$) sample. Rorschach protocols were scored by two graduate clinical psychology students, each of whom had advanced training in the use of the CS and a minimum of 3 years of experience in CS coding procedures. The results for both samples were similar: the median reliability of CS scores was in the low .80s, the maximum was 1.0, and the minimum was approximately .20. As Acklin and his co-authors pointed out, interrater reliability was acceptable and at times even excellent for many CS scores. However, about 50% fell below .85. Furthermore, reliability was low for several widely used CS scores. For example, reliability coefficients for the Schizophrenia Index (SCZI) were .45 and .56 in the two samples. Similarly, interrater reliability was low for Adjusted D (.53 and .68), which is held forth as an important CS index of self-control under stress, and for X-% (.62 and .66), which is considered an indicator of perceptual and mental distortion (Exner, 1991, 1993; Weiner, 1998).

Acklin et al. (2000) concluded that reliability coefficients above .60 for CS variables are "substantial and acceptable." However, this conclusion appears overly sanguine. Although most statistical experts would agree that interrater reliabilities of .60 are minimally acceptable for research involving between-group comparisons (e.g., Fleiss, 1981; Landis & Koch, 1977), there is ample reason to question whether scores with reliabilities lower than .80 should be used to assess individual clients in clinical or forensic work. For example, the subtests of the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III; Wechsler, 1997) have a minimum interrater reliability of .90 and a median reliability of approximately .95 as measured by intraclass correlation (Psychological Corporation, 1997). Aside from clerical errors, interrater unreliability is of course not even a relevant concern for most self-report measures (e.g., the MMPI-2). Interrater reliabilities above .80 or .90 appear especially important in clinical work to ensure that the idiosyncrasies or subjective biases of individual scorers will exert little influence on a client's test scores.

Practitioners who use the Rorschach can be confident that about half of CS variables can potentially be scored at a level of reliability suitable for clinical work. However, it is equally

[2]A meta-analysis by Meyer (1997a, 1997b) yielded somewhat higher estimates of scoring reliability for the CS (range = .72 to .96; median = .89). However, Wood et al. (1997) criticized this meta-analysis on several grounds. Most importantly, the meta-analysis examined not the reliability of individual CS scores, but rather the reliability of Rorschach "segments," which combine numerous scores. Although Meyer claimed that "segment reliability" was a particularly stringent approach to Rorschach reliability, the results of Acklin et al. (2000) do not support this claim. Specifically, Acklin's findings show that although the reliability of a "segment" may seem excellent, the reliability of individual Rorschach scores included in the segment may be quite poor.

important to recognize that scoring reliability is problematic for a substantial number of CS variables and that the use of these variables to assess individual clients is inadvisable. For example, even among psychologists who are highly experienced with the CS or regarded as authorities, Rorschach scoring is not necessarily above challenge. Disagreements can have particularly serious implications if the test results are used to reach important clinical or legal recommendations. Rendering this issue potentially even more troublesome is the fact that the CS's field reliability—that is, the extent to which scores achieve high interrater reliability in actual clinical practice—is essentially unknown. Nevertheless, a study of CS scoring accuracy using alumni of the Rorschach Workshops suggests that field reliability may be problematic (Wood et al., 1996a).

Test-Retest Reliability of the CS

Rorschach proponents have sometimes argued that the test-retest reliability of CS scores is excellent. For example, Viglione (1999, p. 252) claimed that "the great majority of Rorschach Comprehensive System (CS) variables and configurations have shown impressive temporal consistency reliability. . . ." Nevertheless, test-retest results have actually been reported for only about 40% of the variables in the CS (for reviews and detailed citations, see Garb et al., in press; Wood & Lilienfeld, 1999). In books and articles by Exner and his colleagues, the test-retest coefficients have typically ranged from .30 to .90, with median values in the .80s or mid-to-high .70s (Meyer, 1997a, p. 487). However, when researchers other than Exner have reported test-retest coefficients for CS scores, the numbers have often been substantially lower than the figures reported in *TRACS* (e.g., Adair & Wagner, 1992; Erstad, 1995/1996; Perry, McDougall, & Viglione, 1995; Schwartz, Mebane, & Malony, 1990). Because of methodological limitations in the test-retest studies (see discussion by Garb et al., in press), only one firm conclusion can be drawn at present: the test-retest reliability of most CS scores is still an open issue that remains to be resolved by methodologically rigorous studies. In the meantime, the general assertion that CS scores have impressive test-retest reliability is unwarranted (Wood & Lilienfeld, 1999).

The Influence of Response Frequency (R) on CS Scores

For more than half a century, commentators have repeatedly noted that *R,* the total number of responses that clients give to the inkblots, can exert a substantial effect on their other Rorschach scores (Anastasi, 1988; Cronbach, 1949; Holtzman et al., 1961; Meyer, 1992a, 1993). For example, if one client gives 14 responses and a second client gives 28, the second client has twice as many opportunities to report aggressive content (supposedly indicative of aggressive personality characteristics) or morbid imagery (supposedly indicative of depression). Because *R* is higher in certain cultural and educational groups and because it is positively related to intelligence (Anastasi, 1982), certain groups of people may receive higher scores on Rorchach indexes of psychopathology simply because they give more responses.

Some psychologists (e.g., Groth-Marnat, 1984) believe that the CS has eliminated response frequency problems by adjusting for *R* or using ratios. In fact, however, many of the clinical scores and indexes of the CS are unadjusted. Furthermore, in a study of psychiatric patients, Meyer (1993) found that various CS indexes exhibited significant correlations with *R,* ranging from .25 for the Suicide Constellation to .60 for the Hypervigilance Index. In other words, clients who gave more responses on the Rorschach also tended to appear more pathological on various CS indexes.

Various solutions have been offered for the "problem of *R*." In the 1960s, Wayne Holtzman and his colleagues (Holtzman et al., 1961) developed an inkblot test resembling the Rorschach that used 45 cards instead of the traditional 10. Clients were instructed to give precisely one response to each card. Almost three decades later, Meyer (1989/1991) suggested that the 10 original Rorschach cards be retained, but that examinees be instructed to give exactly two responses to each card. Neither solution has met with a favorable reception. Holtzman's inkblot test has been largely ignored by clinicians despite its admirable research base and psychometric properties (see Peixotto, 1980), and Meyer's suggestion has excited little comment in the years since it was published. In general, Rorschach scholars and clinicians appear to believe that the problem of *R* does not exist, that it bears no important practical consequences, or that it is not worth remedying. For example, a recent article by Stricker and Gold (1999) on "Psychometrics and the Rorschach" did not even mention the problem. The words of Bill Kinder (1992, p. 253), current editor of the *Journal of Personality Assessment,* summarize the prevailing attitude:

> To propose limiting *R* when the Rorschach is used with individuals would mean the necessity of developing new normative, reliability, and validity data. In summary, there is very little to gain and a great deal to lose if we seriously propose limiting *R* in individual Rorschach records.

The Factor Structure of Rorschach Scores

The technique of factor analysis can provide guidance in identifying the dimensions that underlie the relationships among a set of test scores. In particular, factor analysis can help to reveal whether the correlations among scores conform to a meaningful pattern that is consistent with theoretical prediction. Several factor analyses of Rorschach scores have been published (see reviews by Costello, 1998, 1999; Meyer, 1989/1991, 1992b). Two important findings have emerged. First, the variables comprising the largest factor of the Rorschach, and

perhaps the second largest as well, load highly on *R* (Meyer, 1992b). In other words, the factor analyses are consistent with the observation offered in the previous section that *R* has a strong and pervasive influence on many Rorschach scores. This finding has important implications for the validity of the instrument:

> . . . the traditional use of the Rorschach, where a subject can give as many or as few responses as desired, seriously compromises the validity of the test, as approximately seventy percent of the common variability among Rorschach scores is simply due to error (response frequency). This fact alone calls into question almost all research conducted on the Rorschach, as most studies do not control for this variable (Meyer, 1989/1991, p. 229).

The second important finding to emerge from factor analyses is that various Rorschach scores usually do not intercorrelate or "hang together" in a way that is consistent with either theories about the test or common clinical practice (Costello, 1998; Meyer 1992b). The most thorough study on this issue was reported by Meyer (1992b). Based on interpretations published by Exner (1986), Meyer predicted that certain CS variables would intercorrelate to form well-defined theoretical factors. For example, Meyer predicted that Morbid responses, inanimate movement (*m*), Vista responses (*FV, VF, V*), diffuse shading (*FY, YF, Y*), and blends of shading (*Sh-Bl*) would intercorrelate and form a factor of "Neuroticism and Negative Affect." However, when Meyer (1992b) performed a factor analysis of the Rorschach, he found that these CS variables did not intercorrelate as predicted to form the expected Neuroticism factor. Similarly, the other intercorrelations and factors that emerged in Meyer's study did not conform to what he had predicted based on interpretations published by Exner.

Meyer concluded that "the Rorschach's internal structure does not clearly correspond to that which would be expected from traditional variable interpretation" (p. 132), and that "it is very doubtful that any theoretical perspective would actually predict the Rorschach structure" (p. 133). Although the factor analyses of Meyer (1992b) seem to require a fundamental reassessment of the construct validity of CS scores, Rorschach experts have been slow to come to grips with the study's implications. For example, the recent article by Stricker and Gold (1999) on "Psychometrics and the Rorschach" did not discuss these factor analytic findings at all.

Rorschach Validity: Global Meta-analyses

Several meta-analyses have compared the average validity of the Rorschach, MMPI, and the Wechsler Adult Intelligence Scales (WAIS) (e.g., Garb, Florio, & Grove, 1998; Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999; Parker, Hanson, & Hunsley, 1988). These global meta-analyses have adopted a "melting pot" approach of computing the correlations of a wide variety of Rorschach, MMPI, and WAIS variables with an equally wide array of criterion variables[3] to yield average validity coefficients for the three instruments.

Five comments can be made concerning these meta-analyses. First, all the meta-analyses have had serious methodological flaws. Their shortcomings will not be enumerated here, but readers are referred to critical discussions by Garb, Florio, and Grove (1998, 1999), Garb, Wood, et al. (in press), Hiller et al. (1999), Hunsley and Bailey (2000), and Parker, Hunsley, and Hanson (1999).

Second, all of these meta-analyses have been based exclusively on published studies. Because published studies often yield larger effect sizes than unpublished studies, an artifact known as the file drawer effect or publication bias (Cooper, DeNeve, & Charlton, 1997; Lipsey & Wilson, 1993), meta-analytic estimates of Rorschach validity may be inflated. Later in this article, we summarize for the first time the results of a meta-analysis of projective techniques designed in part to address the issue of publication bias.

Third, without additional follow-up analyses of specific scores, global meta-analyses of multiple-score instruments like the Rorschach and MMPI are of limited clinical value because they do not address whether any particular test score is valid for any particular purpose (Hunsley & Bailey, 1999; Parker, Hunsley, & Hanson, 1999; Weiner, 1996). As Rorschach proponent Irving Weiner explained (1996, p. 207):

> . . . an overall sense of the validity of multidimensional instruments can be arrived at only by some mathematical or impressionistic averaging of the validity coefficients of their component parts. Such averaging may at times prove useful. . . . There is always some risk, however, that the averaging of validity coefficients will conceal more than it reveals about an instrument, especially if the instrument includes both some highly valid scales and some scales with few, if any, valid correlates.

Fourth, despite the limitations that have just been enumerated, the meta-analyses have converged on more or less the same number: global meta-analyses of published Rorschach studies have generally yielded mean validity coefficients (*r*) of approximately .30 (plus or minus .05). As even Rorschach critics agree, such findings "suggest that some Rorschach indexes can possess moderate validity" (Hunsley & Bailey, 1999, p. 269). However, given the effects of publication bias and other methodological artifacts, the .30 figure may represent an overestimate of the average validity of Rorschach scores.

Fifth, meta-analyses suggest that the average validity of published studies is generally lower for the Rorschach than for the WAIS (Parker et al., 1988). Although this point is more controversial, meta-analyses also suggest that the average va-

[3] Although we use the terms "criterion variables" or "criteria" in this manuscript for the sake of convenience, it should be borne in mind that virtually none of these variables in the domain of personality assessment is strictly a "criterion" in the sense of providing an essentially infallible indicator of its respective construct (see also Cronbach & Meehl, 1955).

lidity of published studies is generally lower for the Rorschach than for the MMPI, although the difference is probably not large and sometimes fails to attain statistical significance (Garb et al., 1998; Hiller et al., 1999; Parker et al., 1988; see also discussion in Garb, Wood, et al., in press). Again, these conclusions must be tempered by the caveat that they are based on published studies only, and that the meta-analyses contained various methodological flaws.[4]

Rorschach Validity: Narrowly Focused Literature Reviews and Meta-analyses

As already noted, global meta-analyses by themselves cannot address which specific Rorschach scores are valid for which specific purposes. Instead, narrowly focused narrative or meta-analytic reviews, which concentrate on the relationship of one or two Rorschach predictors to a few specific criteria, are better suited for such a task. In the present section we summarize the relevant focused reviews and meta-analyses that have been published during the past decade. We do not include "overview" articles like the present one that have briefly summarized the evidence regarding a large number of Rorschach variables (e.g., Hunsley & Bailey, 1999; Viglione, 1999; Wood, Nezworski, & Stejskal, 1996a). In addition, we leave reviews regarding psychiatric diagnoses and self-report tests to the following section.

In a series of brief focused literature reviews, Frank concluded that there are no well-demonstrated relations between (a) color responses and emotional expression or control (Frank, 1976, 1993c), (b) achromatic color responses (C') and depression (Frank, 1993a), (c) shading responses and anxiety (Frank, 1978, 1993d), (d) space responses and oppositionality or hostility (Frank, 1993e), (e) movement responses and intelligence or degree of "inner life" (Frank, 1979b, 1993b, 1997), or (f) Rorschach content and aggressive behavior (Frank, 1994b). On the other hand, Frank concluded that good form quality (F+%) (g) is related to psychotherapy outcome (Frank, 1993f) and (h) differentiates psychotic from nonpsychotic patients, schizophrenic from non-schizophrenic patients, and process schizophrenics from reactive schizophrenics (Frank, 1979, 1980, 1994a). Furthermore, Frank concluded that (i) good form quality, in combination with the form-color ratio (FC:CF +C), may be useful for predicting suicidal or aggressive behavior (Frank, 1994a, 1994b, 1994c, 1997).

Additional negative results can be described. A focused literature review by Nezworski and Wood (1995) (see also updates by Wood, Lilienfeld, et al., 2000; and Wood, Nezworski, Stejskal, & McKinzey, in press) concluded that the Egocentricity Index is probably not significantly related to self-esteem or self-focus, and that Reflection responses do not bear a consistent relationship to either narcissism or psychopathy (but see Exner, 1995). A focused review by Jorgensen et al. (2000) concluded that the CS Schizophrenia Index (SCZI) effectively discriminates between psychotic and non-psychotic patients. In a focused review of research regarding the *D* score and Adjusted *D,* Kleiger (1992, p. 293; but see Exner, 1992) noted "two broad problem areas." First, about half of the empirical reports on major structural concepts in the CS are unpublished and have not undergone peer review. As a result, it is difficult to exclude the possibility that these studies contain methodological flaws that may systematically influence the effect size of CS scores. Second, the findings of the published studies appear "equivocal."

Positive findings have been reported for several Rorschach scores. A focused review by Acklin (1999) concluded that the Thought Disorder Index for the Rorschach (TDIR; Johnston & Holzman, 1979; Solovay et al., 1986) is a valid indicator of thought disorder in schizophrenic and borderline patients. A focused review by Jorgensen et al. (2000) concluded that the CS Schizophrenia Index (SCZI) effectively discriminates between psychotic and non-psychotic patients. A meta-analysis by Meyer and Handler (1997) concluded that the Rorschach Prognostic Rating Scale (RPRS) bears a well-established relationship to treatment outcome. Finally, a meta-analysis by Bornstein (1999; see also Bornstein, 1996) suggested that the Rorschach Oral Dependency Scale (ROD; Masling, Rabie, & Blondheim, 1967) is related to objective behaviors indicative of dependency.

As may be seen, four successful Rorschach scores have been identified in either meta-analyses or focused literature reviews: the TDIR, the SCZI, the RPRS, and the ROD. However, several comments seem appropriate regarding these scales. First, only one of the four, the SCZI, is part of the CS. Thus, the positive performance of the remaining three scores does not address the question of CS validity that has been

[4] Weiner (2000, p. 477) asserted that psychiatric diagnoses correlate more highly with the MMPI than with the Rorschach because diagnoses and the MMPI share "substantial method variance" (i.e., both involve self-report). Despite its seeming plausibility, for four reasons we believe that Weiner's assertion is without a sound basis. (1) Clinicians often use multiple information sources when formulating diagnoses, including self-reports, clinical observation, review of records, and interviews with collateral contacts. Weiner did not cite research evidence that the correlation of the MMPI with diagnoses is due solely to overlapping self-reports, and not to covariation with the other sources of information. (2) As Campbell and Fiske (1959) explained, an information source may contain both true variance and method variance, which by definition do not overlap. To the extent that self-reports are valid, they contain true variance. Thus, if MMPI scores and diagnosticians both use self-reports, this shared information may constitute "shared true variance" and not "method variance." (3) Weiner pointed out that both MMPI scores and diagnoses are based to some extent on shared information (i.e., self-reports). However, the same is true regarding Rorschach scores and diagnoses. For example, scores on the CS SCZI may be based in part on deviant verbalizations by the client, just as diagnoses of schizophrenia may be based in part on deviant verbalizations (i.e., disordered speech) observed during a diagnostic interview. In such a case, the Rorschach and diagnostic interview draw on the same valid source of information (deviant verbalizations), and this shared variance is true variance, not method variance. (4) Even if Weiner's criticism of the MMPI were entirely correct, he still would not be any closer to explaining why the Rorschach generally fails as a diagnostic tool. That is, even if the MMPI were totally invalid, that fact would not make the Rorschach more valid.

raised by such critics as Wood et al. (1996a, 1996b). Second, much of the research supporting the validity of the RPRS is of poor methodological quality (Hunsley & Bailey, 1999, p. 274). Third, two of the four successful scales present special scoring difficulties for clinical practice. Specifically, (a) the RPRS is scored using cumbersome rules from the now rarely used Klopfer system (Meyer & Handler, 1997; Meyer, 2000a) and (b) the TDIR is typically scored by research teams using tape-recorded Rorschach sessions (e.g., Coleman et al., 1993). Fourth, representative or up-to-date norms are not available for the TDIR, RPRS, or ROD, and the current norms for the SCZI (Exner, 1993) often seem to yield an unacceptably high false positive rate, especially with children (Hamel, Shaffer & Erdberg, 2000; Shaffer et al., 1999; but see Jorgensen et al., 2000). Thus, although narrative and quantitative literature reviews regarding the TDIR, SCZI, RPRS and ROD are encouraging, there are problems with their application to clinical practice.

Rorschach Validity: Relationships with Diagnoses and Self-Report Instruments

The Rorschach is often described by its proponents as a helpful diagnostic tool. For example, in a recent discussion of "Differential Diagnosis," Weiner (1997) claimed:

> At present the Rorschach Comprehensive System provides indices for schizophrenia (SCZI) and depression (DEPI) that can prove helpful in identifying these two conditions. . . . Recent work by Gacono and Meloy (1994) suggested that a similarly sound and useful index of psychopathic personality can now be constructed. . . . In addition, although further documentation is needed, accumulating data indicate that there are on the horizon adequately conceptualized and empirically valid Rorschach indices for bipolar disorder, borderline and schizotypal personality disorder, and acute and chronic stress disorder. . . . (pp. 10-11).

However, these same Rorschach proponents have sometimes adopted a considerably different position. For example, just two years later Weiner (1999) asserted that:

> The Rorschach Inkblot Method is not a diagnostic test, it was not designed as a diagnostic test, it is not intended to be a diagnostic test, and it does not in fact work very well as a diagnostic test, especially if what is meant by diagnosis is a DSM category (pp. 336-337).

Because such claims appear to be contradictory, the best course is to turn to the scientific literature for illumination. The authors of the present article and a colleague recently reviewed the research on the Rorschach and psychiatric diagnoses (Wood, Lilienfeld et al., 2000a, 2000b). We reached the following conclusions:

> Despite a few positive findings, the Rorschach has demonstrated little validity as a diagnostic tool. Deviant verbalizations and bad form on the Rorschach, and indexes based on these variables, are related to Schizophrenia and perhaps to Bipolar Disorder and Schizotypal Personality Disorder. Patients with Borderline Personality Disorder also seem to give an above-average number of deviant verbalizations. Otherwise the Rorschach has not shown a well-demonstrated relationship to these disorders or to Major Depressive Disorder, Posttraumatic Stress Disorder (PTSD), anxiety disorders other than PTSD, Dissociative Identity Disorder, Dependent, Narcissistic, or Antisocial Personality Disorders, Conduct Disorder, or psychopathy. (p. 395)

For example, the DEPI has been the most extensively studied Rorschach indicator of depression. According to Exner (1991, p. 146), an elevated score on the DEPI "correlates very highly with a diagnosis that emphasizes serious affective problems." However, independent investigators have usually found that diagnoses of depression are not significantly related to scores on the original or revised version of the DEPI, either in adults or adolescents (for reviews, see Jorgensen, Anderson, & Dam, in press; Viglione, 1999; Wood, Lilienfeld, et al., 2000a).

Similarly, Meloy and Gacono (1995, p. 414) claimed that the Rorschach is "ideally suited" for the assessment of psychopathy, and that through a series of studies "we have validated the use of the Rorschach as a sensitive instrument to discriminate between psychopathic and nonpsychopathic subjects." Yet numerous replication studies by independent researchers have failed to cross-validate the Rorschach indicators of psychopathy proposed by Meloy and Gacono (Wood, Lilienfeld, et al., 2000a; Wood, Nezworski, et al., in press).

Just as CS scores show no replicable relations with most psychiatric diagnoses, neither do they show a consistent relationship to self-report indexes. For instance, after reviewing the relationships between Rorschach and MMPI scores in 37 studies of adults, Archer and Krishnamurthy (1993b, p. 277) concluded, "The results of these studies generally indicate limited or minimal relationships between the MMPI and Rorschach" (see also Archer & Krishnamurthy, 1993a, 1997). Rorschach proponent Gregory Meyer (1996, p. 572) summarized the situation as follows:

> Archer and Krishnamurthy (1993a, 1993b) and Meyer (in press) have established that Rorschach and MMPI constructs do not converge on a common universe of information in unrestricted heterogeneous samples. This finding is so robust that additional efforts to find cross-method correlates in heterogeneous samples would be redundant. . . . Currently, there is enough research to conclude the Rorschach does not consistently or globally measure self-reported characteristics.

