JOURNAL OF PERSONALITY ASSESSMENT, *80*(3), 260–271

A Critique of Lilienfeld et al.'s (2000) "The Scientific Status of Projective Techniques"

Stephen Hibbard
Department of Psychology
University of Windsor

Lilienfeld, Wood, and Garb (2000) published a largely negative critique of the validity and reliability of projective methods, concentrating on the Comprehensive System for the Rorschach (Exner, 1993), 3 systems for coding the Thematic Apperception Test (TAT; Murray, 1943) cards, and human figure drawings. This article is an effort to document and correct what I perceive as errors of omission and commission in the Lilienfeld et al. article. When projective measures are viewed in the light of these corrections, the evidence for the validity and clinical usefulness of the Rorschach and TAT methods is more robust than Lilienfeld et al. represented.

Lilienfeld, Wood, and Garb (2000) recently rendered a largely negative pronouncement on "The Scientific Status of Projective Techniques." They focused primarily on Exner's (1993) Comprehensive System (CS) for the Rorschach, three contemporary coding systems for the Thematic Apperception Test (TAT; Murray, 1943), and various systems for coding human figure drawings. The major point of this critique is to document scientific errors of omission and commission in Lilienfeld et al.'s article. Their piece was placed in a new monograph series that purports to monitor scientific psychology for the sake of the public, and hence, reasoned criticism of their article may itself serve the public interest. On the basis of their arguments, they made recommendations that would dramatically constrain if not eliminate some projective methods in training programs and in practice. Because of this, it is important to investigate the scientific merit of their work. The stated goal of Lilienfeld et al. (2000) was "to examine impartially the best available research evidence concerning the scientific status" (pp. 27–28) of these measures. My criticism can be viewed then as observations on the extent to which they may have failed in that stated goal. In a few instances, I present new data that Lillienfeld et al. could not have had access to in presenting their critique. However, I have ignored published papers, no matter how relevant, that have emerged since the publication of Lilienfeld et al.

In addition to Exner's (1993) CS, Lilienfeld et al. (2000) discussed three TAT coding systems. These are McClelland and followers' (Smith, 1992) motive coding, Westen, Silk, Lohr, & Kerber's (1989) Object Relations and Social Cognition scales (ORSC[1]), and Cramer's (1991b) Defense Mechanism Manual (DMM). Lilienfeld et al. also evaluated figure drawings and referred to other Rorschach indexes. Because I have greater expertise with the Rorschach and TAT than with figure drawings, I do not discuss the last here.

ON THE DESCRIPTORS, PROJECTIVE TECHNIQUES, AND PROJECTIVE RATIONALE

Lilienfeld et al.'s (2000) view of projective tests seems to rest on an outmoded conception of the role of projection. In an early section of Lilienfeld et al.'s (2000) article titled "Primer of Projective Techniques and Their Rationale" (pp. 28–29), they stated the following:

> [T]he rationale underlying most projective techniques is the projective hypothesis. … 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 psycho-

[1]Westen's (Westen, Silk, Lohr, & Kerber, 1989) original manual for coding TATs was called the *Object Relations and Social Cognition* or ORSC manual. Since then, variations on this including Q-sort techniques and techniques for coding interview data have been developed, and the name has been reversed to *Social Cognition and Object Relations Scales* or SCORS.

> logical information (e.g., conflicts, impulses) of which respondents are not consciously aware (Dosajh, 1996). (p. 29)

Thus, Lilienfeld et al. (2000) attributed to unnamed individuals (notably not to those they ultimately criticize) a typical hypothesis about how projective tests work. Lilienfeld et al. cited as the source for this typical model an author (Dosajh, 1996) who has not published on the coding systems targeted for criticism. The journal from which the quote is taken has a very limited distribution in the western hemisphere. Examination of the actual writings of Exner, McClelland, Westen, and Cramer suggests that they did not espouse a view of projection similar to that of Dosajh. That is, these authors do not see projection in Dosajh's sense as having an important role in the Rorschach or the TAT.

Exner (1993, with qualification) and McClelland (and his followers; Smith, 1992) explicitly denied that their systems are projective. Westen (1991; Westen, Klepser, et al., 1991) has an explicit account of representations as "working models" in his TAT system that is foreign to Lilienfeld et al.'s (2000) caricature of the projective rationale, and although Cramer (1991b) at times showed interest in the projective hypothesis, her coding system does not operationalize Lilienfeld et al.'s caricature. None of these authors coded "conflicts or impulses," and none were concerned with privileged access or "bypassing conscious defenses," whatever those might be. Two of the systems that Lilienfeld et al. subsequently criticized explicitly limit or entirely eschew the role of projection in their systems (see, e.g., Exner, 1993, p. 53 and following; Smith, 1992, p. 5; Winter, John, Stewart, Klohnen, & Duncan, 1998, footnote 3). In fact, Lilienfeld et al. (2000, p. 31) later stated as much in the case of Exner, a statement that only makes their rationale passage more confusing. The theoretical underpinnings of Westen's (1991; Westen et al., 1989) system are committed not to projection, but rather to paradigms (the activation of working models) borrowed from the social cognitive and attachment literatures (Westen, Klepser, et al., 1991). Cramer's system, although thoroughly psychoanalytic, does not code any privileged conflicts or impulses. In other words, the projective hypothesis as outlined by Lilienfeld et al. seems to have no application to the coding systems that are subsequently criticized by them.