Rorschach proponents have sometimes attempted to explain the failure of Rorschach scores to correlate with diagnoses or self-report instruments. For instance, Stricker and Gold (1999, p. 244) have stated:

> The profile that emerges from the Rorschach may or may not correspond to the profiles that are obtained from self-report measures, interviews, or behavioral observations. Nor is correspondence or lack of correspondence seen as more or less meaningful or desirable.

Using such reasoning, one can readily dismiss most negative results obtained for the Rorschach and other projective instruments. Nevertheless, the fact remains that hundreds of studies have been carried out by researchers who expected CS scores to correlate with psychiatric diagnoses and self-report instruments. The general failure of CS scores to correlate with such criteria casts doubt on the Rorschach's validity for most purposes (Hunsley & Bailey, 1999, in press; but see Viglione, 1999).

Incremental Validity

Given that administering, scoring, and interpreting a Rorschach takes 2 to 3 hours, one would hope that the addition of the Rorschach to other information (e.g., an interview) would lead to improved validity. However, results on incremental validity offer little support for the use of the Rorschach when other assessment information is available. This has been true for both clinical judgment and statistical prediction studies.

When psychologists made judgments after being given increasing amounts of assessment information, the addition of the Rorschach almost never led to an increase in the validity of their judgments (for reviews, see Garb, 1984, 1998). For example, psychologists did not become more accurate when: (a) the Rorschach was added to demographic data (e.g., Gadol, 1969), (b) a test battery that included the Rorschach and a Sentence Completion Test was added to demographic data (e.g., Cochrane, 1972), and (c) the Rorschach was added to other test results or biographical information (e.g., Bilett, Jones, & Whitaker, 1982; Golden, 1964; Perez, 1976; Whitehead, 1985). In fact, in several studies, validity decreased when the Rorschach was added to other information (e.g., Gadol, 1969; Golden, 1964; Sines, 1959; Whitehead, 1985). Nevertheless, the results from clinical judgment studies are not definitive. Although the study by Whitehead (1985) used the CS, it is not clear how many of the other investigations did so. However, it is safe to conclude that these results offer negligible support for the use of the Rorschach in clinical settings, particularly when other readily obtained demographic or assessment information is available. Moreover, at least one study using the CS (Whitehead, 1985) yielded negative results similar to those from studies that used other Rorschach methods.

In statistical prediction studies, there has been support for the incremental validity of a few Rorschach scores. Specifically, (1) predictions of future psychotic symptoms were significantly improved when the Rorschach Thought Disorder Index was added to information from a clinical interview (O'Connell, Cooper, Perry, & Hoke, 1989, (2) the amount of variance accounted for in a laboratory measure of prepulse inhibition (which assesses the inability of patients to screen out irrelevant stimuli) was increased when X-% (a Rorschach index of perceptual inaccuracy) was added to an interview rating of delusional symptoms (Perry, Geyer, & Braff, 1999), (3) prediction of the severity of psychopathology and social competence among psychiatric patients were significantly improved when Rorschach scores for *R,* X+%, X-%, and the Ego Impairment Index were added to MMPI scores (Perry, Moore, & Braff, 1995, reanalyzed by Dawes, 1999; Perry & Viglione, 1991), (4) predictions of treatment outcome were significantly improved when scores from the RPRS were added to IQ scores and scores from the MMPI Ego Strength scale (Meyer, 2000a; Meyer & Handler, 1997), and (5) predictions of schizophrenia diagnoses and psychotic conditions were improved significantly when the CS Schizophrenia Index was added to MMPI scores (Meyer, 2000b). In addition, (6) predictions of depression diagnoses showed a small but statistically significant improvement (from $R = .33$ to $R = .35$) when the CS DEPI was added to MMPI scores (Meyer, 2000b), although in another study hit rates for depression diagnoses did not significantly improve when the DEPI or other CS variables (i.e., Vista or Affective Ratio) were added to MMPI-A scales (Archer & Krishnamurthy, 1997). In addition, no significant incremental validity was found when Rorschach scores were added to MMPI scores for diagnoses of conduct disorder (Archer & Krishnamurthy, 1997).

Although the findings regarding the RPRS have been replicated by independent researchers, the other findings have not and should thus be regarded as tentative. In addition, studies of statistical incremental validity are of direct clinical relevance only if practitioners rely exclusively and precisely on the output of statistical prediction rules. They do so only very rarely (Dawes, Faust, & Meehl, 1989). Overall, incremental validity has not been studied for the vast majority of Rorschach scores. Thus, for nearly all Rorschach scores, including scores comprising the CS, there is no evidence for incremental validity above and beyond other psychometric information.

Summary and Discussion

Despite its continued widespread use by clinicians, the Rorschach Inkblot Test remains a problematic instrument from a psychometric standpoint. Although many psychologists initially believed that the CS (Exner, 1974) remedied the Rorschach's primary shortcomings, the scientific status of this system appears to be less than convincing. The CS norms for many Rorschach variables appear to have the effect of misclassifying normal individuals as pathological, the possibility of significant cultural bias in the CS has not been excluded, the inter-rater and test-retest reliabilities of many CS variables are either problematic or unknown, the factor structure of CS variables does not correspond to investigators' theoretical predictions, and the validities of most Rorschach indexes rest on a weak scientific foundation.

At the same time, dismissing the Rorschach in broad brush as invalid oversimplifies the genuine state of affairs. Meta-analyses of published research on the Rorschach (e.g., Garb et al., 1999; Hiller et al., 1999) suggest that at least some Rorschach indexes possess above-zero validity, although the clini-

cal utility of these indexes remains to be demonstrated. Moreover, narrowly focused literature reviews have identified several Rorschach variables that appear to possess validity in the identification of schizophrenia, BPD, and perhaps schizotypal personality disorder and bipolar disorder. Four other Rorschach variables appear to be positively correlated with thought disturbance, psychotherapy prognosis, and dependency. Nevertheless, the substantial majority of Rorschach variables have not demonstrated consistent relations to psychological disorders or personality traits. Perhaps most important, few Rorschach variables have demonstrated consistent incremental validity in the assessment of psychologically meaningful construct indicators above and beyond other, more readily acquired, psychometric information.

THEMATIC APPERCEPTION TEST

The Thematic Apperception Test (TAT) is a construction technique (Lindzey, 1959) developed by Henry Murray and his student Christiana Morgan to assess reactions to ambiguous interpersonal stimuli (Morgan & Murray, 1935; Murray, 1943). Murray chose the term "apperception" as opposed to perception to denote the fact that respondents actively interpret TAT stimuli in accord with their personality traits and life experiences (Anderson, 1999). The TAT consists of 31 cards depicting ambiguous situations, most of them social in nature (e.g., a young woman grabbing the shoulders of a young man who appears to be attempting to pull away from her). One of these cards, Card 16, represents the epitome of ambiguity: it is entirely blank. Although some TAT cards are intended for males and others for females, neither examinee sex nor gender role has been found to be significantly associated with the content of TAT stories (Katz, Russ, & Overholser, 1993). The TAT has spawned a variety of cognate apperception tests developed for different age groups, such as the Children's Apperception Test (Bellak & Bellak, 1991), the Roberts Apperception Test for Children (McArthur & Roberts, 1990), the Adolescent Apperception Cards (Silverton, 1993), and the Senior Apperception Test (Bellak, 1975). Because the research literature on these techniques is considerably less extensive than that on the TAT, these techniques will not be reviewed here (see Hayslip & Lowman, 1986, and Kroon, Goudena, & Rispens, 1998, for reviews of apperception techniques for the elderly and children/adolescents, respectively).

With regard to TAT administration, the respondent is asked to look at each card and construct a story. Each story should describe what (a) led up to the events depicted on the card, (b) events are occurring on the card, (c) events will occur in the future, and (d) the characters on the card are thinking and feeling (Murray, 1943). Murray assumed that the respondent typically identifies with the primary protagonist featured in each card (the "hero") and creates the story from the vantage point of the hero.

Murray recommended that TAT examiners select approximately 20 cards whose themes appear particularly relevant to the respondent's presenting difficulties, and administer these cards across two sessions. Nevertheless, these recommendations are almost never followed today. There is considerable variability in the number of TAT stimuli administered by different examiners, and most administer between 5 and 12 cards and do so in only one session (Vane, 1981). Moreover, the specific cards selected and order of card administration vary greatly across examiners (Groth-Marnat, 1997; Worchel & Dupree, 1990). The modal time needed for administering, scoring, and interpreting the TAT is approximately 1.5 hours (Ball, Archer, & Imhof, 1994).

General Overview of TAT Research: Problems and Findings

Although a variety of TAT quantitative scoring schemes have been developed, such as those of Bellak (1975), Dana (1955), and Arnold (1962; see Vane, 1981, for a review), few clinicians use these schemes with any regularity (Rossini & Moretti, 1997). Instead, most interpret the TAT on an impressionistic basis using clinical judgment and intuition. For example, a survey of nearly 100 North American psychologists practicing in juvenile and family courts revealed that only 3 percent use any standardized TAT scoring system (Pinkerman, Haynes, & Keiser, 1993; see also Wade & Baker, 1977).

More pertinent to the present review is the fact that there is little consistency regarding which TAT cards are used in published research. In a review of 69 published studies of the TAT over a 10 year period, Keiser and Prather (1990) found enormous variability across investigations in the cards used and even in whether these cards were in the original TAT set. They concluded that the extent to which TAT findings can be generalized across investigations is unknown. The wide variety of stimulus sets used in TAT research also implies that adequate norms for virtually all TAT scoring systems are unavailable.

In addition to the substantial variability in stimulus sets and scoring schemes, there are at least two major obstacles to evaluating claims regarding the TAT's validity. It is largely these problems that render the TAT "a clinician's delight and a statistician's nightmare" (Vane, 1991, p. 319). The first interpretive problem has been termed the "Walter Mitty" effect (Loevinger, 1987), and refers to the fact that some respondents may display high levels of a given attribute (e.g., achievement motivation) on the TAT not because they possess high levels of this attribute, but because they are fantasizing about possessing high levels of this attribute. Conversely, some TAT proponents have maintained that individuals can exhibit low levels of an attribute on the TAT not because they possess low levels of this attribute, but because they are repressing or otherwise inhibiting the expression of this attribute. We term this purported phenomenon the "inhibition effect." Because both the Walter Mitty and inhibition effects can be invoked as ad hoc immunizing tactics (Popper, 1959) to explain away negative find-

ings, they can render certain predictions regarding the TAT's validity difficult or impossible to falsify. For example, Vane (1981) argued that:

> . . . an individual may tell stories in which much aggression is present, but does not show this in his actions. Many clinicians would have no difficulty reconciling this discrepancy because the protocol would be interpreted to mean that the individual really is aggressive, but is afraid to be so. Because of this fear, he has built a successful defense against aggression and thus appears meek and mild . . . On the other hand, if an aggressive individual told stories with many aggressive themes, he would be considered an aggressive individual (pp. 332-333).

Second, proponents of the TAT have sometimes offered conflicting assertions regarding the relations between the TAT and self-report indexes, which are sometimes used as indicators with which to validate the TAT. For example, Cramer (1999) argued that the characteristics assessed by the TAT "are, by definition, inaccessible to consciousness" and that "attempts to establish concurrent validity between TAT and self-report measures are not likely to be successful" (p. 85). McClelland, Koestner, and Weinberger (1989) similarly maintained that because the TAT and self-report instruments assess different types of motives (implicit and self-attributed, respectively; see section entitled TAT-based need scoring schemes), the correlations between these two types of instruments should be very low. In contrast, Westen and his colleagues (e.g., Barends, Westen, Leigh, Silbert, and Byers, 1990) have adduced positive correlations between their TAT index of object relations and self-report indicators of psychopathology as evidence for the validity of this TAT index. As a consequence, it is unclear whether findings (e.g., Emmons & McAdams, 1991) indicating significant positive correlations between TAT indexes and self-report indicators of the same constructs (e.g., achievement motivation) argue for or against the validity of the former indexes.

Rendering a systematic review of the TAT literature even more difficult is that the fact that a plethora of different scoring schemes for the TAT have been developed for certain studies, many on an ad hoc basis (Ryan, 1985; Vane, 1981). The great diversity of these schemes, not to mention the diversity of stimulus materials on which they are based, all but precludes a systematic examination of their replicated psychometric properties (Hunsley et al., in press). With some notable exceptions that we will discuss, the track record of standardized scoring schemes for the TAT has been at best mixed. On the one hand, some investigators have reported encouraging results. For example, Mussen and Naylor (1954) reported that levels of expressed aggression on the TAT were significantly associated with overt aggression among a sample of 29 male adolescent delinquents (but see Gluck, 1955, and Kagan, 1956, for failures to replicate this finding). Karon and O'Grady (1970) asked clinical psychology graduate students blind to patient identity to make ratings of emotional health from the TATs of inpatient schizophrenics. After statistically transforming these ratings using a scaling procedure, Karon and O'Grady found these ratings to display high predictive validity ($r = .63$ and .64 in two studies) for the number of days that patients remained in the hospital during a 6 month period.

In contrast to these fairly isolated positive findings, most of which have not been replicated, the TAT literature is replete with strikingly negative findings. In a study of 150 U.S. male war veterans, Eron (1950) reported that a TAT index of affect tone was unable to distinguish significantly among psychotic individuals, neurotic individuals, and normals. Sharkey and Ritzler (1985) found that the TAT was unable to differentiate samples of psychotic individuals, depressed individuals, and normals on the basis of perceptual distortions, unusual story interpretations, or affect tone. In fact, the affect tone of TAT stories was nonsignificantly more positive for depressed individuals than for normals. Murstein and Mathes (1996) found essentially no association ($r = .03$) between a self-report neuroticism questionnaire and rated psychopathology on the TAT among a nonclinical sample. Lilienfeld, Hess, and Rowland (1996) found that a TAT-derived index of future time perspective adapted from the work of Ricks, Umbarger, and Mack (1964) exhibited perfect interrater reliability (intraclass $r = 1.00$), but correlated negligibly and nonsignificantly with a host of indexes of personality and psychopathology as well as with alternative (e.g., self-report, projective) indexes of future time perspective.

Incremental Validity

The incremental validity of the TAT above and beyond either demographic or test information has typically been unimpressive (Gallucci, 1990; Garb, 1984, 1998). Soskin (1954) found that the addition of TAT protocols did not add significantly to the validity of clinicians' personality ratings of normal participants above and beyond basic demographic information. In contrast, Golden (1964) found that clinicians' judgments concerning the personality traits of participants (both psychiatric and nonpsychiatric patients) increased significantly when the TAT was added to demographic information. Nevertheless, as Golden pointed out, these judgments probably increased in validity simply because the demographic data did not provide information on whether individuals were psychiatric patients. Adding TAT data to the interpretive mix may have increased validity because these data permitted judges to determine whether individuals exhibited psychopathological symptoms (Garb, 1984). Once having made this determination, they were able to describe participants in terms of personality traits that characterize most or all psychiatric patients (see also Horowitz, 1962). Golden also found that adding the TAT to either the MMPI or to the Rorschach generally led to slight, but nonsignificant, increases in validity. In contrast, adding the TAT to both the Rorschach and MMPI led to essentially no increases in validity. Wildman and Wildman

(1975) asked a group of clinical psychologists to determine, on the basis of various combinations of test results, whether respondents were psychiatric patients or nonpatients (nurses) who had been matched to the patients for age and education. Although adding the MMPI to the TAT resulted in an increase in accuracy from 57% to 80%, adding the TAT to the MMPI resulted in a decrease in accuracy from 88% to 80%.

To our knowledge, only one study has examined the incremental validity of judgments made by statistical decision rules (in the studies described in the previous paragraph, judgments were made by clinicians). Winch and More (1956) examined the extent to which numerical information derived from the TAT contributed to the predictions of participants' (members of 25 married couples) scores on 12 of Murray's (1938) needs (e.g., achievement, dominance, hostility) above and beyond interview information. Virtually none of the increments in variance corresponding to the entry of the TAT in hierarchical multiple regression equations was statistically significant, and all were small in magnitude (range of 0 to 2%).

Because the great diversity of TAT stimuli and scoring schemes renders a comprehensive review of this literature impractical, we have elected to focus on three systematic scoring approaches to the TAT that appear potentially promising. These three approaches are (a) need scoring schemes, (b) the assessment of object relations, and (c) the assessment of defense mechanisms.

TAT-Based Need Scoring Schemes

The best known need-based scoring scheme developed for the TAT is the quantitative system developed by McClelland, Atkinson, Clarke, and Lowell (1953) to assess Murray's (1938) need for achievement. Respondents are asked to write stories in response to several (e.g., four) cards, some of which are drawn from the original TAT and some of which (e.g., a photograph of a schoolboy at a desk with a book in front of him) are drawn from other sources. Each of these stimuli was selected by McClelland et al. (1953) to "pull" for achievement motivation. Examinees' written stories are coded according to a detailed and explicit scoring scheme.

McClelland, Koestner, and Weinberger (1989; see also McClelland, 1980) asserted that the TAT (as well as other projective techniques) assesses implicit motives, viz., needs of which the respondent are not consciously aware. In contrast, they contended that self-report instruments assess self-attributed motives, viz., needs of which the respondent are consciously aware. McClelland and his colleagues posited that TAT and self-report indexes of needs should therefore correlate negligibly. Moreover, McClelland (1980) hypothesized that TAT and self-report instruments should correlate with different outcome variables. Liberally adapting terminology introduced by Skinner (1938), McClelland maintained that TAT-based indexes of needs should best predict operant behavior, viz., behavior that is not highly constrained by environmental variables. In contrast, self-report instruments should best predict respondent behavior, viz., behavior that is elicited by highly structured stimuli. Thus, TAT indexes of achievement motivation should correlate most highly with long-term achievement (e.g., occupational success), whereas self-report indexes of achievement motivation should correlate most highly with performance on highly structured laboratory measures of achievement (e.g., anagram tasks). Finally, McClelland et al. (1989) distinguished between task and social incentives. The former (e.g., task difficulty) are intrinsic to achievement-oriented tasks, whereas the latter (e.g., achievement-oriented instructions from an experimenter) are extrinsic to such tasks. McClelland et al. hypothesized that task incentives should interact statistically with implicit motives of achievement (i.e., motives derived from the TAT), whereas social incentives should interact statistically with self-attributed motives of achievement (i.e., motives derived from self-report instruments). As we will soon see, a number of researchers have endeavored to test these hypotheses.

Reliability. The reliability of TAT indexes of achievement motivation has been a longstanding bugbear for proponents of these techniques. Although the interscorer reliabilities of these techniques have typically been in the .80 to .90 range (Fineman, 1977), their internal consistency and test-retest reliability have been notoriously problematic. In an influential critique of the TAT, Entwisle (1973) concluded on the basis of numerous published studies that the internal consistency of the McClelland et al. (1953) scoring scheme (as measured by Cronbach's alpha) rarely exceeded .30 to .40. There has been little subsequent data to challenge Entwisle's conclusion, although some authors have questioned the relevance of internal consistency statistics for evaluating the TAT's reliability. For example, Cramer (1999) asserted that "measures of reliability based on internal consistency . . . are not appropriate for the TAT. TAT cards are not the same as a series of items on a personality scale, all of which are intended to measure the same personality trait" (p. 89). Nevertheless, this argument undermines the rationale for aggregating responses to different TAT items into a total score, which assumes that each response is a fallible but useful indicator of the latent construct assessed by this total score (Epstein, 1979).

Moreover, the test-retest reliabilities of TAT-based achievement indexes over intervals of several weeks have generally been in the .30 range (Entwisle, 1973; Fineman, 1977; Winter & Stewart, 1977). Winter and Stewart (1977; see also Cramer, 1999) contended that the low test-retest reliabilities of TAT-based need indexes are artifactual. Specifically, they maintained that on retesting respondents often feel obliged to create different stories. To examine this possibility, Winter and Stewart (1977) used the TAT to assess the test-retest reliability of undergraduates' need for power (Winter, 1973), a motive to be discussed subsequently, when given instructions to write stories that were either unique or as similar as possible to their

earlier stories. Need for power test-retest correlations were significantly higher in the latter ($r = .61$) than in the former ($r = .27$) condition. Nevertheless, Kraiger, Hakel, and Cornelius (1984) failed to replicate Winter and Stewart's results in a sample of undergraduates, instead finding higher correlations in the unique condition ($r = .52$) than in the similar as possible condition ($r = .38$). Disturbingly, the test-retest correlation in an additional condition in which respondents were given no explicit instructions on retesting was essentially zero ($r = .02$). Serious problems concerning the test-retest reliability of TAT-based need indexes thus appear to be largely unresolved (cf. McClelland et al., 1989).

Validity. Many investigators have examined the construct validity of TAT indexes of achievement motivation. In a meta-analysis of correlations ($n = 36$) between TAT-based achievement motivation indexes and self-report achievement motivation indexes, Spangler (1992) found a mean correlation of .09. This correlation, although statistically significant, is low in magnitude and provides support for McClelland's (1980) contention that projective and self-report indexes of needs are not assessing the same construct (but see Emmons & McAdams, 1991, for data suggesting moderate positive correlations between TAT and self-report need indexes). Some authors, in contrast, have argued that the low correlations between projective and self-report indexes of achievement motivation suggest that the former possess poor convergent validity (Entwisle, 1972; Fineman, 1977).

A number of studies, however, suggest that these two sets of instruments correlate with different outcome variables, as predicted by McClelland's (1980) operant-respondent distinction. Spangler (1992) conducted a meta-analysis of 105 studies examining the relations between behavioral outcomes and both TAT and self-reported indexes of achievement motivation. The mean correlations between TAT achievement motivation indexes and operant (e.g., occupational success, income) and respondent (e.g., school performance, measured intelligence) outcome measures were .22 and .19, respectively, whereas the mean correlations between self-report achievement motivation indexes and these two classes of outcome measures were .13 and .15, respectively. Both the operant and respondent correlations were slightly but significantly higher for the TAT indexes than for the self-report indexes, contradicting previous claims (e.g., Fineman, 1977) that the latter possess superior validity for achievement-related outcomes. Nevertheless, all of these mean correlations are low in magnitude. In addition, Spangler found a significant interaction between TAT achievement motivation indexes and task incentives, as predicted by McClelland et al.'s (1989) model, although the predicted interaction between self-report achievement motivation indexes and social incentives did not materialize. Several other findings lend support to the construct validity of TAT achievement motivation indexes. For example, high scores on these indexes tend to be associated with participants' selection of tasks of moderate difficulty (see Fineman, 1977), as predicted by a large theoretical and empirical literature on the level of aspiration construct (McClelland, 1951).