I now take up Lilienfeld et al.'s (2000) critical remarks about the Rorschach and the TAT in very much the same order they were elaborated in the original monograph.

THE RORSCHACH

CS Normative Data

Lilienfeld et al. (2000) criticized the CS normative data. Lilienfeld et al. claimed that (a) the norms for many CS variables are in error, (b) these discrepancies overpathologize normal individuals, and (c) the use of the CS norms in clinical or forensic settings may harm clients and be unethical (p. 32). Indeed, at the end of the 1990s, Rorschach workshops set out to develop new normative data for the CS, in part because of reports by CS users that the extant normative data erred in the direction of being too "healthy" in various respects. The re-norming effort is ongoing.

With regard to the extent and basis of error in the norms, evidence on this issue is yet inconclusive and wise observers will withhold judgment. Greene (2000) concluded that the restandardization of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) resulted in surprisingly little difference in the norms. Discrepancies between the CS norms and data in disparate reports or in local normative studies may have many explanations. Lilienfeld et al. (2000) cited Shaffer, Erdberg, & Haroian (1999) as an example of such a study, and the latter paper reported discrepancies between the Shaffer et al. data and Exner (1993) norms for several CS indexes and scores. Salient examples include Reflection responses (an index of narcissism), Schizophrenia Index (SCZI; a disordered thinking index) and T (an index of comfort or discomfort in close relationships), the first two of which were discussed by Lilienfeld et al. One possible explanation for such discrepancies might be that Exner's (1993) normative sample likely is somewhat healthier than the general population because individuals with any psychiatric history were excluded from Exner's normative group. Again, for some of the most important discrepancies, the normative data have not kept up with (i.e., not been rescored for) changes in certain scoring criteria (viz., especially form quality). That is, changes have been made for how to score ordinary and unusual form level in particular, and (because of the enormity of the task) the normative data have not been rescored according to changes in contemporary scoring criteria (Meyer & Richardson, 2001). Therefore, the published norms overstate to some unknown extent normative X + % and understate normative Xu%.

Lilienfeld et al. (2000) were concerned that CS norms overpathologize "normals." Changes in normative data as such do not always entail that the use of older norms results in overpathologizing contemporary evaluees. This is especially likely when societal changes reflect psychopathology, that is, in essence, if society becomes more pathological. Lilienfeld et al. did not consider that if norm changes drift in pathological directions, this may be a result of pathological forces in the society. It is possible that new normative data (higher or lower means or modal scores) may in fact reflect increased psychopathology in the society. In fact, increases in the past few decades in some forms of pathology are quite possible. A good case can be made that the relevant extra test correlates of at least two CS variables, Texture and Reflections, have been affected by just such societal changes. Of importance, these are two of the CS indexes for which

Shaffer et al. (1999) found discrepancies. Increases in the divorce rate over the relevant decades are well documented (United States Census Bureau, 2001, Table 68, p. 59). According to the empirical studies cited by Exner (1993, pp. 383–385), more than one Texture response is found in the protocols of individuals who have recently divorced or separated. Also, zero Texture response protocols are more often rendered by those who as children have had extended parental absences and those who were in foster home placements. Increases in divorce rates, then, might quite reasonably account for the decrease in frequency of T = 1 protocols. T = 1 was described in Exner (1993) as modal and normative. Such changes, however, would not make the conditions putatively indexed by T > 1 or < 1 any less pathological. Likewise, personality and social psychologists have written extensively on increases in narcissism (Lasch, 1979; Wallach & Wallach, 1983), which is associated with problems in self-regard. The presence of a Reflection response in a protocol is interpreted in the CS as precisely above average levels of self-regard. Hence, at least in respect to Textures and Reflections, when Lilienfeld et al. claimed that increases in normative levels of these variables leads to overpathologizing examinees, they overlooked the possibility that these forms of pathology have in fact increased. If this is so, to use new normative data reflective of these increases would itself lead to faulty interpretation. Of course I am not arguing here that these same considerations (regarding societal change) apply to all of the CS indexes for which Shaffer et al. (1999) found discrepancies. However, I am pointing to plausible cases.

Beyond this, Lilienfeld et al. (2000) did not consider that the quality of the CS normative data as described in Exner (1993) is superior to any of the data sets that they cited. The normative data were gathered by well-trained administrators and at sites throughout the United States. On these bases, at least, the presumption would be in favor of the normative data as being more representative than loosely assembled samples.

The Question of Cultural Bias in the CS

Lilienfeld et al. (2000) argued that the evidence in support of the generalizability of the CS across different cultures is very limited. This point is true: There is a need to greatly expand the range of validity studies of the CS variables in different cultures. However, much the same could have been said about the MMPI–2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) up until the past few years as newer studies have begun to emerge. The lack of a large number of cultural generalizability studies does not in itself provide evidence of bias in a test, and a frequent practice is to apply a normed test (such as the MMPI–2 or the Wechsler tests) to ethnically different populations with caution (Jones & Thorne, 1987). A naive reader of Lilienfeld et al. might get the impression that the CS violates expectable ethical constraints regarding its use in non-White populations: "Use of the CS with American minorities and non-Americans can be highly problematic" (p. 33). Yet in a sense, of course, the use of any test on a population ostensibly different from the one for which we have validity data can be highly problematic. This alone does not establish that the test should not be used because it is biased.

Lilienfeld et al. (2000) offered as empirical evidence of their fears about the generalizability of the Rorschach's validity not differential validity studies (they rightly said these are needed) but rather three studies (Boscan, 1999/2000; Glass, Bieba, & Tkachuk, 1999; Krall et al., 1983) in which the authors compared some ethnic group of interest (inner-city Black children, Alaskan Native prisoners, Mexican college students) to the extant CS norms and found divergence between the minority scores and the norms. However, these citations do not support findings of cultural bias with the Rorschach, nor do they provide a reason to believe that the CS norms are not culturally generalizable. Lilienfeld et al. cited these mean difference studies as if they believed them to be relevant to the cultural generalizability of the CS, but they gave no further explanation as to how they might be relevant. Therefore, it is important for one to examine the methodology of the studies to see whether they were in fact intended to show cultural bias and whether they succeeded.

Construed as evidence for doubting the cultural generalizability of the CS, these studies actually would commit three methodological errors. First, four separate articles in a recent CS research volume (Exner, 1995) point out that statistical comparisons between some research group of interest and the CS normative group as a control are methodologically inappropriate. (By the way, the same point is generalizable to the use of the normative sample of any widely normed test.) This is because narrowly defined groups (inner-city Blacks, Alaskan Native prisoners, Mexican college students) are likely to differ from the test's normative group on all manner of demographics such as age, sex, regional location, interest, acculturation, socioeconomic status, and so forth. It could very well be that these demographic variables mediate the observed differences rather than the relevant group differences such as culture. Curiously, Lilienfeld et al. (2000) showed elsewhere that they were aware of this methodological point. It is unclear as to why they did not choose to mention this point in the present context. Second, Boscan (1999/2000), Glass et al. (1999), and Krall et al. (1983; and seemingly Lilienfeld et al.) regarded prisoners' mean scores being discrepant with CS norms as evidence of racial bias in the CS. Although they simultaneously express the view that they think the CS norms are inaccurate or invalid, they contradict themselves in appealing to the validity and accuracy of those norms to support the meaningfulness of their claim of group differences. Surely it is methodologically wrong to impugn the integrity of data on one hand and to simultaneously appeal to that data to argue for another point (nongeneralizability) on the other. If the norms are mistaken, they cannot reasonably be ap-

pealed to to establish a discrepancy between normative data and some other group of interest.

Of the studies cited, only the comparison of the Alaskan native prisoners with the nonprisoners (Glass et al., 1999) does not rely on comparisons with the CS norms. Also, in this study, these same two groups likewise differed on the Millon Clinical Multiaxial Inventory–II (MCMI–II; Millon, 1987). Does this constitute evidence that use of the MCMI–II is problematic with Alaskan Natives, as such? No, it does not. As the third methodological error, even if these studies had made comparisons to appropriate groups (i.e., not to the CS normative groups), differences between group means on some variable of interest are not psychometrically sound evidence for bias with the test (Anastasi & Urbina, 1997; Hibbard, Tang, et al., 2001; McNulty, Graham, Ben-Porath, & Stein, 1997; Timbrook & Graham, 1994). Rather, they are prima facie evidence for a difference between the two groups. Thus, Johnson and Sikes (1965) scored TAT stories told by African Americans, Mexican Americans, and Whites on a scale of family cohesion. Johnson and Sikes found that the Mexican Americans had the highest scores. They correctly interpreted their findings as indicative of greater family cohesion among the Mexican Americans, not as evidence of cultural bias in the coding system.

As many texts point out (Anastasi & Urbina, 1997), test bias is investigated by testing how differences in group membership (different ethnicities, gender, etc.) moderate the relation between the test and the criterion or whether the test predicts different levels of the criterion in the two groups. None of the empirical studies Boscan (1999/2000), Glass et al. (1999), and Krall et al (1983) adduced by Lilienfeld et al. (2000) in regard to the CS investigates differential validity; rather, each cites group mean differences. These studies are not methodologically appropriate to investigate the cultural generalizability of CS variables.