Despite these modestly encouraging results, unresolved questions remain. In particular, the potential confounding role of intelligence, which was not examined in Spangler's (1992) meta-analysis and which has been examined in few studies of TAT achievement motivation, requires clarification (see Entwisle, 1972). Because TAT indexes of the need for achievement tend to be moderately and positively correlated with IQ (Caplehorn & Sutton, 1965; Hundal & Jerath, 1972), and because the operant outcomes (e.g., income) that correlate with these TAT indexes are positively correlated with intelligence (Willerman, 1979), intelligence should be routinely examined as a covariate in future studies of achievement motivation.

Other TAT-based need scoring schemes. A number of other TAT-based need scoring schemes have been developed, among which Winter's (1973) system for assessing need for power has been the most influential. Koestner, Weinberger, McClelland, and Healey (1988; cited in McClelland et al., 1989) presented undergraduates with a social perception task developed by Sternberg (1986) consisting of photographs of two individuals in an job-related setting, and asked them to determine which of these individuals was the boss. Scores on a TAT-derived index of the need for power were significantly related to success on this task. Need for power scores derived from the TAT have also been found to be significantly associated with occupational success among a sample of managers at the American Telephone and Telegraph Company 8 and 16 years after their initial hiring (McClelland & Boyatsis, 1982). Winter, John, Stewart, Klohnen, and Duncan (1998) reported statistically significant interactions between the need for power and extraversion in predicting important life outcomes (e.g., choice of careers affording influence over others) in two longitudinal studies of female college students.

TAT indexes have been developed for other needs, including need for affiliation (Atkinson, Heyns, & Veroff, 1954). In a study of participants who were beeped randomly throughout the day, a TAT index of need for affiliation was significantly associated with how often participants were speaking with other people (see McClelland, 1985). Further support for TAT-based need indexes derives from investigations of biological variables. Students high in TAT-assessed need for power show greater releases of norepinephrine after a major examination than students low in the need for power (McClelland, Ross, & Patel, 1985). Moreover, individuals high in TAT-assessed (but not self-reported) need for affiliation show increases in dopamine release in response to a romantic film (McClelland, 1989). This finding provides construct validity for this TAT index because dopamine has been linked to the experience of reward (Depue & Collins, 1999). Nevertheless, the incremental validity of TAT measures of needs for power and affiliation

above and beyond self-report indexes of these needs has received little or no investigation.

The Assessment of Object Relations with the TAT

Over the past decade, Westen and his colleagues have embarked on an ambitious research program designed to assess object relations (i.e., the mental representation of other people) from TAT protocols. They have been especially interested in assessing aspects of object relations that are relevant to psychopathological conditions, such as borderline personality disorder (BPD), that have been posited some authors (e.g., Kernberg, 1985) to result from early disturbances in parent-child relationships. Based on a subset (typically four to seven) of TAT cards, Westen and his colleagues (e.g., Westen, Lohr, Silk, Gold, & Kerber, 1990) constructed a detailed scoring scheme, the Social Cognition and Object Relations Scale (SCORS), which assesses four dimensions of object relations: (a) Complexity of Representations of People, (b) Affect-tone of Relationships, (c) Capacity for Emotional Investment in Relationships and Moral Standards, and (d) Understanding of Social Causality. This latter scale assesses understanding of the causes of others' thoughts, emotions, and behaviors. Each dimension is scored on a 1-5 scale, with 5 putatively representing the most developmentally advanced set of responses. More recently, Westen (1995) revised the SCORS by subdividing the third dimension into two dimensions of Relationships and Moral Standards, and by adding three dimensions of Aggression, Self-esteem, and Identity and Coherence of Self. In the latest version of the SCORS, responses are scored on a 1-7 scale. There appears to be considerable variability in the TAT cards administered across published investigations of the SCORS (e.g., Barends et al., 1990; Ordnuff & Kelsey, 1996), rendering generalizations across studies difficult because the SCORS may be useful when one set of cards is administered but not another. Similarly, optimal cut-off scores may differ depending on which subset of cards is used. Moreover, adequate norms for the SCORS in general population samples are not yet available.

Reliability. Although interrater reliabilities for the dimensions of the SCORS have been high (e.g., Westen, Ludolph, Lerner, Ruffins, and Wiss, 1989, reported intraclass correlations of approximately .90 for most dimensions), their internal consistencies have been less impressive. Westen, Lohr, Gold, and Kerber (1990), for example, reported internal consistencies (Cronbach's alphas) ranging from .59 to .77 across several clinical samples. In a sample of 96 undergraduates, the internal consistency of the Affect-tone scale was reported to be .56 (Barends et al., 1990). Although the four scales of the original SCORS are positively intercorrelated (e.g., Ornduff, Freedenfeld, Kelsey, & Critelli, 1994, reported significant scale intercorrelations ranging from .30 to .73), it is not known whether the SCORS total score corresponds to a single higher-order dimension. We are unaware of published test-rest reliability studies on the SCORS. Test-retest reliability studies of this scoring method, which should ideally include alternate forms, therefore appear warranted.

Validity. The SCORS dimensions have demonstrated encouraging construct validity in several investigations of differential diagnosis. Westen, Lohr, et al. (1990) used the SCORS to compare 35 patients diagnosed with BPD, 25 patients diagnosed with major depressive disorder, and 30 normals; all groups were matched on sex. BPD patients scored significantly lower on all four original SCORS scales than normals, and significantly lower than the major depressive group on the Affect-tone (suggesting more malevolent object representations) and Understanding of Social Causality scales. Westen, Ludolph, et al. (1990) extended these findings to a study of 33 adolescent BPD patients, 21 adolescent psychiatric patients with mixed diagnoses, and 31 normals; all groups were matched on age, sex, and race. As predicted, BPD patients scored significantly lower than the other groups on the Affect-tone scale, with differences in the medium to large range (d = .55 for difference from non-borderline psychiatric patients, d = .68 for difference from normals). In addition, BPD patients scored significantly lower than the normal group, but not the mixed psychiatric group, on the Capacity for Emotional Investment and Understanding of Social Causality scales. The differences between the BPD patients and normals were in the medium range (ds = .60 and .59, respectively). Contrary to prediction, however, BPD patients scored significantly higher than the mixed psychiatric group, but not the normal group, on the Complexity of Representations scale. Moreover, this difference was medium in magnitude (d = .59). This finding calls into question the construct validity of the Complexity of Representations Scale, although Westen and his colleagues have argued that this finding might point to the presence of a subset of BPD patients with highly differentiated object representations. Interestingly, Westen, Lohr, et al. (1990) had found marked heterogeneity in this variable among BPD patients, with approximately half exhibiting complex object representations on at least one TAT card. The capacity of the SCORS to differentiate BPD patients from either normals or other psychiatric patients has been replicated by investigators not affiliated with Westen's research team (e.g., Gutin, 1997; Malik, 1992).

Supportive findings for the revised version of the SCORS were reported in a recent investigation of 58 patients with BPD, antisocial personality disorder (ASPD), narcissistic personality disorder (NPD), and Cluster C (anxious, fearful) personality disorders (Ackerman, Clemence, Weatherill, & Hilsenroth, 1999). As predicted, BPD patients exhibited the lowest Affect-tone scores of the 4 groups, with the differences from patients with NPD and Cluster C disorders reaching significance. In addition, BPD patients obtained significantly lower scores on the Identity and Coherence of Self variable

than NPD patients, which is consistent with the identity disturbance often believed to be central to BPD (Kernberg, 1985). Nevertheless, several group differences raise serious questions concerning the construct validity of certain indexes derived from the SCORS. For example, patients with ASPD scored nonsignificantly higher than the BPD group on the Moral Standards variable and did not differ significantly from any of the other groups on this variable, despite the fact that individuals with ASPD are typically characterized by a weak or ineffectual conscience. Similarly paradoxical results were found for the Aggression variable, despite the centrality of aggression to the ASPD diagnosis (American Psychiatric Association, 1994).

Several investigators have examined the validity of the SCORS using other external criteria. In a sample of 96 undergraduates, Barends et al. (1990) found that the SCORS Affect-tone scale correlated significantly but modestly ($r = .23$) with affect tone as assessed by a semi-structured interview, as well as with a self-reported index assessing faith in people (r also $=$ .23). Nevertheless, a number of other predicted correlates of the SCORS Affect-tone scale, such as a Rorschach indicator of the perceived malevolence of others (Blatt, Wein, Chevron, & Quinlan, 1979) and a self-report index of social adjustment, were not statistically significant. There is also some support for the use of the SCORS in the assessment of the impact of early adverse environmental events. In a sample of 36 hospitalized female adolescents, Westen, Ludolph, Block, Wixom, and Wiss (1990) found that the Affect-tone scale was negatively and significantly correlated ($r = -.46$) with the number of mother surrogates. In addition, the proportion of poorly differentiated responses on the Complexity of Representations scale was positively and significantly correlated ($r = .70$) with the self-reported duration of sexual abuse among the 12 patients in this sample who had reported a history of such abuse. Nevertheless, a number of predicted correlations between SCORS dimensions and various early childhood risk factors (e.g., number of early moves) were not significant, although low statistical power may have precluded the detection of some genuine associations. Ordnuff et al. (1994) examined the capacity of the SCORS to detect sexual abuse in a sample of 17 female children with a documented history of sexual abuse and 25 female children with no abuse history. Mean SCORS levels were significantly lower in the former than the latter group ($r = .40$), although separate significance tests were not reported for individual SCORS dimensions.

Thus, the SCORS appears to be significantly associated with certain psychopathological conditions, particularly BPD, and perhaps the impact of early adverse experiences. Nevertheless, several issues concerning the validity of the SCORS warrant further examination. First, several of the predicted correlates of the SCORS dimensions run counter to prediction. In particular, findings suggesting (a) more complex object representations among BPD patients than other patients and (b) relatively low or unremarkable levels of immorality and aggression among ASPD patients than patients with other personality disorders require replication. If these findings can be replicated, it will be necessary for Westen and his colleagues to demonstrate that BPD patients and ASPD patients demonstrate these attributes on other psychological instruments, or for them to offer a revision of current conceptualizations of these conditions.

Second, measured intelligence has been found to be positively and significantly correlated with the Complexity of Representations scale ($r = .33$), whereas SCORS word count (i.e., the total number of words produced by respondents) has been found to be positively and significantly correlated ($r = .34$ and .29) with the Affect-tone and Understanding of Social Causality scales, respectively (Ordnuff et al., 1994; but see Ordnuff & Kelsey, 1996, for a different pattern of correlations). The extent to which these covariates account for positive findings on the SCORS is unclear. Ornduff et al. (1994) found that the differences in SCORS between abused and non-abused children remained significant after controlling for general intelligence and word count. In contrast, Westen, Ludolph, et al. (1990) found that controlling for both word count and verbal intelligence eliminated significant differences between the BPD group and other groups on two of three SCORS scales, although controlling for word count did not (the authors did not report analyses controlling for verbal intelligence alone). Measures of verbal intelligence and word count should therefore be routinely examined as covariates in studies of the SCORS so that the potential confounding role of verbal ability can be clarified.

The Assessment of Defense Mechanisms with the TAT

Drawing on the work of Vaillant (1977) and others on the developmental progression of defense mechanisms, Cramer (1991) constructed the Defense Mechanisms Manual (DMM), a TAT-based index of the defense mechanisms of denial, projection, and identification. According to Vaillant, defense mechanisms can be arrayed along a continuum of developmental maturity, with some mechanisms (e.g., denial) being immature, others (e.g., projection) being more advanced, and others (e.g., identification) being still more advanced.

In Cramer's approach, participants are administered several TAT cards, and their stories are scored for the presence of numerous characteristics hypothesized to reflect the presence of one of these three defense mechanisms (see Cramer, 1991). As in the case of the SCORS, there is considerable variability across published studies in the number of TAT cards and the specific cards used, with some studies based on only two cards (one of which is not in the original TAT set; see Cramer, 1997) and other studies based on six cards from the original TAT (e.g., Porcerelli, Thomas, Hibbard, & Cogan, 1998). As a result, generalization across studies is potentially problematic and population norms for the DMM are not available.

Reliability. Interrater reliabilities for the DMM defenses have been adequate, although not extremely high. Across 17 different samples, median interrater reliabilities as assessed by Pearson correlations were .81, .80, and .64 for denial, projection, and identification, respectively (see Cramer & Block, 1998). The relatively low interrater reliability for identification may reflect the greater complexity involved in scoring this defense mechanism (Cramer, 1991). It is also worth noting that unlike intraclass correlations, Pearson correlations can overestimate the magnitude of interrater agreement because they can be high even when raters differ markedly in the absolute elevations of their ratings. The internal consistency and test-retest reliability of the DMM have been even more troublesome. Cramer (1991) reported Cronbach's alphas of .57, .63, and .83 for the denial, projection, and identification scales, respectively, in a sample of 40 undergraduates. In addition, she reported alternate form test-retest reliabilities (over a 2 to 3 week interval) of .07, .30, and .41 for these three scales, respectively, in a sample of 32 6th graders. The corresponding test-retest correlations (again over a two to three week interval) for a sample of 32 2nd graders were .46, .24, and .24. Nevertheless, because children in both samples underwent either success or failure feedback after the initial test administration, these figures may underestimate the DMM's typical test-rest reliability. Further investigation of the DMM's alternate-form reliability is clearly necessary.

Validity. Cramer and her colleagues have examined the construct validity of the DMM in several ways. First, they have explored the capacity of the DMM to differentiate among individuals at different ages. Cramer (1987) administered the DMM to four groups of school children with mean ages of 5.8, 9.1, 14.6, and 16, respectively. As predicted by the hypothesized developmental course of defense mechanisms (Valliant, 1977), there were significant decreases in denial from the first age group onward and significant increases in identification from the second group onward. In addition, projection peaked significantly in the second and third age groups. In a study of 2nd, 5th, 8th, and 11th grade students and college students, Hibbard et al. (1994) similarly found decreases in denial from the 2nd grade onward. Nevertheless, there was an unanticipated and significant increase in denial in the 11th grade. Contrary to Cramer's (1991, p. 34) developmental hypotheses, which predict an increase in projection from ages 7 to 10, projection showed a pattern of decline across all grade levels. Finally, as predicted, identification increased across all grade levels. The findings of Hibbard et al. provide mixed support for the construct validity of the DMM, although it is instead possible that Cramer's developmental hypotheses are partially in error.

Cramer and Block (1998) extended Cramer's previous work on the development of defense mechanisms by examining the validity of the DMM in a sample of 90 nursery school children evaluated at ages 3 and 4 and again at age 23. Personality ratings in nursery school were completed by a set of teachers using the California Q set (Block & Block, 1980), and DMM scores were obtained from participants approximately 2 decades later. The analyses revealed that the inappropriate use of denial in early adulthood was predicted by moodiness, stress reactivity, and poor impulse control in nursery school, but only among males. There were few or no consistent childhood predictors of the use of adult projection or identification in either sex. Cramer and Block's findings are difficult to interpret given that they made few explicit predictions concerning the specific Q-sort correlates of DMM scores (see p. 160).

Hibbard et al. (1994) examined the capacity of the DMM to differentiate 29 psychiatric patients at a Veterans Administration hospital from 40 undergraduates. In contrast to the methods used in Cramer's studies, participants' scores for each defense were computed as a percentage of their total defense scores. The differences between groups in their use of either denial or projection were nonsignificant and negligible in magnitude. Identification, in contrast, was significantly higher among undergraduates. These findings provide mixed and largely negative evidence for the construct validity of the DMM, although Hibbard et al.'s ipsative method of computing defense scores, which eliminated between-subject differences in the total use of defense mechanisms, may have attenuated between-group differences.

Finally, Cramer and her colleagues have examined the extent to which DMM scores increase after stressful experiences, as would be predicted by psychodynamic models of defense mechanisms (see Cramer, 1999, for a review). For example, Cramer and Gaul (1988) randomly assigned 64 elementary school children (in the 2nd and 6th grades) to receive either success or failure feedback following a perceptual-motor task. For reasons that are unclear, the authors administered two TAT cards prior to feedback and three TAT cards following feedback. As predicted, the use of denial and projection, but not the more developmentally advanced defense of identification, was significantly greater in the failure than in the success condition. Nevertheless, inspection of means reveals that although the use of denial increased after failure feedback, the use of projection actually decreased slightly. The significant difference in the use of projection between success and failure conditions seems to have been due to a marked decrease in the use of projection following success feedback.

Dollinger and Cramer (1990) examined the use of defense mechanisms in a sample of 29 adolescent males who had witnessed a lightning strike that killed one boy and injured several others. They found that boys who obtained higher scores on all three DMM defense mechanisms, particularly projection, exhibited significantly lower levels of psychological symptoms (e.g., fears, somatic complaints) than other children. These findings were interpreted by Dollinger and Cramer (see also Cramer, 1999) to mean that more defensive children were better able to protect themselves from the psychological impact of the trauma. Nevertheless, the implications of these results for

the DMM's construct validity are unclear. Because Cramer (1991) argued that a variety of forms of psychopathology are positively associated with the use of defense mechanisms, it would seem that almost any pattern of findings (e.g., greater or lesser use of defense mechanisms associated with psychopathology) could be interpreted as supporting the validity of the DMM.

In summary, the evidence for the construct validity of the DMM is at best mixed. Cramer's (1991) hypothesized developmental progression of defense mechanisms has been only partly supported, the relation between childhood personality problems and the use of defenses in early adulthood has not been corroborated across sexes, and associations between defense use and psychopathology have been inconsistent or difficult to interpret. Moreover, with the possible exception of a study examining the developmental course of defense mechanisms (Porcerelli et al., 1998), there seem to be no replications of the DMM's correlates by independent researchers. Further investigations of the DMM should focus on generating more clearly falsifiable predictions concerning the relations between defense mechanism use and relevant outcome variables. In addition, a standardized set of DMM cards should be developed and used across studies so that results can be more meaningfully synthesized across different investigations and adequate norms can be developed.

Summary and Discussion

There is modest support for the construct validity of several TAT scoring schemes, particularly those assessing need for achievement and object relations. Nevertheless, a number of unresolved issues, particularly potential confounds with intelligence and a lack of stimulus standardization across studies, require additional attention. The use of the TAT to assess defense mechanisms has received limited and inconsistent support. A number of other potentially useful TAT scoring schemes have been developed in recent years. For example, Ronan and his colleagues (Ronan, Colavito, & Hammontree, 1993; Ronan et al., 1996) have derived an index of personal problem-solving from the TAT that correlates significantly with a performance measure involving the generation of means-end solutions to problems, and that significantly distinguishes psychiatric patients from normals. In addition, scores on this index have been found to increase significantly following training in the generation of alternative solutions (Ronan, Date, & Weisbrod, 1995).

Even the few promising TAT scoring systems, however, are not yet appropriate for routine clinical use. For all of these systems, (a) adequate norms are not available, (b) test-retest reliability is either questionable or unknown, (c) field reliability (Wood et al., 1996a) is untested, and (d) there is almost no research to ensure that such systems are not biased against one or more cultural groups. In addition, there is no convincing evidence that TAT scoring schemes for object relations or defense mechanisms possess incremental validity above and beyond self-report indexes of these constructs.

Adequate TAT norms are needed so that clinicians will not overdiagnose psychopathology. In a classic study, Little and Schneidman (1959) asked psychologists to rate normal individuals on a 1-9 scale of maladjustment, with 9 indicating severe psychopathology. Psychologists were more likely to perceive psychopathology in normal individuals when ratings were based on the TAT (mean = 4.8) than when ratings were based on either a case history (mean = 1.6) or the MMPI (mean = 3.3). More recently, Murstein and Mathes (1996) found that a measure of vocabulary was significantly correlated (r = .36) with TAT-rated pathology in a nonclinical sample. Other analyses in this sample and in a sample of psychiatric patients revealed positive correlations between vocabulary and a TAT-derived index of projection, which assessed the extent to which respondents revealed personally relevant material. Although these findings are open to several interpretations, one possibility is that verbose respondents will tend to be judged as more pathological. This bias may be especially pronounced when using impressionistic scoring of the TAT, which is by definition conducted without reference to normative data.

Although there is modest support for the construct validity of several TAT scoring schemes, the relevance of these findings to clinical practice is doubtful, because an overwhelming majority of clinicians rely solely on impressionistic interpretations of the TAT (Hunsley et al., 2000). As Ryan (1985) observed,

> Practitioners interpreting the TAT are likely to use different systems, an idiosyncratic combination of systems, or no system at all. This is the bane of the psychometrician, and it also suggests that in common usage the interpretation of the TAT is based on strategies of unknown and untested reliability and validity, a potentially dangerous outcome (p. 812).

A final major unresolved issue concerns what might be termed the fungibility of different projective methods. We have seen that several standardized coding schemes for the TAT probably can be characterized as possessing modest construct validity. At the same time, a considerable body of literature indicates that stimulus materials other than the TAT can be used to score the dimensions assessed by the TAT. For example, need for achievement measures obtained from a variety of written materials have demonstrated validity for achievement-related outcomes (McClelland, 1961). Westen and others have found that object relations indexes patterned after the SCORS can be obtained from a number of sources other than the TAT, including early memories and stories told during the administration of the WAIS-R Picture Arrangement subtest. Moreover, like the SCORS, these methods of scoring object relations have been found to distinguish BPD patients from nonpatients and other psychiatric patients (see Barends et al., 1990; Nigg, Lohr, Westen, Gold, & Silk, 1992). As Zeldow and McAdams (1993) concluded in their brief review of the

comparative validities of the TAT and free speech samples, "various forms of narrative speech samples [including the TAT] may be highly correlated, so long as psychologically meaningful, well-validated, and higher-order content categories are used" (p. 181). Thus, although the results reviewed here offer encouraging support for the validity of certain scoring systems derived from the TAT, they leave open the question of whether the TAT (or even any formal projective technique) *per se* is necessary for achieving such validity.

HUMAN FIGURE DRAWING METHODS

The controversy surrounding human figure drawings has been nearly as contentious and polarized as that surrounding the Rorschach. Proponents of these construction techniques (Lindzey, 1959), such as Riethmiller and Handler (1997a), have maintained that "figure drawing tests have enormous potential that should be cultivated" (p. 460) and that "drawings provide something that a series of scores cannot provide" (p. 466). In contrast, detractors have gone so far as to opine that "Approximately a century behind in time, the DAP [Draw-A Person Test] might well be described as phrenology for the twentieth century" (Gregory, 1992, p. 468) and that the human figure drawing method "more properly belongs in a museum chronicling the history of simple-minded assessment practices in school psychology" (Gresham, 1993, p. 185).

Although there is a wide variety of human figure drawing techniques, all require the examinee to draw one or more people. These techniques can be divided into kinetic methods, which ask the respondent to draw people performing an activiity, or nonkinetic methods, which do not (Knoff & Prout, 1985). In contrast with other projective techniques discussed here, most human figure drawing methods can be administered and scored relatively quickly. The mean time for administration of the commonly used Goodenough-Harris Draw-A-Person (DAP) Test (Harris, 1963), is approximately 5 minutes, with another 5 minutes required for scoring by experienced clinicians (Kamphaus & Pleiss, 1991). The scoring time differs considerably, however, across different scoring systems.