Aggregated Interrater Reliability Scores Versus Reliability at the Level of the Response

Lilienfeld et al. (2000) criticized the interrater reliability of the CS, a discussion I have not joined in this article. However, in footnote 2, Lilienfeld et al. seem not to have understood a point previously made by Meyer (1997). This is the point that aggregated scores, that is, the scores of each individual summed up over his or her protocol, are likely to show higher reliabilities than when reliability is assessed at the level of the individual response. For the former evaluation, the intraclass correlation (ICC) is the appropriate statistic, whereas for the latter, κ is appropriate. Meyer's view was that summary scores are more reliable than response-level data. Lilienfeld et al. claimed that the findings of Acklin, McDowell, Verschell, and Chan (2000) refute Meyer's claim. (Curiously, Lilienfeld et al. [2000, p. 55] elsewhere showed clearly that they understand the general principle.) An empirical test of this issue is easily available by comparing the κ values of those variables in Acklin et al.'s Table 1 (response level data) to the ICC values of those same variables in their Table 2 (aggregated across individual protocols). I found 46 such pairs of coefficients, and for 36 of them (78%), the ICC is larger than the κ, supporting Meyer's view. Indeed, Acklin et al. (2000, pp. 29, 32) reported means for two samples of both ICCs and κs, and the ICCs are the higher means in both samples. It is not clear whence Lilienfeld et al. derived their conclusions on this matter, but they seem to be in error.

Test–Retest Studies

Lilienfeld et al. (2000) were critical of Exner's (1993) claims of good test–retest reliability coefficients for many CS variables. Lilienfeld et al. implied that the generalizability of Exner's figures is questionable by claiming that when researchers other than Exner report test–retest reliability, the results are lower. To support this claim, Lilienfeld et al. cited four studies, three of which appear in refereed journals. However, when one consults the cited studies, none of them is a bona fide test–retest study. Of the three, the first (Adair & Wagner, 1992) is a study of CS Rorschach variables indicative of thought disorder. Participants for the study are outpatient schizophrenics, all under treatment. The retest interval ranged from 1 to 18 years, with a mean of 6.4 years. The second study (Perry, McDougall, & Viglione, 1995) considers stability of a non-CS variable, which is nonetheless a composite of CS variables. Patients in this study are individuals who were first tested while hospitalized in acute states of serious mental illness and who were retested 5 years later after extensive treatment. Participants in the third study (Schwartz, Mebane, & Malony, 1990) were all deaf, and what Lilienfeld et al. described as a test–retest coefficient was computed over two different and nonstandard methods of Rorschach administration (signing and written English).

These are not bona fide test–retest studies of the CS. In no case did the authors of the original studies themselves regard the indexed coefficients as test–retest coefficients, nor were they reported as such. In the computation of test–retest reliability, standard practice (Anastasi & Urbina, 1997) is to identify possible extra-test sources of unreliability such as treatment and maturation. Alternatively, if the purpose (as it was for Adair & Wagner, 1992, and for Perry et al., 1995) is to demonstrate the stability of a trait or condition over time, interventions, maturation, and so forth, then the proper thing to report is not test–retest reliability of the test but stability of the trait or condition. Lilienfeld et al. (2000) omitted this information. In contrast, the Schwartz et al. (1990) study was an effort to explore which nonstandard method of administration might be best to use with a population whose disability precludes standard administration. Lilienfeld et al.'s description of the data as evidence for test–retest reliability is in error.

Response Frequency and CS Indexes

Lilienfeld et al. (2000) noted that the free response format of the Rorschach creates the possibility that variable scores will be determined not by the presence in the respondent of the features indexed by a variable but by irrelevant influences determining the number of responses (R). Lilienfeld et al. discounted the position of Groth-Marnat (1984) and others who pointed out that because most CS interpretation is based primarily on indexes, ratios, and percentages, free response on the Rorschach has an attenuated effect on CS scores. Lilienfeld et al. cited Meyer (1993) who, in a clinical sample, found Pearson correlations ranging from .23 to .60 between R and the sums of various indexes developed by Exner to provide cutoffs for clinically relevant conditions. These indexes were the SCZI, DEPI, CDI, HVI, OBS, and the SCON. Lilienfeld et al., however, seem to have overlooked the points that (a) Meyer selected these participants in such a way that high and low R scorers were overrepresented, and (b) therefore, the correlations he reported are not accurate indicators of the associations in the population.

Lilienfeld et al. (2000) also seem to have overlooked an even more crucial point regarding these indexes. For the most part, interpretively, Exner did not treat these as continuous variables but rather dichotomous (to some extent, the SCZI is an exception; see Exner, 1993). The indexes are positive if a certain cutoff is attained and otherwise negative. To see the difference this makes, in a data set of 1,152 adult CS protocols taken from various sources,[2] I first computed correlations between R and the continuous form of the previously mentioned indexes. They ranged from –.18 (CDI) to .45 (HVI and OBS). When the appropriate transformations were made of the indexes into dichotomous (negative–positive) values, the correlations with R ranged from –.15 (CDI) to .26 (SCZI). Thus, only SCZI shares any appreciable amount of its variance with R.