Broadly speaking, there are two major approaches to human figure drawing scoring and interpretation. One approach, which we term the sign approach, is rooted largely in the theorizing of Machover (1949) and others, and draws inferences from isolated drawing features (e.g., large eyes). According to Machover (1949), a variety of signs derived from the DAP are associated with specific personality and psychopathological characteristics. For example, Machover linked large eyes to suspiciousness or paranoia, long ties to sexual aggressiveness, the absence of facial features to depression, heavy shading to aggressive impulses, and multiple erasures to anxiety. Machover further hypothesized that the person drawn by the respondent in the DAP embodies the central psychological and physical attributes of the respondent (the "body-image hypothesis").

The global approach, in contrast, stems largely from work by Koppitz (1968), who developed a system for scoring 30 indicators from children's drawings. These indicators are then summed to yield a total maladjustment score. As we will see, there is some evidence that the sign and global approaches differ in their psychometric properties.

Normative data are available for at least some human figure drawing methods. For example, a recently revised version of the DAP (Naglieri, 1988) has been normed on 2622 children (age range = 5 to 17). These children were sampled using 1980 U.S. Census data, with the sample stratified on the basis of age, sex, race, geographical area, ethnicity, socioeconomic status, and size of community (Kamphaus & Pleiss, 1991).

Reliability

The interrater reliabilities of specific human figure drawing signs are quite variable across studies. Kahill (1984) found that the interrater reliabilities of the majority of figure drawing indicators (as measured by various correlational statistics) were above .80, with approximately two-thirds above .70. Nevertheless, some interrater reliabilities reported in the literature have been low (see also Swenson, 1968). For example, in Kahill's review, interrater reliabilities ranged from .45 to .96 for shading, .54 to .99 for head size, and −.13 to .60 for facial expression. In a study of individuals' House-Tree-Person (H-T-P; Buck, 1948) scores, Palmer et al. (2000) reported interrater reliabilities ranging from .01 to 1.0, with a median of .52. Similarly, Vass (1988) reported interrater reliabilities for H-T-P scores that ranged from .27 to .75, with a mean of .54. He concluded that there are "serious reliability and validity problems with the empirical investigations of projective drawing tests" (p. 611). Thus, although some figure drawing indexes possess high interrater reliability, reliability may often be poor. Consequently, acceptable reliability cannot be assumed without corroborating evidence. Moreover, there is relatively little evidence regarding the inter-rater reliabilities of clinicians' interpretations of figure drawing characteristics (Thomas & Jolley, 1998). Because many interpretations do not flow directly or inexorably from figure drawing scores, this type of reliablity must be investigated separately, even for characteristics that can be reliably scored.

The test-retest reliabilities of figure drawing indexes have also been variable across studies. For global indexes (e.g., overall body image, overall drawing quality), Swenson (1968) reported test-retest reliabilities ranging from .74 to .90 across nine studies, and Kahill (1984) reported reliabilities ranging from .81 to .99 across four more recent studies. There is evidence, however, that the test-retest reliabilities of specific drawing features are sometimes problematic. Swenson (1968), for example, reported test-retest reliabilities of .54 for omissions and test-retest reliabilities ranging from .21 to .85 for the height of drawn figures (see also Hammer & Kaplan, 1964, and

Thomas & Jolley, 1998, for data indicating questionable temporal stability for height).

The internal consistencies of human figure drawing global indexes have generally been acceptable, although some have been only moderate. The median internal consistencies (Cronbach's alphas) of a recently revised version of the DAP (Naglieri, 1988) were .86 for the total score and .70 (range of .56 to .78) for each drawing scored separately. Naglieri, McNeish, and Bardos (1992) reported Cronbach's alphas of .76, .77, and .71 across three age groups of children and adolescents (age range of 6 to 17) for the Draw-A-Person: Screening Procedure for Emotional Disturbance (DAP: SPED), a 55-item DAP scoring system designed to identify children and adolescents with emotional difficulties. In a sample of undergraduates, Groth-Marnat and Roberts (1998) reported Cronbach's alphas of .76 for the H-T-P total score and .69 and .50 for scores derived from male and female DAP figures.

Validity

As with the TAT, a major obstacle to evaluating the validity of human figure drawings is the fact that many of the hypotheses generated by investigators seem difficult to falsify. For example, in attempting to explain negative findings for certain DAP signs, Hammer (1959) argued that in contrast to normals, pathological individuals can draw figures that are either too small or too large, draw lines that either too heavy or too light, or produce either too many or too few erasures. Although Hammer's (1959) speculations (p. 31) imply (a) a bimodal distribution of certain DAP indexes in pathological, but not normal, groups and (b) higher levels of variance (and other measures of dispersion) of these indexes in pathological than normal groups, we are unaware of any systematic effort to test these hypotheses (but see Joiner & Schmidt, 1997). Handler and Reyher (1965, p. 308) similarly contended that shading, line reinforcement, and erasures can reflect either the presence of anxiety or the presence of successful coping efforts against anxiety (and therefore the absence of overt anxiety). More recently, Waehler (1997) asserted that "We should not always be dissuaded by negative findings" because "sometimes a drawing might not be the medium through which people choose to communicate their distress" (p. 486). Nevertheless, Waehler did not explicate how to predict which medium of expressing distress respondents will select.

These caveats concerning the difficulty of falsifying investigators' predictions notwithstanding, an enormous body of research has examined the validity of specific human figure drawing signs. Beginning with Swenson (1957), a parade of reviewers over the past four decades has converged on one virtually unanimous conclusion: the overwhelming majority of human figure drawing signs possess negligible or zero validity (Kahill, 1984; Klopfer & Taulbee, 1976; Motta, Little, & Tobin, 1993; Roback, 1968; Suinn & Oskamp, 1969; Swenson, 1968; Thomas & Jolley, 1998). In particular, published research offers very little support for Machover's (1949) DAP signs of personality and emotional disturbance. In a "box score" review of the published literature from 1967 to 1982, Kahill (1984), for instance, found support for only 2 of 30 Machover indexes reviewed: round (as opposed to square) torsos as an indication of feminine personality features and colored drawings as an indication of anxiety. In contrast, and contrary to Machover's (1949) hypotheses, studies revealed no consistent relationships between ear emphasis and paranoia; internal organs and schizophrenia; inanimate drawing props (e.g., guns, knifes) and delinquency; and hair emphasis and sexual concerns, among many other purported associations. In addition, studies yielded mixed and largely negative findings concerning Machover's body-image hypothesis. For example, Viney, Aitken, and Floyd (1974) reported no significant differences in height, waist width, or breast width between pregnant and non-pregnant women (but see Tolor & Digrazia, 1977), and Thomas, Jones, and Ross (1968) reported no significant correlations between figure size and the height, weight, or girth of the drawer. Broadly similar conclusions regarding Machover's hypotheses were reached in box score reviews of the earlier literature [e.g., Roback, 1968; Swenson, 1968; see also Handler & Habenicht, 1994, for a review of the validity of the Kinetic Family Drawing Test (KFD); Burns & Kaufman, 1970].

Of specific drawing signs, size of figure has been among the most extensively investigated. There is some suggestion that overall drawing size is related to the perceived likeability or importance of the drawn figure. For example, there is evidence that the size of Santa Claus in children's drawings increases as Christmas approaches (Craddick, 1961; Sechrest & Wallace, 1964). Nevertheless, these findings may be due to the tendency of children to see more pictures and photographs of Santa Claus at Christmastime, which could lead them to produce larger and more detailed drawings (Thomas & Jolley, 1998). Moreover, although Thomas, Chaigne, and Fox (1989) reported that children drew a man described as "nasty" as smaller than a neutral control man, these findings have not been consistently replicated (Jolley, 1995).

Results from a recent study on size, level of detail, and line heaviness were also negative. Joiner, Schmidt, and Barnett (1996) examined the relations among these variables derived from the KFD and the Kinetic-House-Tree-Person Drawings (Burns, 1987) in 80 psychiatric inpatient children (age range = 6 to 16 years). The latter projective technique asks respondents to "Draw a house, a tree, and a person, all in the same picture, with the person doing something." Although interrater scoring reliabilities were high (range = .91 to .95), none of the three indicators was significantly related to self-report indexes of depression or anxiety. For example, drawing size, which has been found in some studies to be negatively associated with depression (e.g., Lewinsohn, 1964), was nonsignificantly correlated at $r = -.10$ with a self-report index of depression. Amount of detail, which has been posited to be negatively

associated with anxiety (Handler, 1967), was positively, although nonsignificantly, correlated at $r = .12$ with a self-report index of anxiety. Moreover, the correlations between these figure drawing indexes and both depression and anxiety indexes derived from another projective instrument (the Roberts Apperception Test for Children; McArthur & Roberts, 1960) were virtually all nonsignificant and low in magnitude.

It is possible, of course, that certain drawing signs possess slight validity for features of personality and psychopathology that were obscured by the box score method used by Kahill (1984) and previous reviewers. Because they do not take statistical power into account, box score reviews tend to err toward mistakenly negative conclusions (Schmidt, 1992). Consequently, we suspect that questions concerning the validity of human figure drawing signs will not be conclusively settled until meta-analyses of research on these signs are conducted. An overwhelmingly negative box score does, however, increase the likelihood that any effect, if present, is small in magnitude. Moreover, because a number of findings concerning the signs of Machover (1949) and others have been in the direction opposite from those predicted, it is unlikely that negative findings regarding human figure drawing signs are due entirely to low statistical power. For example, Dudley, Craig, Mason, and Hirsch (1976) found that depressed individuals were less likely than nondepressed individuals to draw faces in profile, directly contradicting one of Machover's hypotheses. Again contra Machover, Cvetkovic (1979) reported that schizophrenics were less likely than normals to draw disembodied heads. Moreover, in a review of research on human figure drawing signs and anxiety, Handler and Reyher (1965) found that 30 of 255 findings were statistically significant in the direction opposite from those predicted (see Riethmiller & Handler, 1997a, for discussions of these negative findings).

Some authors have responded to these negative findings by maintaining that clinicians in actual practice rarely, if ever, use isolated drawing signs. For example, Riethmiller and Handler (1997a) argued that reliance on specific figure drawing indicators "is definitely not congruent with the way in which most clinicians use the DAP" (p. 467). We are unaware of any research directly examining the modal uses of drawing methods in clinical practice. Nevertheless, a study by Smith and Dumont (1995) raises serious concerns regarding clinicians' overreliance on DAP signs. These authors provided a sample of 36 clinical and counseling psychologists (58% of whom had received formal training in the use of projective techniques) with DAP protocols, and tape-recorded their comments as they interpreted these protocols. Of 22 practitioners who used the DAP to draw clinical inferences, 20 based at least some of their inferences on specific signs. Among the statements made by experts with training in human figure drawing tasks were:

> "His eyes are strange and overemphasized. I think he may have problems with men, with some paranoid suspiciousness"; "The only thing that's curious is how broad the shoulders are, which indicates that he feels he's carrying a terrible and heavy load"; and "There are indications for (sic) dependency, lots of buttons and buckles" (Smith & Dumont, 1995, p. 301).

Although it would be inappropriate to generalize these findings beyond Smith and Dumont's (1985) sample, they raise the possibility that many clinicians use the sign approach in real world settings despite compelling evidence against its validity. In particular, many clinicians may rely on figure drawing signs (e.g., large eyes) that bear a strong semantic or associative connection with psychopathology (e.g., paranoia). Classic research by Chapman and Chapman (1967) demonstrates that clinicians are often convinced of the validity of such figure drawing signs despite the pronounced lack of evidence for their validity (a phenomenon known as "illusory correlation").

Is there any silver lining to the large black cloud of research evidence looming over the human figure drawing literature? Perhaps. There is suggestive evidence that global approaches can achieve modest validity. It has long been known, for example, that poor overall quality of figure drawings is a rough barometer of psychopathology (Swenson, 1957). More recently, evidence from controlled studies suggests that certain global scoring approaches to figure drawings may distinguish individuals in certain diagnostic groups from normals. Tharinger and Stark (1990) administered the DAP and KFD to 52 children with mood disorders, anxiety disorders, both mood and anxiety disorders, or no disorder. For each figure drawing technique, they examined the validity of two global indexes: a quantitative method based on the Koppitz scoring scheme and a qualitative method based on global judgments of psychopathology (e.g., lack of emotional well-being in the drawn figure, inhumanness of the drawn figure). The quantitative scores did not distinguish significantly among the diagnostic groups. In contrast, the DAP qualititative score significantly distinguished (a) normal children from children with mood disorders and (b) normal children from children with both mood and anxiety disorders. In addition, the KFD qualitative score significantly distinguished normal children from children with mood disorders. Corroborating previous findings (e.g., Kahill, 1984), virtually no diagnostic differences emerged when isolated DAP and KFD signs were used.

Naglieri and Pfeiffer (1992) examined the capacity of a global quantitative scoring system, the DAP: SPED (Naglieri, 1988), to differentiate 54 children and adolescents with conduct and oppositional defiant disorders from 54 normal children and adolescents. DAP: SPED scores of the former group were significantly higher than those of the latter. Moreover, the effect size for this difference was large ($d = .76$). Naglieri and Pfeiffer's findings, in contrast to those of Tharinger and Stark (1990), suggest that certain quantitative scoring systems may be valid for diagnostic purposes. Other authors have reported that the Koppitz and other quantitative scoring systems differentiate adjusted from maladjusted individuals (e.g., Currie, Holtzman, & Swartz, 1974).

Nevertheless, the overall picture for quantitative and qualitative global scoring systems cannot be described as consistently positive. In a sample of 40 undergraduates, Groth-Marnat and Roberts (1998) reported that total scores on the H-T-P and HFD derived from a published quantitative scoring system were not significantly correlated with either of two self-esteem indexes. In addition, in a study already described, Tharinger and Stark (1990) found that a qualitative scoring system for the DAP did not significantly distinguish normal children from children with anxiety disorders, and that a qualitative scoring system for the KFD did not significantly distinguish normal children from either children with anxiety disorders or children with both mood and anxiety disorders.

Incremental Validity

Serious questions can be raised concerning the incremental validity of human figure drawings. In particular, there is reason to question whether the addition of human figure drawing scores to measures of intelligence and artistic ability will lead to increases in validity. With respect to intelligence, total scores on the DAP have generally been found to be moderately correlated with scores on children's IQ measures (median r = .57; Kamphaus & Pleiss, 1991). Indeed, human figure drawings are sometimes used as screening measures for global intelligence, although their relatively modest correlations with IQ measures render this use questionable (Kamphaus & Pleiss, 1991; Motta et al., 1993). Positive correlations with IQ measures have also been reported for scores derived from other human figure drawing techniques. As Knoff (1993) pointed out, "the variance in a HFD [human figure drawing] attributable to 'intellectual maturity' is likely to overlap with the variance related to 'personality'" (p. 191). This overlap is essential to consider in diagnostic studies, because many psychopathological conditions are associated with lower than average IQ. For example, children with conduct disorder, who were examined in Tharinger and Stark's (1990) study, have mean IQs approximately 8 points lower than other children (Wilson & Herrnstein, 1985). In addition, patients with schizophrenia, who have been examined in a number of studies of human figure drawings (see Kahill, 1984), have significantly lower average IQs than normals (Aylward, Walker, & Bettes, 1984). Nevertheless, the incremental validity of human figure drawings above and beyond IQ has rarely been examined (see Kahill, 1984). In one of the few studies to examine this issue (Schneider, 1978), the KFD contributed no significant validity increment in the assessment of the severity of children's school problems above and beyond age and IQ.

Another variable that has received attention in the human figure drawing literature is artistic ability. Although one might legitimately use figure drawings to assess artistic ability, artistic ability is a potential nuisance variable in studies examining the relations of these drawings to personality and psychopathology. Some early proponents of human figure drawing methods asserted that artistic quality was not a major threat to their validity (e.g., Hammer, 1958). Nevertheless, the problem of artistic ability has never been satisfactorily resolved. In an important study, Feldman and Hunt (1958) asked art teachers to rate which body parts were most difficult to draw. They then asked clinicians to rate the drawings of 65 undergraduates for psychological adjustment. Feldman and Hunt reported a significant correlation of $r = -.53$ between the rated difficulty of drawing a given body part and the rated psychological adjustment indicated by that body part, demonstrating that body parts that are more difficult to draw are also more likely to be viewed as reflecting maladjustment. This finding raises the disturbing possibility that examinees with poor artistic skill may often be erroneously labeled as pathological.

Factor analyses of human figure drawing signs by Nichols and Strumpfer (1962), Adler (1970), and Cressen (1975) all revealed that a factor most parsimoniously interpreted as artistic ability accounted for the majority of the variance among quantitatively rated drawing signs. These findings suggest that artistic ability is a major determinant of individual differences in human figure drawings.

Moreover, in the study by Cressen (1975), psychologists' ratings seemed to be influenced by artistic ability. Psychologists were asked to classify participants as schizophrenic or normal. They did not perform better than chance, and tended to make diagnoses of schizophrenia when given drawings of low artistic quality, even when the drawings were done by normals. In addition, there was little association between artistic quality and actual diagnostic status (schizophrenic vs. normal). These findings again suggest that poor artistic quality may lead clinicians to make false positive judgments of pathology. In a similar vein, Carlson, Quinlan, Tucker, and Harrow (1973) found that a factor labeled Body Disturbance derived from DAP protocols correlated significantly (r = .53) with rated artistic ability among a sample of psychiatric patients, even though this factor was not significantly related to psychiatric diagnoses.

It is worth noting that in the aforementioned studies, the relation between artistic quality and actual psychopathology was weak (Adler, 1970; Carlson et al., 1973; Cressen, 1975, Nichols & Strumpfer, 1962). This finding introduces the possibility that artistic quality could be used as a suppressor variable (Conger & Jackson, 1972), as artistic quality often correlates highly with total DAP scores but negligibly with actual psychopathology. This idea receives some indirect support from the study by Cressen (1975), who found that judges' classifications of DAP protocols as pathological or nonpathological improved somewhat when the artistic quality of drawings was held constant. Nevertheless, we are unaware of any investigations that have explicitly incorporated artistic quality as a suppressor variable in predictive equations. Moreover, suppressor variables have proven notoriously difficult to replicate in most psychological domains (e.g., see Greene, 2000,

for a review of research on the use of the K scale as a suppressor variable in the MMPI).

Finally, there are few data bearing on the incremental validity of human figure drawings above and beyond either psychometric or demographic information. Wildman and Wildman (1975) found that adding the H-T-P to the Bender-Gestalt figure drawing test decreased the accuracy of clinicians' classifications (of individuals as either psychiatric patients or nurses) from 62% to 53%. We are unaware of any studies demonstrating that human figure drawings offer psychologically useful information over and above the MMPI-2, psychiatric interviews, demographic data, or other information that is often readily available in clinical settings.

Summary and Discussion

The scientific status of scores derived from human figure drawings can best be described as weak. Although test-retest and interrater reliabilities are sometimes high, there is marked variation across studies. In addition, field reliability has not been studied. Moreover, despite hundreds of investigations, there are no well replicated relationships between specific drawing signs and either personality or psychopathology. Although approaches using global scoring methods have sometimes distinguished psychopathological individuals from normals, these approaches have not been uniformly successful (e.g., Tharinger & Stark, 1990). The role of artistic quality in human figure drawings has not been satisfactorily resolved, although there is reason to believe that poor artistic ability can often result in false positive classifications of psychopathology. Perhaps most important, there is no convincing evidence that human figure drawings possess incremental validity above and beyond other readily available demographic or psychometric data. Unless and until these issues are resolved, there is ample reason to question the continued widespread use of human figure drawings in clinical practice (Gresham, 1993; Motta et al., 1993). Nevertheless, we encourage further research on global scoring approaches, as these systems, in contrast to sign approaches, have displayed modest validity in at least some studies.

Despite the host of negative findings, many proponents of human figure drawing techniques continue to maintain that indexes derived from these techniques possess adequate validity. For example, some proponents aver that these techniques are valid in the hands of qualified clinicians, such as those with high levels of empathy (Scribner & Handler, 1987) or extensive experience with these techniques. However, in studies of human figure drawings, validity has not generally been significantly related to clinical training or clinical experience. For example, Stricker (1967) found that clinicians experienced in the use of figure drawings were significantly less accurate than clinical psychology graduate students when using the DAP to distinguish normality from abnormality. Levenberg (1975) reported no significant differences among doctoral-level practitioners, predoctoral interns, and even hospital secretaries in their levels of success when using the KFD to differentiate normal and abnormal children, although doctoral-level practitioners were slightly more accurate. The overall accuracy rates for these three groups (where chance accuracy was 50%) were 72%, 61%, and 61%, respectively. Because the judges in Levenberg's study had access to respondents' verbal statements concerning the content of their drawings (p. 390), however, these percentages may overestimate the extent to which judges can accurately assess examinees based solely on information contained in their drawings.

Disturbingly, Levenberg found that an expert on human figure drawings who had authored two books on the KFD was markedly less accurate than any of the other three groups, and obtained a hit rate slightly below chance (47%). Additional studies indicate that training and clinical experience are not significantly related to validity when judgments are based on human figure drawings (Cressen, 1975; Hiler & Nesvig, 1965; Wanderer, 1969; see Garb, 1989, 1998, for reviews). In view of evidence that clinicians often attend to invalid figure drawing signs—particularly those bearing a strong semantic or superficial association with features of psychopathology (Chapman & Chapman, 1967)—it is perhaps not entirely surprising that clinicians are no more accurate than individuals without psychological training.

META-ANALYSIS OF PROJECTIVE TECHNIQUES FOR DETECTING CHILD SEXUAL ABUSE

To what extent are projective methods useful in applied settings? To begin to address this complex and important issue, we elected to examine quantitatively the validity of projective techniques in one scientifically and socially important domain, namely the detection of child sexual abuse. We also used this analysis as an opportunity to obtain a preliminary estimate of the file drawer effect (publication bias) in one major psychological domain of research involving projective techniques.

Even though most forensic mental health professionals believe that projective techniques are useful for detecting child sexual abuse (Oberlander, 1995), their validity for this purpose has not been established (Garb, Wood, & Nezworski, 2000; Trowbridge, 1995). For example, although West (1998) conducted a meta-analysis and claimed that projective instruments can be used to detect child sexual abuse, she included only statistically significant results and systematically excluded nonsignificant results (for a critique, see Garb et al., 2000). It is important that we ascertain the validity of projective techniques for this task, because incorrect judgments can cause enormous suffering for children, their families, and those who are wrongly accused.

To determine whether projective methods can be used to detect child sexual abuse at clinically useful levels and to explore the possibility of publication bias in this literature, we

conducted a series of analyses (Garb, Wood, & Lilienfeld, 2000). Although West (1998) located only 12 appropriate studies, we were able to locate 47 studies. Median effect sizes (using the *d* statistic) were calculated for each study, and these effect sizes were aggregated using D-STAT (Johnson, 1994). In addition, results for individual test scores were examined to determine whether positive findings have been consistently replicated by independent investigators.

Results of the meta-analyses are presented in Table 2. Meta-analyses were conducted separately for the Rorschach, TAT, and human figure drawings, the three major techniques examined in this review. Listed are values for *d* and the number of comparisons on which each *d* value is based. For the Rorschach, the average effect size is less than small in magnitude for the comparisons of sexually abused children with nonabused children receiving psychological treatment and is small in magnitude for the comparison of sexually abused children with nonabused children in the community. The effect size is medium-large when sexually abused children are compared with the Exner CS norms. However, as discussed earlier, even normal individuals in the community seem pathological when compared with the CS norms. Thus, the results for the Rorschach are largely negative. The results are better for the TAT and projective drawings. For the TAT, effect sizes range from small-medium to medium-large. For human figure drawings, effect sizes range from small to medium in size.

In an additional meta-analysis, publication bias was examined. As discussed earlier, publication bias, also referred to as the file drawer effect, is present when results in published studies are larger in magnitude than those obtained in unpublished studies. By pooling all of the results across instruments, 19 median effect sizes from published studies and 43 median effect sizes from unpublished studies were gathered. These analyses yielded evidence of publication bias. The average effect size for published studies is $d = .51$, and the average effect size for unpublished studies is $d = .24$. These two values for *d* correspond to correlation coefficients of .25 and .12, respectively. As these figures indicate, studies of projective instruments in this literature were less likely to be published when results were small in magnitude.

Table 2. *Meta-Analytic Results for the Rorschach, TAT, and Human Figure Drawings*

Test	*d*	Number of Comparisons[a]
Rorschach		
Clinical versus CSA groups	.08	8
Normal versus CSA groups	.23	7
CS norms versus CSA groups	.60	5
TAT		
Clinical versus CSA groups	.41	9
Normal versus CSA groups	.57	3
Human Figure Drawings		
Clinical versus CSA groups	.30	13
Normal versus CSA groups	.24	18

Note. CSA = child sexual abuse.
[a]Some studies contained more than one type of comparison.

There are several reasons why publication bias may occur. For example, editors may prefer to accept manuscripts that include statistically significant findings. In addition, investigators may be especially inclined to submit manuscripts that include statistically significant findings, either because they believe these manuscripts are more likely to be accepted or because they believe these results are more important. Finally, it is possible that studies with small or zero effect sizes tend to be of low methodological quality, although this seems unlikely because criterion contamination and the inappropriate comparison of sexually abused children to the CS norms are both likely to lead to inflated effect sizes.

In addition to describing average effect sizes, results for individual test scores were examined to determine whether positive findings have been replicated by independent investigative teams. Specifically, we examined whether positive findings have been replicated for the comparison of sexually abused children and nonabused children receiving psychological treatment. This comparison is particularly important because clinicians who are confronted with the task of detecting a history of sexual abuse typically assess children who are referred for evaluation, treatment, or both. For the Rorschach and human figure drawings, positive findings were never consistently replicated.

For example, in studies of human figure drawings, positive results were reported for the use of the tongue as an indicator of sexual abuse, but these results have not been replicated. According to Drachnik (1994),

> Because of the number of tongues I had seen in the drawings of sexually abused children . . . , I decided to review my collection of drawings that I had accumulated over the past 15 years. [For 43 children] identified as sexually abused, the drawings of 14 children displayed one or more tongues. . . . Of the other 194 clients (not known to be sexually abused) seen over this 15-year period, only two drawings displayed a protruding tongue (p. 60).

Drachnik (1994) also discussed the potential significance of these findings:

> If the tongue is a graphic symbol of sexual abuse in children's drawings, what is its purpose? Could children be using this symbol to work through the sexual abuse? Could they be unconsciously communicating the abuse to the therapist? Or could they be using the symbol as a a protective device (as some cultures relate the tongue to protection as a way to ward off danger) to prevent further sexual abuse (p. 60)?

These conjectures seem premature. Although positive findings were reported by Drachnik (1994; $d = 1.44$), negative findings were reported by Chase (1987; $d = .09$, 34 sexually abused children, 26 nonabused children at a mental health

clinic) and Grobstein (1996; $d = .08$, 81 sexually abused children, 82 nonabused children at a mental health clinic).

Results were slightly better for the TAT. Again, we limit our conclusions to the comparison of sexually abused children with nonabused children receiving psychological treatment. Using the SCORS (Westen, Lohr, Silk, Kerber, & Goodrich, 1985), positive findings were replicated by independent investigators for the Affect-tone scale (Ornduff & Kelsey, 1996; Westen, Ludolph, Block, Wixom, & Wiss, 1990). Positive findings reported for the other three scales comprising the original version of the SCORS (Ornduff & Kelsey, 1996) were not replicated by independent investigators (Westen, Ludolph et al., 1990).

Although positive findings were replicated for the Affect-tone scale, normative data are not available to help clinicians use this scale. To understand why normative data are needed, it is helpful to examine the mean scores obtained for this scale. Mean scores for sexually abused and nonabused children receiving psychological treatment were 2.71 and 3.32 (Ornduff & Kelsey, 1996) and 2.48 and 2.64 (Westen, Ludolph et al., 1990), respectively. Sexually abused children should score lower, but the sexually abused children in the Ornduff and Kelsey (1996) study scored higher (mean = 2.71) than the nonabused children in the Westen, Ludolph et al. (1990) study (mean = 2.64). Thus, even for the only test score that seemed to do well for detecting child sexual abuse, it is not clear what cutoff score should be used for determining that it is likely that a child has, or has not, been sexually abused.

In conclusion, the use of projective techniques for detecting child sexual abuse received relatively little support. In our meta-analysis, average effect sizes for the Rorschach were either small or negligible, except when sexually abused children were compared with the CS norms. However, incorrect judgments will likely be made if CS norms are used to interpret results, because as discussed earlier the CS norms tend to classify too many nonpathological indivduals as pathological. With the exception of the SCORS Affect-tone scale, positive findings for individual projective scores have not been consistently replicated by independent investigators. Moreover, because the prevalence of child sexual abuse is likely to be considerably lower than 50 percent in most settings, the predictive utility of projective indexes in the real world is likely to be lower than that found in the studies we examined, most of which made use of an approximately equal number of abused and nonabused children (see Dawes, 1993).

Finally, our results provide the first clear evidence of publication bias in the projectives literature. Previous meta-analyses of projective techniques have not included results from unpublished studies. Thus, our findings raise important questions concerning all other published meta-analyses on projective techniques. For example, as noted earlier, the mean validity coefficient of $r = .30$ that has been reported in meta-analyses of the Rorschach could represent a subsantial overestimate of this instrument's actual validity. Similarly, our findings raise the possibility that many articles and books paint an overly positive picture of projective techniques because authors are familiar with published but not unpublished results. The file drawer effect must now be carefully considered in evaluating the validity of all projective indexes.

CONCLUSIONS REGARDING EMPIRICALLY SUPPORTED PROJECTIVE INDEXES

Reflecting on the state of current attitudes toward projective techniques in academia, Westen, Lohr, et al. (1990) wrote that "Generations of psychologists, including personality and clinical psychologists, have been trained with a deeply ingrained assumption that projective techniques are inherently invalid and unreliable" (p. 362). As we have seen and will discuss further in this section, it is evident that certain projective instruments, as well as scores derived from these measures, can indeed achieve acceptable levels of reliability and validity. Consequently, dismissing in broad brush all projective techniques as unreliable and invalid is unwarranted.

At the same time, the research literature we have reviewed provides ample justification for skepticism concerning most widely used projective techniques. Many of these techniques yield indexes with negligible or undemonstrated validity, and some proponents of these techniques continue to make strong claims regarding their predictive powers that are not supported by research (e.g., Kubisyzn et al., 2000). Although the use of projective techniques seems to have declined somewhat in recent years (Piotrowski et al., 1998), these techniques continue to be administered in clinical and forensic settings with considerable frequency. Given the limited validity of many of the indexes derived from these techniques, it is virtually inevitable that the inferences routinely drawn from them by practitioners are often unjustified, erroneous, or both. For example, although our meta-analysis demonstrated that the Rorschach performs only slightly better than chance in the detection of child sexual abuse, it continues to be used commonly for this purpose (Pinkerman et al., 1993). The Rorschach is also used commonly for diagnostic purposes, even though its validity for detecting psychiatric conditions not characterized by thought disorder seems to be quite limited (Wood et al., 2000a). This state of affairs is deeply troubling and raises significant challenges for clinical psychology and allied professions.

Early in the manuscript we delineated three criteria that should be fulfilled before projective indexes are regarded as empirically supported (see also Wood et al., 1996b): (a) consistent relations with one or more specific external criteria (e.g., personality traits, psychological symptoms or disorders) in (b) multiple methodologically sound validation studies that have been (c) conducted by independent investigators. We can now revisit these three criteria in light of the research we have reviewed on the Rorschach, TAT, and human figure drawings. Specifically, we conclude that the following projective indexes can be regarded as empirically supported:

(1) Rorschach: (a) Thought Disorder Index for the Rorschach (TDIR) in the assessment of thought disorder, (b) Rorschach Prognostic Rating Scale (RPRS) in the prediction of treatment outcome, (c) Rorschach Oral Dependency Scale in the assessment of objective behaviors related to dependency, and (d) deviant verbalizations and poor form (as well as the CS SCZI, and other indexes derived from these variables) in the assessment of schizophrenia (and perhaps schizotypal personality disorder and bipolar disorder) and borderline personality disorder.

(2) TAT: (a) McClelland's et al.'s (1953) scoring system for the need for achievement in the assessment of achievement-related outcomes and (b) Westen's (1991) SCORS in the identification of child sexual abuse history and BPD (although in the case of child sexual abuse history, the SCORS Affect-tone scale only).

(3) Human figure drawing indexes: Other than the use of certain global indexes (e.g., overall quality of drawing) to distinguish psychopathology from normality, no indexes have achieved empirical support. These global indexes also tend to have moderate correlations with measures of intelligence, although we do not endorse them as substitutes for standard IQ measures (see also Kamphaus & Pleiss, 1991).

For three reasons it is important to emphasize that our classification of these indexes as empirically supported should be regarded as provisional. First, some of these indexes, such as Westen's (1990) SCORS, have been examined in relatively few published studies. Thus, it is entirely possible that future negative findings could overturn these tentative conclusions. Second, the empirical foundation for some of these techniques has been criticized. As noted earlier, for example, the research support for the RPRS derives largely from studies that can be faulted on methodological grounds (Hunsley & Bailey, 1999). Third, our meta-analysis of the child sexual abuse literature points to the presence of substantial file drawer effects for projective methods. If this publication bias extends to other substantive areas in the projectives literature, the published research may yield an overly sanguine picture of the validity of projective indexes, including those that received empirical support. One major recommendation emanating from our review is clear: estimating the magnitude of the file-drawer effect across different domains should become a major priority among researchers in the literature on projective techniques.

The following projective indexes did not satisfy our three criteria for empirical support: the overwhelming majority of Rorschach indexes, most TAT scoring systems (including the DMM of Cramer, 1991), all isolated signs derived from human figure drawings, and global scoring approaches to human figure drawings that are intended to detect specific conditions (e.g., mood disorders) and child sexual abuse history. It is crucial to note that the projective indexes that received empirical support comprise only a very small percentage of those used routinely in clinical practice. As a consequence, most practitioners who use projective instruments are basing many of their inferences on indexes that are lacking in solid research support.

We should also emphasize that "empirically supported" does not equate to "ready or appropriate for routine clinical use." Even for projective indexes that received empirical support, (a) adequate population norms are typically unavailable, (b) field reliability is untested, and (c) evidence of cultural and ethnic bias has not been clearly ruled out. In addition, with the potential exception of the Rorschach RPRS and McClelland et al.'s (1953) system for scoring achievement needs from the TAT, there is little convincing evidence that these indexes (d) possess clinically meaningful incremental validity above and beyond data provided by other psychological instruments that tend to be readily available to clinicians (e.g., commonly administered questionnaires, interviews). Moreover, as discussed earlier, many of the investigations reviewed here probably overestimate the predictive utility of projective techniques in most clinical settings, because they are based on study designs in which the sample sizes of the pathological and nonpathological groups are approximately equal (Dawes, 1993). In contrast, the prevalence of many of the clinical phenomena (e.g., schizophrenia, borderline personality disorder, child sexual abuse history) assessed by the projective indexes reviewed here is considerably less than 50 percent in many clinical settings.

Finally, we uncovered no evidence for the treatment utility (Hayes et al., 1987) of any projective technique. In other words, there is no research evidence that any projective instrument used for assessment purposes enhances treatment outcome (see also Hunsley & Bailey, 1999). Although absence of evidence is not evidence of absence, there is scant justification for the use of projective techniques in the treatment context unless these techniques can be shown to contribute to therapeutic efficacy. We strongly recommend that researchers in this area undertake studies using the technique of manipulated assessment (Hayes et al., 1987; see also Meehl, 1959). This method treats therapists as participants and randomly assigns them to either receive information from an assessment device (in this case, a projective instrument) or no information from this device. The extent to which the provision of this information enhances treatment outcome is a direct test of the instrument's treatment utility. In recommending such studies, we should make clear that we are not holding projective methods to higher standards than structured methods (e.g., the MMPI-2), whose treatment utility is similarly undemonstrated (Hunsley & Bailey, 1999).[5] Nevertheless, unless the treatment

[5]Although Finn and his colleagues (Finn, 1996; Finn & Tonsager, 1992) reported data indicating that feedback from the MMPI-2 can decrease psychological distress, we do not regard such data as compelling evidence for the MMPI-2's treatment utility (cf., Kubisyzn et al., 2000). Halperin and Snyder (1979) showed that snake-phobic clients provided with Barnum feedback after taking two psychological tests showed enhanced treatment outcome relative to

utility of projective techniques can be demonstrated in studies of manipulated assessment, the rationale for their administration in psychotherapeutic contexts will remain doubtful.

RECOMMENDATIONS FOR RESEARCH, FORENSIC AND CLINICAL PRACTICE, AND EDUCATION AND TRAINING

In this final section of the manuscript, we offer recommendations concerning three major issues: (a) research on the construction of projective methods with demonstrated validity, (b) the forensic and clinical use of projective methods, and (c) the education and training of students on projective methods.

Recommendations for Building a Valid Projective Technique

On the basis of the research reviewed here, we are strongly inclined to agree with Westen, Lohr et al. (1990) that projective techniques are not inherently unreliable or unvalid. Because some projective indexes can attain satisfactory psychometric properties, it is unlikely that projective techniques per se possess intrinsic or ineluctable shortcomings. Instead, we suspect that the poor validity of most projective techniques for their intended purposes stems from their suboptimal design and construction. On the basis of the literature, can we offer any principles or guidelines for constructing projective techniques that are likely to possess adequate validity?

To a limited extent we can. First, most of the projective techniques with reasonable validity rely either implicitly or explicitly on the principle of aggregation across multiple items (see also Lilienfeld, 1999; Riethmiller & Handler, 1997b). With rare exceptions, single items tend to possess a substantial component of "situational uniqueness" (Epstein, 1979) and are therefore highly fallible indicators of the latent construct they are intended to assess. By aggregating across a number of items designed to assess the same construct, measurement error is typically averaged out, thereby resulting in a more reliable and construct valid index. TAT indexes of achievement needs and object relations, for example, make use of this aggregation principle. In contrast, many Rorschach indicators (which are often based on very small numbers of responses) and isolated human figure drawing signs do not. Interestingly, the only figure drawing methods that exhibit indications of modest validity are global indexes, most of which (quantitative indexes) combine many items into a total score. Even qualitative indexes derived from these drawings often rely implicitly on aggregation, as they require the assessor to consider many aspects of the drawing before arriving at a global judgment.

Second, many successful projective techniques consist of ambiguous stimuli that are especially relevant to the construct being assessed. TAT measures of achievement motivation, for example, are based on cards that are preselected to be pertinent to achievement needs, and Westen's (1991) use of the TAT to assess object relations is based on cards that emphasize interpersonal themes. Indeed, the classic concept of "stimulus pull" (see Murstein & Easter, 1965) in the TAT literature implies that certain cards are more likely than others to elicit certain needs and personality traits. A rarely used but reasonably well validated projective technique not reviewed here, the Rosenzweig Picture Frustration Study (Rosenzweig et al., 1947), relies on this relevance principle to assess aggression: the stimuli (cartoons) are explicitly selected to elicit vicarious frustration in the respondent. In contrast, although Hermann Rorschach selected inkblots that seemed to differentiate schizophrenic patients from other individuals, these inkblots were not otherwise preselected to elicit particular classes of responses.

Interestingly, a number of well validated measures of cognitive bias implement this relevance principle, although they are not traditionally classified as projective techniques. For example, Dodge and his colleagues (e.g., Dodge, Murphy, & Buchsbaum, 1984; see also Waldman, 1996) have had considerable success with videotapes depicting children engaged in ambiguous social interactions that "pull" for attributions of aggressive intent (e.g., child A stepping over and ruining the play materials of child B). Children who interpret the intentions of child A as hostile (e.g., "He meant to do it") are more likely than other children to exhibit high levels of real-world aggression (Dodge & Frame, 1982). In addition, both homophone and sentence disambiguation tasks have been used with success in the anxiety disorders literature. For example, when presented with homophones that have both a threatening and nonthreatening meaning (e.g., mourning-morning, dye-die) individuals with some anxiety disorders (e.g., generalized anxiety disorder) are more prone than other individuals to hear the threatening meaning; they are also more likely to interpret ambiguous sentences that have both a threatening and nonthreatening meaning (e.g., "The doctor examined little Emma's growth") as threatening (see McNally, 1996, for a review).

Third, we believe that future development of projective instruments would benefit from an iterative and self-correcting approach to test construction (Loevinger, 1957; Tellegen & Waller, 1994). Using this approach, which is captured nicely by Cattell's (1957) term, the "inductive-hypothetico-deductive spiral," the test developer begins with a tentative formulation of the constructs to be assessed and then progressively revises these constructs (as well as the stimuli assessing them) on the basis of new data. If performed thoughtfully and carefully, the end result will often be both a clarified set of constructs and a psychometrically superior pool of stimuli to assess them. To our knowledge, this iterative approach has been used only rarely to develop projective instruments.

The Washington University Sentence Completion Test

snake-phobic clients who received no feedback after taking these tests. Thus, the work of Finn and his colleagues demonstrates only that some form of feedback to clients can be therapeutic, but it does not demonstrate the treatment utility of a given assessment device nor that this feedback must be accurate.

(WUSCT) is a projective measure of ego development developed by Loevinger (1976), who adhered to all three of the aforementioned guidelines in the process of test construction. The WUCST presents examinees with 36 sentence stems, and the responses to these stems are scored and then combined using a complex algorithm. The sentence stems selected by Loevinger are especially useful for eliciting various aspects of ego development. In addition, the WUCST was constructed and revised throughout numerous cycles of test development. In each cycle, (a) preliminary scoring instructions were devised and applied to previous samples, (b) the data from these samples were used to revise the scoring instructions and, in some cases, the items and conceptualization of the ego development stages themselves, and (c) the revised scoring instructions and items were applied to new samples (see Loevinger, 1993, 1998). The most recent version of the WUCST places respondents in 1 of 8 major stages of ego development ranging from Impulsive (Lowest) to Integrated (Highest).

The WUSCT has demonstrated impressive construct validity in numerous studies by independent investigators (see Hauser, 1976; Loevinger, 1993, for reviews), and fulfills our criteria for empirical support. For example, scores on this instrument correlate (a) moderately to highly with ego level as assessed by interviews (e.g., Lucas, 1971), (b) moderately with scores on Kohlberg's (1981) moral judgment test even after controlling statistically for age (e.g., Lambert, 1972), (c) negatively and substantially with indexes of delinquency and antisocial behavior (Frank & Quinlan, 1976), (d) positively with successful adaptation after divorce (Bursik, 1991), (e) positively with the openness to experience dimension of the "Big Five" personality taxonomy (McCrae & Costa, 1980), and (f) positively with observer ratings of ego resiliency and morality (Westenberg & Block, 1993). In addition, the WUSCT has demonstrated substantial incremental validity above and beyond intelligence measures in the prediction of personality traits among nonclinical participants (Westenberg & Block, 1993) and length of stay and problematic ward behavior among psychiatric inpatients (Browning, 1986). In a sample of twins reared apart, Newman and Bouchard (1998) also found that WUSCT possesses considerable genetic variance even after controlling statistically for the effects of intelligence measures. Finally, as predicted by Loevinger's model of ego development, WUSCT scores have shown curvilinear relations with measures of conformity (Hoppe & Loevinger, 1977; Westenberg & Block, 1993).

Although a number of questions regarding the construct validity of the WUSCT remain to be resolved (e.g., Costa & McCrae, 1993; Jackson, 1993), this instrument is arguably the most extensively validated projective technique. The research evidence for the WUSCT demonstrates that when carefully conceptualized and constructed, projective instruments can indeed meet scientifically acceptable standards for zero-order and incremental validity (for another example of a thoughtfully constructed projective instrument with promising psychometric properties, see Holtzman et al., 1961, and Peixotto, 1980, for discussions of the Holtzman Inkblot Test).

Recommendations Regarding the Forensic and Clinical Use of Projective Techniques

The research literature provides numerous reasons why psychologists should exercise considerable caution in the use of projective instruments in forensic contexts (e.g., custody disputes, sentencing evaluations, parole reviews) and in clinical practice. First, as the present article has documented, the scoring of many projective techniques can often be unreliable, so there is considerable room for subjectivity and error from one psychologist's scores to the next. Second, among the projective scores that can be scored reliably, only a handful have well-demonstrated validity. Third, for the small group of projective scores that possess both adequate scoring reliability and validity, normative data are generally either non-existent or problematic.

Considering these problems, we recommend that forensic and clinical psychologists either refrain from administering the Rorschach, TAT, and human figure drawings, or at least limit their interpretations to the very small number of indexes derived from these techniques that are empirically supported. Whenever possible, forensic and clinical evaluations should be based on more dependable assessment techniques, such as structured psychiatric interviews and well-validated self-report indexes. Moreover, practitioners should use these empirically supported indexes only when (a) adequate population norms are available, (b) there is compelling evidence for incremental validity above and beyond more readily acquired sources of information (e.g., well validated self-report instruments, demographic data) and (c) the base rate of the phenomenon in question (e.g., child sexual abuse) is sufficiently high to render these indexes potentially clinically useful.