Lilienfeld et al. (2000) also appealed to Meyer's (1992) factor analysis of the Rorschach and the conclusions he drew from it concerning the strength of association between R and other Rorschach variables. Lilienfeld et al. quoted Meyer to support their view that because R is strongly associated with unacceptable amounts of variance in Rorschach scores, it ought to be controlled. This conclusion was based on the extent of R's loadings on the first and sometimes also the second factor in factor analytic studies (including Meyer, 1992) of the Rorschach. However, Meyer reached this conclusion based on the amount of common variance accounted for by these factors, not the amount of total test variance. This implies that of the variance in common to those Rorschach variables that were included in these analyses, R is indeed associated with the majority of it. However, this fact does not support the conclusion that Lilienfeld et al. endorsed, namely, that R's influence on Rorschach scores is unacceptably high and amounts to measurement error. This is because the factor analytic considerations leave unexplained the amount of true unique variance that remains in the variables. R's loading on a factor is uninformative about the amount of total test variance associated with these factors. In most Rorschach factor analyses (including Meyer, 1992), R typically is associated with a small amount (about 10%) of the total variance across all the variables. Hence, what Lilienfeld et al. proffered as an estimate of all the variance associated with R is in fact inflated, and the conclusion they drew would seem to be mistaken.

TAT

Reliability of Three TAT-Based Coding Systems

General considerations about reliability. Lilienfeld et al. (2000) called into question the reliability of the three types of TAT scales they considered in the article, beginning with the scales for need for achievement, need for intimacy, and so forth, contemporary versions of a number of which can be found in Smith (1992). Lilienfeld et al. endorsed the view of Entwisle (1972) that there can be no point in aggregating scores into a scale in the absence of applying internal consistency reliability criteria. This view seems to be mistaken. Consider that each subunit of an aggregated group of predictors of a construct could be unrelated to the other, but when found in combination, they might well predict important variance in a construct. For example, bipedalism and featherlessness are probably largely unrelated, but their combination predicts membership in the species. One might likewise discover various weighted aggregations of the Big Five personality factors (which are by definition uncorrelated) to predict various personality features, for example, to predict one or more personality disorder scores or to predict scores on other popular personality scales (O'Conner, 2002). One could then develop Big Five scales consisting of five uncorrelated items that would in turn predict these other scores. As Lundy (1985) noted long ago, these considerations underlie regression approaches to measurement in which scores on variables are aggregated to predict a sum, even though the variables themselves are unrelated. As well, this conception can be applied to diagnostic systems that are based on the specification of different features or symptoms, the satisfaction of which entails a diagnosis, even if the symptoms themselves may be relatively uncorrelated.[3] Inter-

[2]Contributing clinicians are highly skilled, but not all protocols were available for checking interrater reliability. I thank the following contributors: Jacqueline Singer, Carl Hoppe, Thomas Kordinak, Jay Livingston, Margaret Lee, Nancy Oleson, and Marjorie Walters.

[3]For example, the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed. [*DSM–IV*]; American Psychiatric Association, 1994) is such a system. Although it is possible to think of *DSM* fea-

nal consistency is not a consideration in these cases. Indexes with cutoffs such as the CS indexes themselves or Goldberg's MMPI-based index (Greene, 2000) is another example. There seems to be no shortage of instances in clinical psychology in which there is, in fact, utility in aggregating data in the absence of internal consistency.

There are other commentators in the literature on the limits to the claims of internal consistency whose remarks are not considered by Lilienfeld et al. (2000). Karon (1968) argued that because according to classical test theory, reliability constrains validity, if validity is adequate, independent evidence for reliability is a secondary issue. Lundy (1985) gave an empirical demonstration of the point that α need not limit other forms of reliability such as test–retest. Using four cards, Lundy demonstrated greater levels of test–retest reliability (*r*s = .48 to .56 over 1 year) than α reliability (*M* = .16) for motive measures. These test–retest *r*s are comparable to those for standard paper-and-pencil measures (Roberts & DelVecchio, 2000), even though α for the four TAT cards is not as high as typical self-report scales. In other words, adequate levels of reliability for the TAT are not necessarily constrained by the α level.

Lundy's (1985) work was inspired by and also replicated an earlier article by Winter and Stewart (1977) in which they argued (and it was observed that) higher retest reliability would accrue to motive measures if the retest instructions permitted participants to tell stories with the same content as previously. Lilienfeld et al. (2000) cited as counter evidence to Winter and Stewart a different study (Kraiger, Hakel, & Cornelius, 1984) that only partially replicated Winter and Stewart's findings, but Lilienfeld et al. failed to cite Lundy's work. The test–retest sample size in Lundy's study was over twice that of Kraiger et al.'s. Lilienfeld et al. also failed to discuss Reuman's (1982) related studies that cast doubt on the usefulness of internal consistency measures for TAT scales.