We realize that our advice is not likely to be universally heeded. The historical record of the past half century strongly suggests that many psychologists will continue to use inadequately validated projective indexes, even when confronted with negative scientific evidence and despite the risk of harm to clients (Dawes, 1994). In this section we therefore offer advice, recommendations, and general comments that can be of help when dealing with experts who have used a projective technique in a forensic or clinical context (for a more detailed discussion, see Wood et al., in press). Although our suggestions are targeted primarily to professionals who operate in the courtroom or in other forensic arenas, many of these suggestions are applicable to clinical practitioners in general.

1. *Projective techniques are highly controversial.* This simple, undeniable piece of information should always be conveyed to judges and juries who have been offered an expert opinion based on projective techniques. For instance, many judges may be impressed by the "mystique" of the Ror-

schach unless they learn how scientifically controversial this technique is. Expert witnesses should not be allowed to state or imply that projective techniques are widely accepted by the scientific community. Psychologists who use projective techniques in forensic settings have an ethical obligation to describe the limitations of these techniques and the controversy that surrounds them (American Psychological Association, 1992, Standards 2.08a, 7.04b).

2. *Projective techniques are susceptible to faking, as well as to subtle situational influences.* Although we have not reviewed this literature here because of space constraints, recent research suggests that, early claims to the contrary (e.g., Fosberg, 1938, 1941), the Rorschach and perhaps other projective techniques are susceptible to malingering (i.e., "faking bad"). In particular, there is increasing evidence that schizophrenia, depression, and probably posttraumatic stress disorder can be faked on the Rorschach (e.g., Perry & Kinder, 1990; Schretlen, 1997) and that such faking cannot be detected using existing Rorschach indexes. Moreover, there is virtually no methodologically sound research on the susceptibility of the Rorschach to impression management (i.e., "faking good"; Schretlen, 1997), although the results of one study indicate that untrained participants can readily simulate a high need for achievement on the TAT (Holmes, 1974). Experts who present projective techniques in court should be forthright about the potential effects of malingering and impression management, as well as the absence of research evidence that these response sets can be detected.

   In addition, it is well known that many projective techniques, including the Rorschach and TAT, are highly susceptible to situational influences, including subtle verbal reinforcement (e.g., saying "mmm-hmm" following certain responses), the mood and even hunger level of the examinee, and the gender, perceived status, and physical characteristics of the examiner (see Masling, 1960, 1966, 1997, for reviews). In an amusing illustration of the lattermost set of variables, female participants in a study of human figure drawings were more likely to draw male figures with mustaches if the examiner himself had a mustache than if he was clean shaven (Yagoda & Wolfson, 1964). Such seemingly minor situational variables may attenuate the validity of projective techniques in some real world settings.

3. *Projective techniques are routinely used for purposes for which they are invalid or poorly supported by research.* Whenever an expert witness uses a projective technique, a well-informed opposing attorney, assisted by a well-informed consulting psychologist, can often mount a withering challenge to the validity or legal "relevance" of specific scales or scores. Such challenges have been rare in the past (Weiner, Exner, & Sciara, 1996). In the future, however, challenges may become more common, as attorneys and the psychologists who assist them begin to recognize the vulnerability of projective techniques to legitimate criticism.

4. *The scoring of many projective techniques can be unreliable or poor.* Even highly regarded experts can disagree about the scoring of a Rorschach or certain human figure drawing signs. Furthermore, our personal observations suggest that scoring errors may be fairly common in forensic and clinical contexts, although we are unaware of any formal research on the prevalence of such errors. For this reason, in both contexts it is often advisable to have the projective materials re-scored by a second expert who does not know the first expert's scores. This procedure may often reveal errors or discrepancies in scoring that would substantially modify the original examiner's conclusions and interpretations.

5. *Norms for projective techniques are often non-existent, poor, or misleading.* When norms are absent, experts have substantial latitude in interpreting a client's scores on a projective technique. As a result, different experts may arrive at widely varying interpretations of the same projective scores (such differing interpretations may be exacerbated by differences in how clinicians intuitively combine and weight scores). When norms for projective instruments are misleading, clinicians' judgments and predictions are likely to be erroneous. The recently noted problems with the CS norms render the Rorschach particularly vulnerable to legal challenge. It is often possible to show that supposedly "pathological" Rorschach scores are actually well within the normal range (Shaffer et al., 1999; Wood, Nezworski, et al., 2000).

6. *Projective techniques may be biased against North American minority groups and individuals who live outside North America.* As we have discussed, the use of the Rorschach with American minority groups and non-Americans is problematic. There is little recent research to provide guidance concerning other projective techniques. However, studies from the 1950s and 1960s indicate that cross-cultural use of tests like the TAT is fraught with pitfalls and potential problems (Holtzman, 1980; Kaplan, 1961; Klineberg, 1980). In addition, there are often substantial differences in the characteristics of human figure drawings across ethnic and cultural groups (e.g., Handler & Habernicht, 1994). The preponderance of the evidence suggests that the use of the Rorschach and other projective techniques to evaluate American minorities and non-Americans may lead to erroneous interpretations. Experts who use these tests to evaluate minorities or non-Americans should be challenged to demonstrate that they have used appropriate norms, and that the interpretation of scores is valid for the group of clients in question.

7. *Projective techniques and the Daubert criteria.* In 1993, the U.S. Supreme Court articulated the "Daubert criteria" for the admissibility of scientific evidence in federal courts (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993). These criteria have been adopted by many state courts. Considerable doubt exists regarding whether commonly used projective techniques are legally admissible under the Daubert criteria. Although we cannot explore this issue in the depth that it deserves, we will mention four relevant points. First, scholars disagree on whether the CS meets the Daubert standards. Specifically, McCann (1998) argued that the CS does meet the Daubert criteria, whereas Grove and Barden (1999) reached the opposite conclusion.

Second, it is very unlikely that Rorschach systems other than the CS meet the Daubert criteria. For example, although McCann (1998) adopted an optimistic view regarding the admissibility of the CS, he was far less sanguine regarding other Rorschach systems.

Third, no peer-reviewed articles have argued that indexes derived from the TAT, human figure drawings, or other projective methods meet the Daubert criteria. Given the limited or negative scientific evidence regarding these techniques, it is doubtful that they could withstand close scrutiny under either Daubert or the consensually adopted professional standards applied to assessment techniques used in forensic or clinical settings (Heilbrun, 1992; Hunsley et al., in press).

Fourth, the Daubert criteria notwithstanding, many judges will probably continue to admit the Rorschach and other projective techniques into court (McKinzey & Ziegler, 1999). However, a hearing to determine a projective technique's admissibility under Daubert can still serve a useful purpose by alerting a judge to the problems described in the present article. Furthermore, even if a projective technique is admitted into court, it may prove a liability to the side that uses it. As we have indicated, projective techniques are vulnerable to challenge on numerous grounds, and the expert who uses them may be highly vulnerable if cross-examined by a well-informed attorney.

Recommendations for Education and Training

On the basis of the research reviewed here, what suggestions can we offer for the training and education of the next generation of clinical and counseling psychologists? In closing, we present three recommendations. First, given the relatively weak evidence for the zero-order and incremental validity of most projective indexes, the amount of time devoted to educating and training students in the administration and scoring of projective techniques should be reduced (see also Garb, 1998). This recommendation is consistent with that of the American Psychological Association (APA) Division 12 (Clinical Psychology) Task Force on Assessment, whose model graduate assessment curriculum for the 21st century excluded training in projective techniques (Grove et al., 2000).

Second, if instructors intend to cover projective techniques in their courses, they should expose students to the research and meta-analytic literature regarding their psychometric properties. In particular, instructors should teach students to distinguish between projective indexes that do and do not have empirical support. They should also expose students to research concerning variables that can contribute to the low validity of projective techniques in some real-world settings, such as response sets (Schretlen, 1997) and situational influences (Masling, 1967). In addition, instructors should discuss in detail the forensic and ethical implications of relying on projective indexes that are not well validated.

Third, all graduate students in clinical and counseling psychology should be systematically exposed to the extensive body of research on clinical judgment and decision making. This recommendation has also been put forth by the APA Division 12 Task Force on Assessment (Grove et al., 2000). For example, graduate students should be made aware of the weak or negligible relation beween the amount of prior experience with an assessment technique and predictive accuracy (Garb, 1998). In addition, by becoming familiar with research on clinical judgment and decision making, graduate students will become aware of factors that can lead practitioners to become erroneously convinced of the validity of projective methods. For example, Chapman and Chapman (1967, 1969) demonstrated that even when Rorschach and human figure drawing signs are paired randomly with psychopathological characteristics, individuals will tend to perceive statistical relationships between signs and psychopathological characteristics that share strong semantic or associative connections (see Starr & Katkin, 1969, for similar findings regarding sentence completion tests). Moreover, this phenomenon of illusory correlation may even be more powerful in real world than in experimental settings, as there is evidence that the magnitude of illusory correlation between human figure drawing stimuli and psychopathological features increases as information processing load increases (Lueger & Petzel, 1979). Foremost and finally, graduate students should be taught the crucial and sometimes painful lesson that this research literature imparts: clinical experience and clinical intuition can sometimes be misleading. As one of us observed elsewhere, the long and difficult task of training the scientifically-minded clinician necessitates mastering "a skill that does not come naturally to any of us: disregarding the vivid and compelling data of subjective experience in favor of the often dry and impersonal results of objective research" (Lilienfeld, 1999, p. 38).

We thank Drs. William Grove, Lee Sechrest, and John Hunsley for their extremely detailed and valuable comments on an earlier draft of this manuscript. In addition, we are grateful to Dr. Elaine Walker for her invaluable assistance and advice

throughout this project, Drs. M. Teresa Nezworski and R.K. McKinzey for their insights, and Dr. George Alliger for his statistical advice.

REFERENCES

Ackerman, S.J., Clemence, A.J., Weatherill, R. & Hilsenroth, M.J. (1999). Use of the TAT in the assessment of DSM-IV Cluster B personality disorders. *Journal of Personality Assessment*, *73*, 422–448.

Ackerman, M.J., & Ackerman, M.C. (1997). Custody evaluation practices: A survey of experienced professionals (revisited). *Professional Psychology: Research and Practice*, *28*, 137–145.

Acklin, M.W. (1999). Behavioral science foundations of the Rorschach test: Research and clinical applications. *Assessment*, *6*, 319–326.

Acklin, M.W., McDowell, C.J., Verschell, M.S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment*, *74*, 15–47.

Adair, H.E., & Wagner, E.E. (1992). Stability of unusual verbalizations on the Rorschach for outpatients with schizophrenia. *Journal of Clinical Psychology*, *48*, 250–256.

Adler, P.T. (1952). Evaluation of the figure drawing technique: Reliability, factorial structure, and diagnostic usefulness. *Journal of Consulting and Clinical Psychology*, *33*, 52–57.

Aiken, L.R. (1996). *Personality assessment: Methods and practices*. Seattle: Hogrefe and Huber.

Aldridge, N.C. (1998). Strengths and limitations of forensic child sexual abuse interviews with anatomical dolls: An empirical review. *Journal of Psychopathology and Behavioral Assessment*, *20*, 1–41.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, D.C.: Author.

American Psychological Association (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, *47*, 1597–1611.

Anastasi, A. (1982). *Psychological testing*. NY: MacMillan.

Anderson, J.W. (1999). Henry A. Murray and the creation of the Thematic Apperception Test. In L. Geiser & M.I. Stein (Eds.), *Evocative Images: The Thematic Apperception Test and the art of projection* (pp. 23–38). Washington, D.C.: American Psychological Association.

Arnold, M.B. (1962). *Story sequence analysis*. NY: Columbia University Press.

Archer, R.P. (1999). Introduction to a special section: Perspectives on the Rorschach. *Assessment*, *6*, 307–311.

Archer, R.P., & Krishnamurthy, R. (1993a). Combining the Rorschach and the MMPI in the assessment of adolescents. *Journal of Personality Assessment*, *60*, 132–140.

Archer, R.P., & Krishnamurthy, R. (1993b). A review of MMPI and Rorschach interrelationships in adult samples. *Journal of Personality Assessment*, *61*, 277–293.

Archer, R.P., & Krishnamurthy, R. (1997). MMPI-A and Rorschach indices related to depression and conduct disorder: An evaluation of the incremental validity hypothesis. *Journal of Personality Assessment*, *69*, 517–533.

Aronow, E., & Reznikoff, M. (1973). Attitudes toward the Rorschach test expressed in book reviews: A historical perspective. *Journal of Personality Assessment*, *37*, 309–315.

Aronow, E. & Reznikoff, M. (1976). *Rorschach content interpretation.* NY: Grune & Stratton.

Aronow, E., Reznikoff, M., & Moreland, K. (1995). The Rorschach: Projective technique or psychometric test? *Journal of Personality Assessment*, *64*, 213–228.

Aylward, E., Walker, E., & Bettes, B. (1984). Intelligence in schizophrenia. *Schizophrenia Bulletin*, *10*, 430–459.

Ball, J.D., Archer, R.P., & Imhof, E.A. (1994). Time requirements of psychological testing: A survey of practitioners. *Journal of Personality Assessment*, *63*, 239–249.

Barends, A., Westen, D., Byers, S., Leigh, J., & Silbert, D. (1990). Assessing affect-tone of relationship paradigms from TAT and interview data. *Psychological Assessment*, *2*, 329–332.

Bellak, L. (1975). *The T.A.T., C.A.T., and S.A.T. in clinical use* (3rd ed.), NY: Grune & Stratton.

Beyerstein, B.L., & Beyerstein, D.F. (1992). *The Write Stuff: Evaluations of graphology − the study of handwriting analysis*. Buffalo, NY: Prometheus Books.

Bilett, J.L., Jones, N.F., & Whitaker, L.C. (1982). Exploring schizophrenic thinking in older adolescents with the WAIS, Rorschach, and WIST. *Journal of Clinical Psychology*, *38*, 232–243.

Blatt, S.J., Wein, S.J., Chevron, E.S., & Quinlan, D.M. (1979). Parental representatives and depression in normal young adults. *Journal of Abnormal Psychology*, *88*, 388–397.

Blum, G.S. (1950). *The Blacky Pictures: Manual of Instructions*. NY: Psychological Corporation.

Board of Professional Affairs (1998). Awards for distinguished professional contributions: John Exner. *American Psychologist*, 53, 391–392.

Bornstein, R.F. (1996). Construct validity of the Rorschach Oral Dependency Scale: 1967–1995. *Psychological Assessment*, *8*, 200–205.

Bornstein, R.F. (1999). Criterion validity of objective and projective dependency tests: A meta-analytic assessment of behavioral prediction. *Psychological Assessment*, *11*, 48–57.

Borstellmann, L.J., & Klopfer, W.G. (1953). The Szondi Test: A review and critical evaluation. *Psychological Bulletin*, *50*, 112–132.

Boscan, D.C. (2000). The Rorschach test: A Mexican sample using the Comprehensive System. (Doctoral dissertation, The Fielding Institute, 1999). *Dissertation Abstracts International*, *60*, 4285B.

Browning, D.L. (1986). Psychiatric ward behavior and length of stay in adolescent and young adult inpatients: A developmental approach to prediction. *Journal of Consulting and Clinical Psychology*, *54*, 227–230.

Bruhn, A.R. (1992). The Early Memories Procedure: A projective test of autobiographical memory: I. *Journal of Personality Assessment*, *58*, 1–15.

Buck, J.N. (1964). *The House-Tree-Person manual supplement*. Beverly Hills, CA: Western Psychological Services.

Burns, R.C. (1987). *Kinetic-House-Tree Person Drawings (K-H-T-P): An interpretive manual*. NY: Bruner-Mazel.

Bursik, K. (1991). Adaptation to divorce and ego development in adult women. *Journal of Personality and Social Psychology*, *60*, 300–306.

Butcher, J.N., Nezami, E., & Exner, J.E. (1998). Psychological assessment of people in diverse cultures. In S.S. Kazarian & D.R. Evans (Eds.), *Cultural Clinical Psychology* (pp. 61–105). NY: Oxford University Press.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Caplehorn, W.F., & Sutton, A.J. (1965). Need achievement and its relation to school performance, anxiety, and intelligence. *Australian Journal of Psychology*, *17*, 44–51.

Carlson, K., Quinlan, D., Tucker, G., & Harrow, M. (1973). Body disturbance and sexual elaboration factors in figure drawings of schizophrenic patients. *Journal of Personality Assessment*, *37*, 56–63.

Cattell, R.B. (1957). *Personality and motivation: Structure and measurement*. NY: World Book Company.

Chapman, L.J., & Chapman, J.P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*, 193–204.

Chapman, L.J., & Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic observations. *Journal of Abnormal Psychology*, *74*, 271–280.

Cohen, J.E. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Chase, D.A. (1987). An analysis of Human Figure and Kinetic Family Drawings of sexually abused children and adolescents. *Dissertation Abstracts International*, *48* (2A), 338.

Cochrane, C.T. (1972). Effects of diagnostic information on empathic understanding by the therapist in a psychotherapy analogue. *Journal of Consulting and Clinical Psychology*, *38*, 359–365.

Coleman, M.J., Carpenter, J.T., Warenaux, C., Levy, D.L., Shenton, M.E., Perry, J., Medoff, D., Wong, H., Monach, D., Meyer, P., O'Brian, C., Valentino, C., Robinson, D., Smith, M., Makowski, D., & Holzman, P.S. (1993). The Thought Disorder Index: A reliability study. *Psychological Assessment*, *5*, 336–342.

Conger, A.J., & Jackson, D.N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement*, *32*, 579–599.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447–452.

Costa, P.T., & McCrae, R.R. (1993). Ego development and trait models of personality. *Psychological Inquiry*, *4*, 20–23.

Costello, R.M. (1998). Psychometric definition of Rorschach determinant component structure. *European Journal of Psychological Assessment*, *14*, 116–123.

Craddick, R. (1961). Size of Santa Claus drawings as a function of time before and after Christmas. *Journal of Psychological Studies*, *12*, 121–125.

Cramer, P. (1991). *The development of defense mechanisms: Theory, research, and assessment*. NY: Springer-Verlag.

Cramer, P. (1999). Ego functions and ego development: Defense mechanisms and intelligence as predictors of ego level. *Journal of Personality*, *67*, 735–760.

Cramer, P. (1999). Future directions for the Thematic Apperception Test. *Journal of Personality Assessment, 72*, 74–92.

Cramer, P. & Block, J. (1998). Preschool antecedents of defense mechanism use in young adults: A longitudinal study. *Journal of Personality and Social Psychology*, *74*, 159–169.

Cramer, P., & Gaul, R. (1988). The effects of success and failure on children's use of defense mechanisms. *Journal of Personality*, *56*, 729–742.

Cressen, R. (1975). Artistic quality of drawings and judges' evaluations of the DAP. *Journal of Personality Assessment*, *39*, 132–137.

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

Currie, S.F., Holtzman, W.H., & Swartz, J.D. (1974). Early indicators of personality traits viewed retrospectively. *Journal of School Psychology*, *12*, 51–59.

Cvetkovic, R. (1979). Conception and representation of space in human figure drawings by schizophrenic and normal subjects. *Journal of Personality Assessment*, *43*, 247–256.

Dana, R.H. (1955). Clinical diagnosis and objective TAT scoring. *Journal of Consulting Psychology*, *20*, 33–36.

Dana, R.H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn & Bacon.

Dana, R.H. (Ed.). (2000). *Handbook of cross-cultural and multicultural personality assessment*. Mahwah, New Jersey: Erlbaum.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 113 S.Ct. 2786 (1993).

Dawes, R.M. (1993). Prediction of the future versus an understanding of the past: A basic assymetry. *American Journal of Psychology*, *106*, 1–24.

Dawes, R.M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: The Free Press.

Dawes, R.M. (1999). Two methods for studying the incremental validity of a Rorschach variable. *Psychological Assessment*, *11*, 297–302.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Depue, R.A., & Collins, P.F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences*, *22*, 491–569.

Dodge, K.A., & Frame, C.L. (1982). Social cognitive deficits and biases in aggressive boys. *Child Development*, *53*, 8–22.

Dodge, K.A., Murphy, R.R., & Buchsbaum, K. (1984). The assessment of intention-cue detection skills in children: Implications for developmental psychopathology. *Child Development*, *55*, 163–173.

Dollinger, S., & Cramer, P. (1990). Children's defensive responses and emotional upset following a disaster: A projective assessment. *Journal of Personality Assessment*, *54*, 116–127.

Dosajh, N.L. (1996). Projective techniques with particular reference to inkblot tests. *Journal of Projective Psychology and Mental Health*, *3*, 59–68.

Drachnik, C. (1994). The tongue as a graphic symbol of sexual abuse. *Art Therapy*, *11*, 58–61.

Dudley, H.K., Craig, E.M., Mason, M., & Hirsch, S.M. (1976). Drawings of the opposite sex: Continued use of the Draw-A-Person test and young state hospital patients. *Journal of Adolescence*, *5*, 201–219.

Durand, V.M., Blanchard, E.B., & Mindell, J.A. (1988). Training in projective testing: A survey of clinical training directors and internship directors. *Professional Psychology: Research and Practice*, *19*, 236–238.

Emmons, R.A., & McAdams, D.P. (1991). Personal strivings and motive dispositions: Exploring the links. *Personality and Social Psychology Bulletin*, *17*, 648–654.

Entwisle, D.R. (1972). To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin*, *77*, 377–391.

Epstein, S. (1979). The stability of behavior: I. On predicting more of the people most of the time. *Journal of Personality and Social Psychology*, *37*, 1097–1126.

Eron, L. (1950). A normative study of the TAT. *Psychological Monographs*, *64* (Whole No. 315).

Erstad, D. (1996). An investigation of older adults' less frequent human movement and color responses on the Rorschach (Doctoral dissertation, Marquette University, 1995). *Dissertation Abstracts International*, *57*, 4084B.

Exner, J.E. (1969). *The Rorschach systems.* New York: Grune & Stratton.

Exner, J.E. (1974). *The Rorschach: A Comprehensive System. Volume 1.* NY: Wiley.

Exner, J.E. (1986). *The Rorschach: A Comprehensive System. Volume 1: Basic Foundations* (2nd ed.). NY: Wiley.

Exner, J.E. (1989). Searching for projection in the Rorschach. *Journal of Personality Assessment*, *53*, 520–536.

Exner, J.E. (1991). *The Rorschach: A Comprehensive System. Volume 2: Interpretation* (2nd ed.). NY: Wiley.

Exner, J.E. (1992). Some comments on "A conceptual critique of the EA:es comparison in the Comprehensive Rorschach System." *Psychological Assessment*, *4*, 297–300.

Exner, J.E. (1993). *The Rorschach: A Comprehensive System. Volume 1: Basic Foundations* (3rd ed.). NY: Wiley.

Exner, J.E. (1995). Comment on "Narcissism in the Comprehensive System for the Rorschach." *Clinical Psychology: Science and Practice*, *2*, 200–206.

Exner, J.E. (1996). A comment on "The Comprehensive System for the Rorschach: A critical examination." *Psychological Science*, *7*, 11–13.

Exner, J.E. & Weiner, I.B. (1995). *The Rorschach: A Comprehensive System. Volume 3: Assessment of children and adolescents* (2nd ed.). New York: Wiley.

Feldman, M., & Hunt, R.G. (1958). A relation of difficulty in drawing and ratings of adjustment based on human figure drawings. *Journal of Consulting Psychology*, *22*, 217–220.

Fineman, S. (1977). The achievement motive and its measurement. Where are we now? *British Journal of Psychology*, *68*, 1–22.

Finn, S.E. (1996). *Manual for using the MMPI-2 as a therapeutic intervention*. Minneapolis: University of Minnesota Press.

Finn, S.E., & Tonsager, M.E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment*, *4*, 278–287.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed). NY: Wiley.

Fosberg, I.A. (1938). Rorschach reactions under varied conditions. *Rorschach Research Exchange*, *3*, 12–20.

Fosberg, I.A. (1941). An experimental study of the reliability of the Rorschach psychodiagnostic technique. *Rorschach Research Exchange*, *5*, 72–84.

Frank, L.K. (1948). *Projective methods*. Springfield, Ill.: Thomas.

Frank, G. (1976). On the validity of hypotheses derived from the Rorschach: I. The relationship between color and affect. *Perceptual and Motor Skills*, *43*, 411–427.

Frank, G. (1978). On the validity of hypotheses derived from the Rorschach: III. The relationship between shading and anxiety. *Perceptual and Motor Skills*, *46*, 531–538.

Frank, G. (1979a). On the validity of hypotheses derived from the Rorschach: V. The relationship between form level and ego strength. *Perceptual and Motor Skills*, *48*, 375–384.

Frank, G. (1979b). On the validity of hypotheses derived from the Rorschach: VI. *M* and the intrapsychic life of individuals. *Perceptual and Motor Skills*, *48*, 1267–1277.

Frank, G. (1980). New directions in Rorschach research: I. The process-reactive differentiation. *Perceptual and Motor Skills*, *50*, 187–191.

Frank, G. (1993a). *C'* and depression. *Psychological Reports*, *72*, 1184–1186.

Frank, G. (1993b). On the meaning of movement responses on the Rorschach. *Psychological Reports*, *73*, 1219–1225.

Frank, G. (1993c). On the validity of hypotheses derived from the Rorschach: The relationship between color and affect, update 1992. *Psychological Reports*, *73*, 12–14.

Frank, G. (1993d). On the validity of hypotheses derived from the Rorschach:

The relationship between shading and anxiety, update 1992. *Psychological Reports*, *72*, 519–522.

Frank, G. (1993e). On the validity of Rorschach's hypotheses: The relationship of space responses (*S*) to oppositionalism. *Psychological Reports*, *72*, 1111–1114.

Frank, G. (1993f). Use of the Rorschach to predict whether a person would benefit from psychotherapy. *Psychological Reports*, *73*, 1155–1163.

Frank, G. (1994a). On form level on the Rorschach. *Psychological Reports*, *75*, 315–320.

Frank, G. (1994b). On the prediction of aggressive behavior from the Rorschach. *Psychological Reports*, *75*, 183–191.

Frank, G. (1994c). On the prediction of suicide from the Rorschach. *Psychological Reports*, *74*, 787–794.

Frank, G. (1997). Research assessment of the clinical utility of the Rorschach. *Psychological Reports*, *81*, 1255–1258.

Freud, S. (1911). On the mechanism of paranoia. In S. Freud, *General Psychological Theory: Papers on Metapsychology*. NY: Collier Books.

Frueh, B.C., Leverett, J.P., & Kinder, B.N. (1995). Interrelationship between MMPI-2 and Rorschach variables in a sample of Vietnam veterans with PTSD. *Journal of Personality Assessment*, *64*, 312–318.

Gadol, I. (1969). The incremental and predictive validity of the Rorschach test in personality assessment of normal, neurotic, and psychotic subjects. *Dissertation Abstracts*, *29,* 3482-B (University Microfilms No. 69-4469).

Gann, M.K. (1995). The Rorschach and other projective methods. In Jay Ziskin (Ed.), *Coping with psychiatric and psychological testimony*. *Vol. II: Special topics* (5th ed.) (pp. 823–884). Los Angeles, CA: Law and Psychology Press.

Gallucci, N.T. (1990). On the synthesis of information from psychological tests. *Psychological Reports*, *67*, 1243–1260.

Garb, H.N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review*, *4*, 641–655.

Garb, H.N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, *105*, 387–396.

Garb, H.N. (1998a). Recommendations for training in the use of the Thematic Apperception Test (TAT). *Professional Psychology: Research and Practice*, *29, 621–622.*

Garb, H.N. (1998b). *Studying the clinician: Judgment research and psychological assessment.* Washington, D.C.: American Psychological Association.

Garb, H.N. (1999). Call for a moratorium on the use of the Rorschach Inkblot in clinical and forensic settings. *Assessment*, *6*, 313–315.

Garb, H.N., Florio, C.M., & Grove, W.M. (1998). The validity of the Rorschach and the Minnesota Multiphasic Personality Inventory: Results from meta-analyses. *Psychological Science*, *9*, 402–404.

Garb, H.N., Florio, C.M., & Grove, W.M. (1999). The Rorschach controversy: reply to Parker, Hunsley, and Hanson. *Psychological Science*, *10*, 293–294.

Garb, H.N., Wood, J.M., & Lilienfeld, S.O. (2000). *The detection and assessment of child sexual abuse: An evaluation of the Rorschach*, *Thematic Apperception Test, and projective drawings*. Manuscript in preparation.

Garb, H.N., Wood, J.M., & Nezworski, M.T. (2000). Projective techniques and the detection of child sexual abuse. *Child Maltreatment*, *5*, 161–168.

Garb, H.N., Wood, J.M., & Nezworski, M.T., Grove, W.M., & Stejskal, W.J. (in press). Towards a resolution of the Rorschach controversy. *Psychological Assessment.*

Gittelman Klein, R. (1986). Questioning the clinical usefulness of projective psychological tests for children. *Developmental and Behavioral Pediatrics*, *7*, 378–382.

Glass, M.H., Bieber, S.L., & Tkachuk, M.J. (1996). Personality styles and dynamics of Alaska native and nonnative incarcerated men. *Journal of Personality Assessment*, *66*, 583–603.

Gluck, M.R. (1955). The relationships between hostility in the TAT and behavioral hostility. *Journal of Projective Techniques*, *19*, 21–26.

Golden, M. (1964). Some effects of combining psychological tests on clinical inferences. *Journal of Consulting Psychology*, *28*, 440–446.

Greene, R.L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Gresham, F.M. (1993). "What's wrong with this picture?": Response to Motta et al.'s review of human figure drawings. *School Psychology Quarterly*, *8*, 182–186.

Grobstein, G. (1996). Human Figure Drawings and the identification of child sexual abuse. *Dissertation Abstracts International*, *57(8A), 339*.

Gronnerod, C. (1999). Rorschach interrater agreement estimates: An empirical evaluation. *Scandinavian Journal of Psychology*, *40*, 115–120.

Groth-Marnat, G. (1984). *Handbook of psychological assessment.* NY: Van Nostrand Reinhold.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). NY: Wiley.

Groth-Marnat, G., & Roberts, L. (1998). Human figure drawings and House Tree Person drawings as indicators of self-esteem: A quantitative approach. *Journal of Clinical Psychology*, *54*, 219–222.

Grove, W.M. (Chair). APA Division 12 (Clinical) Presidential Task Force "Assessment for the Year 2000." *Report of the Task Force*. Washington, D.C.: American Psychological Association, Division 12 (Clinical Psychology).

Grove, W.M., & Barden, R.C. (1999). Protecting the integrity of the legal system: The admissibility of testimony from mental health experts, under Daubert/Kumho analyses. *Psychology, Public Policy*, *and Law*, *5*, 224–242.

Gutin, N.J. (1997). Differential object representations in inpatients with narcissistic and borderline personality disorders and normal controls. *Dissertation Abstracts International*, *58* (03-B), 1532.

Halperin, K., & Snyder, C.R. (1979). Effects of enhanced psychological test feedback on treatment outcome: Therapeutic implications of the Barnum effect. *Journal of Consulting and Clinical Psychology*, *47*, 140–146.

Hamel, M., Shaffer, T.W., & Erdberg, P. (2000). A study of nonpatient preadolescent Rorschach protocols. *Journal of Personality Assessment*, *75*, 280–294.

Hammer, E.F. (1958). *The clinical application of figure drawings.* Springer, IL: C.G. Thomas.

Hammer, E.F. (1959). Critique of Swensen's "Empirical evaluation of human figure drawings." *Journal of Projective Techniques*, *23*, 30–32.

Hammer, E.F. (1968). *The clinical application of projective drawings*. Springfield, IL.: Thomas.

Handler, L., & Habernicht, D. (1994). The Kinetic Family Drawing: A review of the literature. *Journal of Personality Assessment*, *62*, 440–464.

Handler, L., & Reyher, J. (1965). Figure drawing anxiety indices: A review of the literature. *Journal of Projective Techniques*, *29*, 305–313.

Hauser, S.T. (1976). Loevinger's model and measure of ego development: A critical review. *Psychological Bulletin*, *83*, 928–995.

Hayes, S.C., Nelson, R.O., & Jarrett, R.B. (1987). The treatment utility of assessment: A functional approach to evaluating treatment quality. *American Psychologist*, *42*, 963–974.

Hayslip, B., Lowman, R.L. (1986). The clinical use of projective techniques with the aged: A critical review and synthesis. *Clinical Gerontologist*, *5*, 63–94.

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior*, *16*, 257–272.

Hibbard, S., Farmer, L., Wells, C., Difillipo, E., Barry, W., Korman, R., & Sloan, P. (1994). Validation of Cramer's Defense Mechanism Manual for the TAT. *Journal of Personality Assessment*, *63*, 197–210.

Hiler, E.W., & Nesvig, D. (1965). An evaluation of criteria used by clinicians to infer pathology from figure drawings. *Journal of Consulting Psychology*, *29*, 520–529.

Hiller, J.B., Rosenthal, R., Bornstein, R.F., Berry, D.T.R. Brunell-Neuleib, S. (1999). A comparative meta–analysis of Rorschach and MMPI validity. *Psychological Assessment*, *11*, 278–296.

Holaday, M., Smith, D.A., & Sherry, A. (2000). Sentence completion tests: A review of the literature and results of a survey of members of the Society for Personality Assessment. *Journal of Personality Assessment*, *74*, 371–385.

Holmes, D.S. (1974). The conscious control of thematic projection. *Journal of Consulting and Clinical Psychology*, *42*, 323–329.

Holmes, D.S. (1978). Projection as a defense mechanism. *Psychological Bulletin*, *85*, 677–688.

Holtzman, W.H. (1980). Projective techniques. In H.C. Triandis & J.W. Berry (Eds.), *Handbook of cross-cultural psychology. Methodology*. *Volume 2* (pp. 245–278). Boston: Allyn and Bacon.

Holtzman, W.H., Thorpe, J.S., Swartz, J.D., & Herron, E.W. (1961). *Inkblot perception and personality*. Austin, TX: University of Texas Press.

Horowitz, M.J. (1962). A study of clinicians' judgments from projective test protocols. *Journal of Consulting Psychology*, *26*, 251–256.

Hundal, P.S., & Jerath, J.M. (1972). Correlates of projective measures of achievement motivation and their factorial structure. *Indian Journal of Psychology*, *47*, 15–27.

Hunsley, J., & Bailey, J.M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment*, *11*, 266–277.

Hunsley, J., & Bailey, J.M. (in press). Whither the Rorschach? An analysis of the evidence. *Psychological Assessment*.

Hunsley, J., Lee, C., & Wood, J.M. (in press). Controversial and questionable assessment techniques. In S.O. Lilienfeld, J.M. Lohr, & S.J. Lynn (Eds.), *Science and pseudoscience in contemporary clinical psychology*. NY: Guilford.

Jackson, D.N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, *78*, 229–248.

Jackson, D.N. (1993). Personality development and nonlinear measurement models. *Psychological Inquiry*, *4*, 30–33.

Jensen, A.R. (1965). A review of the Rorschach. In O.K. Buros (Ed.), *Sixth mental measurements handbook* (pp. 501–509). Highland Park, New Hampshire: Gryphon.

Joiner, T.E., & Schmidt, K.L. (1997). Drawing conclusions − or not from drawings. *Journal of Personality Assessment*, *69*, 476–481.

Joiner, T.E., Schmidt, K.L., & Barnett, J. (1996). Size, detail, and line heaviness in children's drawings as correlates of emotional distress: (More) negative evidence. *Journal of Personality Assessment*, *67*, 127–141.

Johnson, B. (1989). *DSTAT: Software for the meta-analytic review of research literature* [Computer program]. Hillsdale, NJ: Erlbaum.

Johnston, M., & Holzman, P.S. (1979). *Assessing schizophrenic thinking*. San Francisco: Jossey-Bass.

Jolley, R.P. (1995). *Children's production and perception of visual metaphors for mood and emotion in line drawings and in art*. Doctoral thesis, University of Birmingham, UK.

Jorgensen, K., Andersen, T.J., & Dam, H. (2000).The diagnostic efficiency of the Rorschach Depression Index and the Schizophrenia Index: A review. *Assessment*, *7,* 259–280.

Kagan, J. (1956). The measurement of overt aggression from fantasy. *Journal of Abnormal and Social Psychology*, *52*, 390–393.

Kahill, S. (1984). Human figure drawing in adults: An update of the empirical evidence, 1967–1982. *Canadian Psychology*, *25*, 269–292.

Kamphaus, R.W., & Pleiss, K.L. (1991). Draw-a-Person techniques: Tests in search of a construct. *Journal of School Psychology*, *29*, 395–401.

Kaplan, B. (1961). Cross-cultural use of projective techniques. In F.L.K. Hsu (Ed.), *Psychological Anthropology* (pp. 235–254). Homewood, IL: Dorsey Press.

Karon, B.P. (1978). Projective tests are valid. *American Psychologist*, *33*, 764–765.

Karon, B.P., & O'Grady, P. (1970). Quantified judgments of mental health from the Rorschach, TAT, and clinical status interview by means of a scaling technique. *Journal of Consulting and Clinical Psychology*, *34*, 229–235.

Katz, H.E., Russ, S.W., & Overholser, J.C. (1993). Sex differences, sex roles, and projection on the TAT: Matching stimulus to examinee gender. *Journal of Personality Assessment*, *60*, 186–191.

Keiser, R.E., & Prather, E.N. (1990). What is the TAT? A review of ten years of research. *Journal of Personality Assessment*, *55*, 800–803.

Kinder, B.N. (1992). The problems of *R* in clinical settings and in research: Suggestions for the future. *Journal of Personality Assessment*, *58*, 252–259.

Kleiger, J.H. (1992). A conceptual critique of the EA:es comparison in the Comprehensive Rorschach System. *Psychological Assessment*, *4*, 288–296.

Klineberg, O. (1980). Historical perspectives: Cross-cultural psychology before 1960. In H.C. Triandis & W.W. Lambert (Eds.), *Handbook of cross-cultural psychology*. *Perspectives*. *Volume 1* (pp. 31–67). Boston: Allyn and Bacon.

Klopfer, W.F., & Taulbee, E. (1976). Projective tests. *Annual Review of Psychology*, *27*, 543–567.

Knoff, H.M., & Prout, H.T. (1985). The Kinetic Drawing System: A review and integration of the Kinetic Family and School Drawing Techniques. *Psychology in the Schools*, *22*, 50–59.

Kohlberg, L. (1981). *The meaning and measurement of moral development*. Worchester, MA.: Clark University Press.

Koocher, G.P., Goodman, G.S., White, C.S., Friedrich, W.N., Sivan, A.B., & Reynolds, C.R. (1995). Psychological science and the use of anatomically detailed in child sexual-abuse assessments. *Psychological Bulletin*, *118*, 119–222.

Koppitz, E.M. (1968). *Psychological evaluation of children's human figure drawing*. NY: Grune & Stratton.

Kostlan, A. (1954). A method for the empirical study of psychodiagnosis. *Journal of Consulting Psychology*, *18*, 83–88.

Kraiger, K., Hakel, M.D., & Cornelius, E.T. (1984). Exploring fantasies of TAT reliability. *Journal of Personality Assessment*, *48*, 365–370.

Krall, V., Sachs, H., Lazar, B., Rayson, B., Growe, G., Novar, L., & O'Connell, L. (1983). Rorschach norms for inner city children. *Journal of Personality Assessment*, *47*, 155–157.

Kroon, N., Goudena, P.P., & Rispens, J. (1998). Thematic Apperception Tests for child and adolescent assessment: A practitioner's consumer guide. *Journal of Psychoeducational Assessment*, *16*, 99–117.

Kubiszyn, T.W., Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., & Eisman, E.J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice*, *31*, 119–130.

Lambert, H.V. (1972). *A comparison of Jane Loevinger's theory of ego development and Lawrence Kohlberg's theory of moral development*. Unpublished doctoral dissertation, University of Chicago, Chicago, IL.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lees-Haley, P.R. (1992). Psychodiagnostic test usage by forensic psychologists. *American Journal of Forensic Psychology*, *10*, 25–30.

Levenberg, S.B. (1975). Professional training, psychodiagnostic skill, and Kinetic Family Drawings. *Journal of Personality Assessment*, *39*, 389–393.

Levy, L.H. (1963). *Psychological interpretation*. New York: Holt, Rinehart, & Winston.

Lewinsohn, P.M. (1964). Relationship between height of figure drawing and depression in psychiatric patients, *Journal of Consulting Psychology*, *11*, 49–54.

Lilienfeld, S.O. (1999). Projective measures of personality and psychopathology: How well do they work? *Skeptical Inquirer*, *23*, 32–39.

Lilienfeld, S.O., Hess, T., & Rowland, C. (1996). Psychopathic personality traits and temporal perspective: A test of short time horizon hypothesis. *Journal of Psychopathology and Behavioral Assessment*, *18*, 285–314.

Lindzey, G. (1959). On the classification of projective techniques. *Psychological Bulletin*, *56*, 158–168.

Lipsey, M.W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park: Sage Publications.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

Little, K.B., & Schneidman, E.S. (1959). Congruencies among interpretations of psychological test and anamnestic data. *Psychological Monographs* (6, Whole No. 476).

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Loevinger, J. (1976). *Ego development: Conceptions and theories*. San Francisco: Jossey-Bass.

Loevinger, J. (1987). *Paradigms of personality*. NY: W.H. Freeman.

Loevinger, J. (1993). Measurement of personality: True or false. *Psychological Inquiry*, *4*, 1–16.

Loevinger, J. (1998). *Technical foundations for measuring ego development: The Washington University Sentence Completion Test*. Mahwah, NJ: Erlbaum.

Lowenstein, L.F. (1987). Are projective techniques dead? *British Journal of Projective Psychology*, *32*, 2–21.

Lucas, R.H. (1971). Validation of a test of ego development by means of a standardized interview (Doctoral dissertation, Washington University, St. Louis). *Dissertation Abstracts International*, *32*, 2204B.

Luscher, M., & Scott, I. (1969). *The Luscher Color Test*. NY: Washington Square Press.

Machover, K. (1949). *Personality projection in the drawing of the human figure*. Springfield, Ill.: Charles C. Thomas.

Magnussen, N.G. (1960). Verbal and non-verbal reinforcers in the Rorschach situation. *Journal of Clinical Psychology*, *16*, 167–169.

Malik, R. (1992). An exploration of object relations phenomena in borderline personality disorder. *Dissertation Abstracts International*, *52* (09B), 4962.

Masling, J. (1960). The influence of situational and interpersonal variables in projective testing. *Psychological Bulletin*, *57*, 65–85.

Masling, J. (1966). Role-related behavior of the subject and psychologist and its effect upon psychological data. In D. Levine (Ed.), *Nebraska symposium on motivation* (pp. 67–104). Lincoln: University of Nebraska Press.

Masling, J. (1997). On the nature and utility of projective tests. *Journal of Personality Assessment*, *69*, 257–270.

Masling, J.M., Rabie, L., & Blondheim, S.H. (1967). Obesity, level of aspiration, and Rorschach and TAT measures of oral dependence. *Journal of Consulting Psychology*, *31*, 233–239.

McArthur, D.S., & Roberts, G.E. (1990). *Roberts Apperception Test for Children manual*. Los Angeles: Western Psychological Services.

McCann, J.T. (1998). Defending the Rorschach in court: An analysis of admissibility using legal and professional standards. *Journal of Personality Assessment*, *70*, 125–144.

McClelland, D.C. (1951). *Personality*. NY: William Sloane Associates.

McClelland, D.C. (1961). *The achieving society*. Princeton: Van Nostrand.

McClelland, D.C. (1980). Motive dispositions: The merits of operant and respondent measures. *Review of Personality and Social Psychology*, *1*, 10–41.

McClelland, D.C. (1989). Motivational factors in health and illness. *American Psychologist*, *44*, 219–233.

McClelland, D.C., Atkinson, J.W., Clark, R.A., & Lowell, E.L. (1953). *The achievement motive*. NY: Appleton-Century-Crofts.

McClelland, D.C., & Boyatsis, R.E. (1982). Leadership motive pattern and long-term success in management. *Journal of Applied Psychology*, *67*, 737–743.

McClelland, D.C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, *96*, 690–702.

McClelland, D.C., Ross, G., & Patel, V.T. (1985). The effect of an examination on salivary norepinephrine and immunoglobulin levels. *Journal of Human Stress*, *11*, 52–59.

McCrae, R.R., & Costa, P.T. (1980). Openness to experiencee and ego level in Loevinger's sentence completion test: Dispositional contributions in developmental models of personality. *Journal of Personality and Social Psychology*, *39*, 1179–1190.

McKinzey, R.K. & Ziegler, T. (1999). Challenging a flexible neuropsychological battery under Kelly/Frye: A case study. *Behavioral Sciences and the Law*, *17*, 543–551.

McNally, R.J. (1996). Cognitive bias in the anxiety disorders. *Nebraska Symposium on Motivation*, *43*, 211–250.

Meehl, P.E. (1945). The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, *1*, 296–303.

Meehl, P.E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology,, 13*, 102–128.

Meehl, P.E. (1986). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In T. Millon & G. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 215–231). NY: Guilford.

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

Meloy, J.R., & Gacono, C.B. (1995). Assessing the psychopathic personality. In J.N. Butcher (Ed.), *Clinical personality assessment* (pp.410–422). NY: Oxford University Press.