Much has been written recently outside TAT circles about the relevance of internal consistency to scale construction, especially about coefficient α (see, e.g., Clark & Watson, 1995, Peterson, 1994, Schmitt, 1996). Lilienfeld et al. (2000) also failed to discuss this literature. Some of these discussions are quite relevant to concerns with internal consistency of the TAT, although they are not specific to the TAT. First, α depends on the number of items in a test. Second, α can be quite high when a test's homogeneity is quite low. Third, item homogeneity and α level can be arbitrarily increased simply by increasing the quantity of semantically redundant items, and correlatively, α can be driven up by defining concepts very narrowly. The converse of this also holds, namely, the broader a construct, the more items will be required to cover it and all else being equal, the lower the alpha. Even if one insists that some minimal level of α be achieved (Clark & Watson suggested .80), striving for higher levels of α actually runs a strong risk of constraining validity by narrowing construct coverage. The upshot is that experts in this field cannot agree on a principled answer or recommendation as to what value of α is satisfactory, if any.

Some of these same considerations transfer to criticisms of test–retest reliability. Low numbers of items in a test constrain not only α but also test–retest reliability. Almost all TAT studies have been done using very few cards, standardly using four cards in McClelland (Smith, 1992) motive measure studies. There have been no test–retest studies of Westen et al.'s (1989) ORSC measure or Cramer's (1991b) DMM; thus, there is no basis for evaluation of them on this ground. However, my colleagues and I (Hibbard, Mitchell, & Porcerelli, 2001) conducted internal consistency analyses on the ORSC scales across various samples. Unpublished findings (Hibbard, 2003) indicate that when student samples are combined (complete data for *N* = 280) using 10 TAT cards, values of α range from .70 to .85 for the four ORSC scales. We (Hibbard, 2003) also examined α levels for the three Cramer DMM scales, again unpublished. Among the same students (complete data for *N* = 301) using 6 cards, αs for the three Cramer DMM scales ranged from .60 to .63. Application of the Spearman–Brown prophecy formula to these data indicate that if 12 to 14 cards were used, minimal αs = .75 would be routinely observed. Of course, this assumes that similar interitems correlations would be observed. Moreover, it does not address the issue of how clinicians actually practice, that is, how many cards they typically use. Because the data suggest that clinicians ought to be using more cards if higher levels of α are desirable, then, as with best practice considerations for any test, so too with the TAT, perhaps they ought to be using more cards.

This is not to advocate (or deny) that TAT reliability should be evaluated by the use of coefficient α. It is rather to suggest that in the absence of test–retest data for TAT measures using adequate numbers of cards, estimates in the range of .75 are the best extant reliability data for what would be achieved if an adequate number of cards were used for the Westen et al. (1989) and Cramer (1991b) scales. Hence, although as Lilienfeld et al. (2000) rightly indicated, further reliability work (especially test–retest reliability) is indicated for these scales, the situation is nowhere near as grim as Lilienfeld et al. might have suggested. Aside from applications to compute estimates of true scores for individual evaluees, the primary reason to be concerned about reliability at all is the constraint it places on validity (Karon, 1968).

---

ture lists as covarying, there is nothing inherent in the classification system that requires features to be related. In fact, any effort to apply orthogonal rotations to factors derived from these feature lists is by its very nature an effort to derive uncorrelated aspects of these diagnostic feature lists. It has become popular in recent years to consider the *DSM* nosology as indexing "fuzzy" categories and hence as not specifying necessary and sufficient conditions for a diagnosis. However, in terms of formal logic, this is not accurate. Logically speaking, the *DSM* specifies sets of disjunctions (this and this and this or this and this and this, etc.), the satisfaction of at least one of which disjuncts is necessary and sufficient for the diagnosis.

According to classical test theory, the maximum value of a test's validity against a criterion is constrained by the reliability index. Specifically, maximum validity is defined as the square root of the reliability coefficient. This means that "even with reliability as low as .49, the upper limit of validity is .70" (Schmitt, 1996, p. 351). This has important implications for criticisms of TAT reliability. Specifically, when clinicians' judgments of clients or when other behavioral indicators are used as the criterion measures, both the Rorschach and our self-report measures account for about the same amount of variance: These coefficients typically account for about 10% of the variance (Meyer & Archer, 2001). At least this same degree of association obtains for the McClelland TAT scales (Spangler, 1992) when the criterion measures used are implicit rather than explicit measures. Similar or somewhat stronger levels of association have been observed for the Westen (Barends, Westen, Leigh, Silbert, & Byers, 1990; Hibbard, Hilsenroth, Hibbard, & Nash, 1995; Leigh, Westen, Barends, Mendel, & Byers, 1992; Westen, Lifton, Boekamp, Huebner, & Silverman, 1991) and Cramer (Cramer, 1995, 1999b; Cramer, Blatt, & Ford, 1988, Tables 2 and 3; Hibbard, Porcerelli, et al., 2003) scales when indexes of association for relevant statistical tests have been reported, although admittedly no meta-analyses have yet been done for the Westen and Cramer scales.