Meloy, J.R., & Gacono, C.B. (1998). The internal world of the psychopath. In T. Millon, E. Simonsen, M. Birket-Smith, & R.D. Davis (Eds.), *Psychopathy: Antisocial, criminal, and violent behavior* (pp. 95–109). NY: Guilford Press.

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

Meyer, G.J. (1991). An empirical search for fundamental personality and mood dimensions within the Rorschach test (Unpublished dissertation. Loyola University of Chicago, 1989). *Dissertation Abstracts International*, *52*, 1071B-1072B.

Meyer, G.J. (1992a). Response frequency problems in the Rorschach: Clinical and research implications with suggestions for the future. *Journal of Personality Assessment*, *58*, 231–244.

Meyer, G.J. (1992b). The Rorschach's factor structure: A contemporary investigation and historical review. *Journal of Personality Assessment*, *59*, 117–136.

Meyer, G.J. (1993). The impact of response frequency on the Rorschach constellation indices and on their validity with diagnostic and MMPI-2 criteria. *Journal of Personality Assessment*, *60*, 153–180.

Meyer, G.J. (1996). The Rorschach and MMPI: Toward a more scientifically differentiated understanding of cross-method assessment. *Journal of Personality Assessment*, *67*, 558–578.

Meyer, G.J. (1997a). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment*, *9*, 480–489.

Meyer, G.J. (1997b). Thinking clearly about reliability: More critical corrections regarding the Rorschach Comprehensive System. *Psychological Assessment*, *9*, 495–498.

Meyer, G.J. (2000a). The incremental validity of the Rorschach Prognostic Rating Scale over the MMPI Ego Strength Scale and IQ. *Journal of Personality Assessment*, *74*, 356–370.

Meyer, G.J. (2000b). On the science of Rorschach research. *Journal of Personality Assessment*, *75*, 46–81.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Kubiszyn, T.W., Moreland, K.L., Eisman, E.J., & Dies, R.R. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part I.* Washington, DC: American Psychological Association.

Meyer, G.J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment*, *69*, 1–38.

Morgan, C.D., & Murray, H.A. (1935). A method for investigating fantasies. *Archives of Neurology and Psychiatry*, *34*, 289–304.

Motta, R.W., Little, S.G., & Tobin, M.I. (1993). The use and abuse of human figure drawings. *School Psychology Quarterly*, *8*, 162–169.

Murray, H. (1938). *Explorations in personality*. NY: Oxford University Press.

Murstein, B.T., & Easter, L.V. (1965). The role of achievement motive, anxiety, stimulus, and expectancy, on achievement motivation in arithmetic and thematic tests. *Journal of Projective Techniques and Personality Assessment*, *29*, 491–497.

Murstein, B.I., & Mathes, S. (1996). Projection on projective techniques = pathology: The problem that is not being addressed. *Journal of Personality Assessment*, *66*, 337–349.

Mussen, P.H., & Naylor, H.K. (1954). The relationships between overt and fantasy aggression. *Journal of Abnormal and Social Psychology*, *49*, 235–240.

Naglieri, J.A. (1988). *Draw A Person: A Quantitative Scoring System*. San Antonio, TX: Psychological Corporation.

Naglieri, J.A. (1992). Review of the Hutt Adaptation of the Bender-Gestalt Test. In J.J. Kramer & J.C. Conoley (Eds.), *The Eleventh Mental Measurements Yearbook* (pp. 103–104). Lincoln, NE: Buros Institute of Mental Measurements.

Naglieri, J.A., McNeish, T.J., & Bardos, A.N. (1991). *Draw-A-Person: Screening Procedure for Emotional Disturbance*. Austin, TX: ProEd.

Naglieri, J.A., & Pfeiffer, S.I. (1992). Performance of disruptive behavior-disordered and normal samples on the Draw-A-Person: Screening Procedure for Emotional Disturbance. *Psychological Assessment*, *4*, 156–159.

Nakata, L.M. (1999). Interrater reliability and the Comprehensive System for the Rorschach: Clinical and non-clinical protocols (Doctoral dissertation, Pacific Graduate School of Psychology). *Dissertation Abstracts International.*

Newman, D.L., Tellegen, A., & Bouchard, T.J. (1998). Individual differences in adult ego development: Sources of influence in twins reared apart. *Journal of Personality and Social Psychology*, *74*, 985–995.

Nezworski, M.T., & Wood, J.M. (1995). Narcissism in the Comprehensive System for the Rorschach. *Clinical Psychology: Science and Practice*, *2*, 179–199.

Nichols, R.C., & Strumpfer, D.J. (1962). A factor analysis of Draw-a-Person test scores. *Journal of Consulting Psychology*, *26*, 156–161.

Oberlander, L.B. (1995). Psycholegal issues in child sexual abuse evaluations: A survey of forensic mental health professionals. *Child Abuse & Neglect*, *19*, 475–490.

O'Connell, M., Cooper, S., Perry, J.C., & Hoke, L. (1989). The relationship between thought disorder and psychotic symptoms in borderline personality disorder. *Journal of Nervous and Mental Disease*, *177*, 273–278.

Ordnuff, S.R., Freedenfeld, R., Kelsey, R.M., & Critelli, J. (1994). Object relations of sexually abused female subjects: A TAT analysis. *Journal of Personality Assessment*, *63*, 223–228.

Ordnuff, S.R., & Kelsey, R.M. (1996). Object relations of sexually and physically abused female children: A TAT analysis. *Journal of Personality Assessment*, *66*, 91–105.

Parker, K.C.H., Hanson, R.K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin*, *103*, 367–373.

Parker, K.C.H., Hunsley, J., & Hanson, R.K. (1999). Old wine from old skins sometimes tastes like vinegar: A response to Garb, Florio, and Grove. *Psychological Science*, *10*, 291–292.

Peixotto, H.E. (1980). The Holtzman Inkblot Technique: Some new directions in inkblot testing. *Academic Psychology Bulletin*, *2*, 47–52.

Perez, F.I. (1976). Behavioral analysis of clinical judgment. *Perceptual and Motor Skills*, *43*, 711–718.

Perry, G.G., & Kinder, B.N. (1990). The susceptibility of the Rorschach to malingering: A schizophrenia analogue. In C.D. Spielberger & J.N. Butcher (Eds.), *Advances in personality assessment*, *Volume 9* (pp. 127–140). Hilldale, N.J.: Erlbaum.

Perry, W., Geyer, M.A., & Braff, D.L. (1999). Sensorimotor gating and thought disturbance measured in close temporal proximity in schizophrenic patients. *Archives of General Psychiatry*, *56*, 277–281.

Perry, W., Moore, D., & Braff, D. (1995). Gender differences on thought disturbance: Measures among schizophrenic patients. *American Journal of Psychiatry*, *152*, 1298–1301.

Perry, W., McDougall, A., & Viglione, D.J., Jr. (1995). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment*, *64*, 112–118.

Perry, W., & Viglione, D.J., Jr. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment*, *56*, 487–501.

Pinkerman, J.E., Haynes, J.P., & Keiser, T. (1993). Characteristics of psychological practice in juvenile court clinics. *American Journal of Forensic Psychology*, *11*, 3–12.

Piotrowski, C., & Belter, R.W. (1999). Internship training in psychological assessment: Has managed care had an impact? *Assessment*, *6*, 381–385.

Piotrowski, C., Belter, R.W., & Keller, J.W. (1998). The impact of "managed care" on the practice of psychological testing: Preliminary findings. *Journal of Personality Assessment*, *70*, 441–447.

Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD clinical psychology programs. *Journal of Personality Assessment*, *61*, 394–405.

Popper, K.R. (1959). *The logic of scientific discovery*. NY: Basic Books.

Porcerelli, J.H., Thomas, S., Hibbard, S., & Cogan, R. (1998). Defense mechanisms development in children, adolescents, and late adolescents. *Journal of Personality Assessment*, *71*, 411–420.

Psychological Corporation (1997). *Wechsler Adult Intelligence Scale*, *Third Edition. Wechsler Memory Scale*, *Third Edition. Technical Manual*. San Antonio: Author.

Rabin, A.I. (1968). Projective methods: An historical introduction. In A.I. Rabin (Ed.), *Projective techniques in personality assessment* (pp. 3–17). NY: Grune-Stratton.

Ricks, D., Umbarger, C., & Mack, R. (1964). A measure of increased temporal perspective in successfully treated adolescent delinquent boys. *Journal of Abnormal and Social Psychology*, *69*, 685–689.

Riethmiller, R.J., & Handler, L. (1997a). Problematic methods and unwarranted conclusions in DAP research: Suggestions for improved research procedures. *Journal of Personality Assessment*, *69*, 459–475.

Riethmiller, R.J., & Handler, L. (1997b). The Great Figure Drawing Controversy: The integration of research and clinical practice. *Journal of Personality Assessment*, *69*, 488–496.

Roback, H. (1968). Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. *Psychological Bulletin*, *70*, 1–19.

Ronan, G.F., Colavito, V.A., & Hammontree, S.R. (1993). Personal problem-solving system for scoring TAT responses: Preliminary validity and reliability data. *Journal of Personality Assessment*, *61*, 28–40.

Ronan, G.F., Date, A.L., & Weisbrod, M. (1995). Personal problem-solving scoring of the TAT: Sensitivity to training. *Journal of Personality Assessment*, *64*, 119–131.

Ronan, G.F., Senn, J., Date, A., Maurer, L., House, K., Carroll, J., & van Horn, R. (1996). Personal problem-solving scoring of TAT responses: Known-groups validation. *Journal of Personality Assessment*, *67*, 641–653.

Rorschach, H. (1921). *Psychodiagnostics: A diagnostic test based on perception.* NY: Grune & Stratton.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenzweig, S., Fleming, E.E., & Clark, H.J. (1947). *Revised Scoring Manual for the Rosenzweig Picture-Frustration Study*. St. Louis: Saul Rosenzweig.

Rossini, E.D., & Moretti, R.J. (1997). Thematic Apperception Test (TAT) interpretation: Practice recommendations from a survey of clinical psychology doctoral program accredited by the American Psychological Association. *Professional Psychology: Research and Practice*, *28*, 393–398.

Ryan, R.M. (1985). Thematic Apperception Test. In D.J. Keyser & R.C. Swetland (Eds.), *Test critiques* (Vol. 2, pp. 799–814). Kansas City, MO: Test Corporation of America.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schneider, G.B. (1978). A preliminary validation study of the Kinetic School Drawing. *Dissertation Abstracts International*, *38* (11-A), 6628.

Schretlen, D.J. (1997). Dissimulation on the Rorschach and other projective measures. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed.; pp. 208–222). NY, NY: Guilford.

Schwartz, N.S., Mebane, D.L., & Malony, H.N. (1990). Effects of alternate modes of administration on Rorschach performance of deaf adults. *Journal of Personality Assessment*, *54*, 671–683.

Schweighofer, A., & E.M. Coles (1994). Note on the definition and ethics of projective tests. *Perceptual and Motor Skills, 79*, 51–54.

Scribner, C.M., & Handler, L. (1987). The interpreter's personality in Draw-A-Person Interpretation: A study of interpersonal style. *Journal of Personality Assessment*, *51*, 112–122.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, *12*, 153–158.

Sechrest, L., & Wallace, J. (1964). Figure drawings and naturally occurring events. *Journal of Educational Psychology*, *55*, 42–44.

Shaffer, T.W., Erdberg, P., & Haroian, J. (1999). Current nonpatient data for the Rorschach, WAIS-R, and MMPI-2. *Journal of Personality Assessment*, *73*, 305–316.

Sharkey, K.J., & Ritzler, B.A. (1985). Comparing diagnostic validity of the TAT and a new picture projection test. *Journal of Personality Assessment*, *49*, 406–412.

Silverton, L. (1993). *Adolescent Apperception Cards: Manual*. Los Angeles: Western Psychological Services.

Sines, L.K. (1959). The relative contribution of four kinds of data to accuracy in personality assessment. *Journal of Consulting Psychology*, *23*, 483–492.

Skinner, B.F. (1938). *The behavior of organisms*. NY: Appleton-Century-Crofts.

Smith, D., & Dumont, F. (1995). A cautionary study: Unwarranted interpretations of the Draw-A-Person Test. *Professional Psychology: Research and Practice*, *26*, 298–303.

Solovay, M., Shenton, M., Gasperetti, C., Coleman, M., Daniels, E., Carpenter, J., & Holzman, P. (1986). Scoring manual for the Thought Disorder Index. *Schizophrenia Bulletin*, *12*, 483–496.

Soskin, W.F. (1954). Bias in postdiction from projective tests. *Journal of Abnormal and Social Psychology*, *49*, 69–74.

Spangler, W.D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin*, *112*, 140–154.

Starr, B.J., & Katkin, E.S. (1969). The clinician as an aberrant actuary: Illusory

correlation and the Incomplete Sentences Blank. *Journal of Abnormal Psychology*, *74*, 670–675.

Sternberg, R.J. (1986). *Intelligence applied: Understanding and increasing your intellectual skills*. NY: Harcourt, Brace Jovanovich.

Stricker, G. (1967). Actuarial, naïve clinical, and sophisticated clinical prediction of pathology from figure drawings. *Journal of Consulting Psychology*, *31*, 492–494.

Stricker, G., & Gold, J.R. (1999). The Rorschach: Toward a nomothetically based, idiographically applicable configurational model. *Psychological Assessment*, *11*, 240–250.

Suinn, R.M., & Oskamp, S. (1969). *The predictive validity of projective measures: A fifteen year evaluative review of research*. Springfield, IL.: Charles C. Thomas.

Sundberg, N. (1977). *Assessment of persons*. Englewood Cliffs, N.J.: Prentice Hall.

Sutherland, S. (1992). *Irrationality: Why we don't think straight!* New Brunswick, NJ: Rutgers University Press.

Swenson, C. (1957). Empirical evaluations of human figure drawings. *Psychological Bulletin*, *54*, 431–466.

Swenson, C. (1968). Empirical evaluations of human figure drawings: 1957–1966. *Psychological Bulletin*, *70*, 20–44.

Szondi, L. (1947). *Experimentelle Triebsdiagnostik*. Bern: Verlag Hans Huber.

Tellegen, A., & Waller, N.G. (1994). Exploring personality through test construction: In S.R. Briggs & J.M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1, pp. 133–161). Greenwich, CT.: JAI Press.

Tharinger, D.J., & Stark, K. (1990). A qualitative versus quantitative approach to evaluating the Draw-A-Person and Kinetic Family Drawing: A study of mood and anxiety-disordered children. *Psychological Assessment*, *2*, 365–375.

Thomas, C.B., Jones, L.W., & Ross, D.C. (1968). Studies on figure drawings: Biological implications of structural and graphic characteristics. *Psychiatric Quarterly Supplement*, *42*, 223–251.

Thomas, G.V., Chaigne, E., & Fox, T.J. (1989). Children's drawings of topics differing in significance: Effects on size of drawing. *British Journal of Developmental Psychology*, *7*, 321–331.

Thomas, G.V., & Jolley, R.P. (1998). Drawing conclusions: A re-examination of empirical and conceptual bases for psychological evaluations of children from their drawings. *British Journal of Clinical Psychology*, *37*, 127–139.

Tolor, A., & Digrazia, P.V. (1977). The body image of pregnant women as reflected in their human figure drawings. *Journal of Clinical Psychology*, *34*, 537–538.

Trowbridge, M.M. (1995). Graphic indicators of sexual abuse in children's drawings: A review of the literature. *The Arts in Psychotherapy*, *22*, 485–493.

Vaillant, G.E. (1977). *Adaptation to life*. Boston: Little, Brown.

Vane, J.R. (1981). The Thematic Apperception Test: A review. *Clinical Psychology Review*, *1*, 319–336.

Vass, Z. (1998). The inner formal structure of the H-T-P drawings: An exploratory study. *Journal of Clinical Psychology*, *54*, 611–619.

Veiel, H., & E.M. Coles (1982). Methodological ambiguities of the projective technique: An overview and attempt to clarify. *Perceptual and Motor Skills*, *54*, 443–450.

Viglione, D.J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment*, *11*, 251–265.

Viney, L.L., Aitken, M., & Floyd, J. (1974). Sel-regard and size of human figure drawings: An interactional analysis. *Journal of Clinical Psychology*, *30*, 581–586.

Wade, T.C., & Baker, T.B. (1977). Opinion and use of psychological tests. *American Psychologist*, *32*, 874–882.

Waehler, C.A. (1997). Drawing bridges between science and practice. *Journal of Personality Assessment*, *69*, 482–487.

Wagner, E.E. (1962). The use of drawings of hands as a projective medium for differentating neurotics and schizophrenics. *Journal of Clinical Psychology*, *18*, 208–209.

Waldman, I.D. (1996). Aggressive boys' hostile perceptual and response biases: The role of attention and impulsivity. *Child Development*, *67*, 1015–1033.

Wanderer, Z.W. (1969). Validity of clinical judgments based on human figure drawings. *Journal of Consulting and Clinical Psychology*, *33*, 143–150.

Watkins, C.E., Campbell, V.L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice*, *26*, 54–60.

Wecshler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Weiner, I.B. (1996). Some observations on the validity of the Rorschach Inkblot Method. *Psychological Assessment*, *8*, 206–211.

Weiner, I.B. (1997). Current status of the Rorschach Inkblot Method. *Journal of Personality Assessment*, *68*, 5–19.

Weiner, I.B. (1998). *Principles of Rorschach interpretation.* Mahwah, NJ: Erlbaum.

Weiner, I.B. (1999). What the Rorschach can do for you: Incremental validity in clinical applications. *Assessment*, *6*, 327–338.

Weiner, I.B. (2000). Using the Rorschach properly in practice and research. *Journal of Clinical Psychology*, *56*, 435–438.

Weiner, I.B., Exner, J.E., & Sciara, A. (1996). Is the Rorschach welcome in the courtroom? *Journal of Personality Assessment*, *67*, 422–424.

West, M.M. (1998). Meta-analysis of studies assessing the efficacy of projective techniques in discriminating child sexual abuse. *Child Abuse & Neglect*, *22*, 1151–1166.

Westen, D. (1991). Clinical assessment of object relations using the TAT. *Journal of Personality Assessment*, *56*, 56–74.

Westen, D., Lohr, N., Silk, K., Kerber, K., & Goodrich, S. (1985). *Measuring object relations and social cognitions using the TAT: Scoring manual.* Ann Arbor: University of Michigan.

Westen, D., Lohr, N., Silk, K.R., Gold, L., & Kerber, K. (1990). Object relations and social cognition in borderlines, major depressives, and normals: A Thematic Apperception Test analysis. *Psychological Assessment*, *2*, 355–364.

Westen, D., Ludolph, P., Block, M.J., Wixom, J., & Wiss, F.C. (1990). Developmental history and object relations in psychiatrically disturbed adolescent girls. *American Journal of Psychiatry*, *147*, 1061–1068.

Westen, D., Ludolph, P., Lerner, H., Ruffins, S., & Wiss, C. (1990). Object relations in borderline adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, *29*, 338–348.

Westenberg, P.M., & Block, J. (1993). Ego development and individual differences in personality. *Journal of Personality and Social Psychology*, *65*, 792–800.

Whitehead, W.C. (1985). *Clinical dissertation making on the basis of Rorschach, MMPI, and automated MMPI report data*. Unpublished doctoral dissertation, University of Texas Health Science Center at Dallas.

Wildman, R.W., & Wildman, R.W. II. (1975). An investigation into the comparative validity of several diagnostic tests and test batteries. *Journal of Clinical Psychology*, *31*, 455–458.

Wilson, J.Q., & Herrnstein, R.J. (1985). *Crime and human nature*. NY: Simon & Schuster.

Winch, R.F., & More, D.M. (1956). Does TAT add information to interviews? Statistical analysis of the increment. *Journal of Clinical Psychology*, *12*, 316–321.

Winter, D.G. (1973). *The power motive*. NY: The Free Press.

Winter, D.G., John, O.P., Stewart, A.J., Klohnen, E.C., & Duncan, L.E. (1998). Traits and motives: Toward an integration of two traditions in personality research. *Psychological Review*, *105*, 230–250.

Winter, D.G., & Stewart, A.J. (1977). Power motive reliability as a function of retest instructions. *Journal of Consulting and Clinical Psychology*, *45*, 436–440.

Wolfner, G., Faust, D., & Dawes, R.M. (1993). The use of anatomically detailed dolls in sexual abuse evaluations: The state of the science. *Applied and Preventive Psychology*, *2*, 1–11.

Woltmann, A.G. (1960). Spontaneous puppetry by children as a projective method. In A.I. Rabin & M.R. Haworth (Eds.), *Projective techniques with children* (pp. 305–312). NY: Grune & Stratton.

Wood, J.M., & Lilienfeld, S.O. (1999). The Rorschach Inkblot Test: A case of overstatement? *Assessment*, *6*, 341–349.

Wood, J.M., Lilienfeld, S.O., Garb, H.N., & Nezworski, M.T. (2000a). The Rorschach Test in clinical diagnosis: A critical review, with a backward look at Garfield (1947). *Journal of Clinical Psychology*, *56*, 395–430.

Wood, J.M., Lilienfeld, S.O., Garb, H.N., & Nezworski, M.T. (2000b).

Limitations of the Rorschach as a diagnostic tool: A reply to Garfield (2000), Lerner (2000), and Weiner (2000). *Journal of Clinical Psychology*, *56*, 441–448.

Wood, J.M., Nezworski, M.T., Garb, H.N., & Lilienfeld, S.O. (in press). The misperception of psychopathology: Problems with the norms of the Comprehensive System for the Rorschach. *Clinical Psychology: Science and Practice*.

Wood, J.M., Nezworski, M.T., & Stejskal, W.J. (1996a). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science*, *7*, 3–10.

Wood, J.M., Nezworski, M.T., & Stejskal, W.J. (1996b). Thinking critically about the Comprehensive System for the Rorschach. A reply to Exner. *Psychological Science*, *7*, 14–17.

Wood, J.M., Nezworski, M.T., & Stejskal, W.J. (1997). The reliability of the Comprehensive System for the Rorschach: A Comment on Meyer (1997). *Psychological Assessment*, *9*, 490–494.

Wood, J M., Nezworski, M.T., Stejskal, W.J., & McKinzey, R.K. (in press). Problems of the Comprehensive System for the Rorschach in forensic settings: Recent developments. *Journal of Forensic Psychology Practice.*

Worchel, F.F. & Dupree, J.L. (1990). Projective story-telling techniques. In C.R. Reynolds & R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior*, *& context* (pp. 70–88). NY: Guilford.

Zeldow, P.B., & McAdams, D.P. (1993). On the comparison of TAT and free speech techniques in personality assessment. *Journal of Personality Assessment*, *60*, 181–185.