In no case, however, do these validity indexes consistently approach even the range of .50 A reliability as low as .70 imposes an upper limit on validity of .84, but a reliability of .80 imposes an upper limit of .89. Given that the validities of all of our measures are far below this, both the projective and objective measures have plenty of "excess capacity" in terms of the level of constraint imposed on validity by reliability. In this regard, it is difficult to argue that a reliability of .80 is adequate, but one of .70 or even .65 is not. To be sure, however, inadequate numbers of TAT cards may have often been used previously in many research situations. Likewise, studies actually using 12 to 14 cards in a bona fide test–retest context need be undertaken.

DMM reliability. Lilienfeld et al. (2000) characterized as "troublesome" a report (which they attribute to Cramer, 1991b) of DMM α coefficients of .83, .63, and .57 for the Identification, Projection, and Denial scales, respectively, for 40 participants using 8 cards. The α coefficients that Cramer (1991b, Table 13.5) actually reported, however, are .76, .67, and .63, respectively. What Lilienfeld et al. described as αs are actually split-half reliabilities from the same study (Cramer, 1991X, Table 13.4). Application of Spearman–Brown here suggests that Cramer would realize αs greater than .75 on all scales had she used 14 cards. Probably fewer cards would be needed to achieve that level with a larger sample. The fact is, Cramer's reported αs are reasonable for research studies using 8 cards.

Lilienfeld et al. (2000) then proceeded to characterize as "alternate form test–retest reliabilities" (p. 45) a report of correlations between two tests that were administered preexperimental and postexperimental manipulations to 32 second graders and to 32 sixth graders. Because experimental manipulations intervened, however, it is inaccurate to describe this as alternate forms of the same test (Anastasi & Urbina, 1997). This echoes Lilienfeld et al.'s earlier mistaken representation of CS reports of trait stability as test–retest reliabilities (see preceding text). The mistake in this latter case is slightly more understandable, however, because Cramer herself similarly characterized these findings. To the credit of Lilienfeld et al., they did mention that significant interventions took place before retest to each of the participants. They did not, however, report the relevant fact that in the first administration, only two cards were used and in the second only three, which means that the tests were not parallel, a necessary condition of alternate forms. Moreover, the use of so few cards is quite likely to further attenuate correlations. Lilienfeld et al. also failed to report the most relevant information about the nature of the interventions as reported in the original study (Cramer & Gaul, 1988): They were two different manipulations (success and failure) predicted and found to produce opposite effects on defense use! Under such circumstances, it is a wonder that measurements of defense at premanipulation and postmanipulation would correlate at all. Of course, no one would actually design an alternate form test–retest study in this way, and it was not so designed by Cramer. It would seem, then, that there are no bona fide alternate form or test–retest studies of the Cramer (1991b) or the Westen et al. (1989) scales as conventionally described in psychometrics texts.

The Validity of Three TAT-Based Coding Systems

Intelligence, the development of cognitive complexity, and verbal productivity. Lilienfeld et al. (2000) pointed out that in some studies (although not in others) some TAT constructs have been noted to correlate with intelligence as measured by (typically verbal) intelligence tests or with number of words used. On this basis, Lilienfeld et al. (2000) recommended that measures of intelligence and word counts always be provided "as covariates … so that the potential confounding role of verbal ability can be clarified" (p. 44).

Contrary to what Lilienfeld et al. (2000) suggested, a correlation of a TAT variable of interest with a measure of intelligence does not mean that the two are confounded in predicting some other variable. The theories underlying both Westen et al.'s (1989) ORSC and Cramer's (1991b) DMM depict object representations and defenses as, like intelligence, developmentally sensitive constructs. It is perhaps best to construe each of them as facets of another developmental and adaptive construct—that of ego function—and as such, each should have some relation to each other (Allen, Coyne, & David, 1986; Browning & Quinlan, 1985; Cramer,

1999a). The best available empirical data confirms this developmental connection (Cramer, 1987, 1997; Porcerelli, Thomas, Hibbard, & Cogan, 1998; Westen, Klepser, et al., 1991). Hence, object representations, increased use of more mature defenses as well as intelligence should increase developmentally as aspects of cognitive development and have been empirically demonstrated to do so. Therefore, it is consistent with the construct validity of both object representations and defenses that they would be correlated with intelligence. It is not a confound, but part of the construct validity of these measures.

Indeed, it might be problematic if in every research study relating ORSC scales to some construct of interest or comparing groups on ORSC scales, significant findings disappeared when measured intelligence scores were covaried out. Lilienfeld et al. (2000) cited one study in the literature (Westen, Ludolph, Block, Wixom, & Wiss, 1990) for which that might be true, although the data are not clear. As I discuss following, it is important to distinguish bona fide criterion validity studies, those attempting to establish a measure's validity, from studies that use a measure to explore other constructs. Westen et al. is a study of the latter type. However, in a criterion validity study correlating the ORSC with an appropriate Rorschach measure of object representations, Hibbard et al. (1995) found that although both TAT and Rorschach object representations measures correlated with measured IQ, the relation between the object representations measures stayed significant even when IQ was covaried out. Contrary to what Lilienfeld et al. said, there need not always be concern about controlling for IQ in doing ORSC studies. Cognitive facets of object representations do and should correlate with other indexes of cognitive complexity (such as IQ) because each of them is a facet of the broad constructs of ego function and ego development. However, as Hibbard et al. showed, the constructs are not identical and IQ does not mediate all the variance common to ORSC scales and relevant criteria.

Sheer verbal productivity is a different matter, and adequate studies have not yet been developed in this regard. It is not clear whether greater verbal productivity is a necessary concomitant of the expression of higher levels of complexity in the expression of more sophisticated object relations and the defense of identification or whether higher scores on these scales is an accident of greater productivity. Only one study (Leigh et al., 1992) thus far has addressed this issue. Leigh et al. examined the relation of word count to ORSC variables and found that, in that data set at least, there was little relation. Yet here again, Lilienfeld et al. (2000) did not consider the study.

In regard to Cramer's (1991b) DMM, examination of my and Porcerelli's data sets (Hibbard & Porcerelli, 1998; Hibbard, Tang, et al., 2000; Hibbard, Porcerelli, et al., 2001) indicates that there are moderate relations between productivity and raw scores on all three DMM defenses, but they are quite different for the three defenses. However, if the TAT variable of interest is not the raw score per se but what Cramer called *relative scores* (e.g., identification over the sum of all defenses), this attenuates the relation with word count. The use of percent scores is the one most central to the underlying theoretical model of defense in relation to development. Hence, Lilienfeld et al.'s (2000) concern about TAT story length is in most cases no cause for alarm.

The ORSC. The Lilienfeld et al. (2000) review of the validity of the Westen et al. (1989) system suffers from a different problem. Lilienfeld et al. systematically failed to distinguish between the validation of the ORSC and its use in discovering (or failing to do so) linkage between object relations phenomena and other aspects of psychopathology. Lilienfeld et al. also omitted from their report most of the important validity studies.

In the validation of a measure, the initial concern should be to investigate the extent to which the measure correlates with other measures of the same purported construct (criterion validity), the extent to which the measure shows the expected pattern of results in groups of people known to have different levels on the constructs (known groups validity), or the extent to which a relevant manipulation produces predicted changes in the measure. This is a fundamentally different endeavor than the use of measures that have been already reasonably well validated to extend one's knowledge of whether and how the construct may be manifested in domains more remote from the central construct itself. Lilienfeld et al. (2000), however, presented studies that are primarily involved with the testing of new hypotheses as if they were validity studies. Lilienfeld et al. then proceeded to indict occasional failures (to reject the null) in these studies as if they somehow weakened the validity of these ORSC variables.

For example, Lilienfeld et al. (2000) pointed to a finding in a single study (Westen et al., 1990) that Complexity of Representations (CR) was higher in a group of teens with borderline personality disorders than in a group of teens with mixed psychiatric diagnoses as if it disconfirmed the validity of the measure, as if somehow it showed that it does not measure what it purports to. This finding (Lilienfeld et al, 2000, p. 43) merely indicates that teens with borderline personality disorder (or a small sample of them) have more complex representations than some (perhaps including Westen) might have thought. It does not mean that the CR scale is not a valid measure of the complexity of object relations. Lilienfeld et al. similarly criticized the validity of the ORSC because it failed to correlate with the number of early moves (changes in residence) of participants in a study of abuse victims! Lilienfeld et al.'s conclusion confuses aspects of scale validity with what most would regard as the use of a measure to test hypotheses regarding the nature or etiology of forms of psychopathology.

Although Lilienfeld et al. (2000) cited these peripheral studies to criticize the ORSC, they overlooked the basic crite-

rion and known groups validity studies done on the ORSC. These include Westen, Klepser, et al. (1991; developmental changes in the ORSC); Leigh, Westen, et al. (1992; correlations with other object representations measures); Westen, Lifton, et al. (1991; known groups validity); Hibbard et al. (1995; correlations with object relations measures); and Barends et al. (1990; correlations with other ego function and interview data). A complete consideration of the validity evidence would include an account of these basic validity studies.

Cramer's DMM. Lilienfeld et al. (2000) gave even lower marks to the validity of the Cramer (1991b) DMM. Lilienfeld et al. did acknowledge at least one study in which Cramer (1987) demonstrated predicted changes in the development of defenses across different ages in childhood. Lilienfeld et al. also cited Porcerelli et al. (1998), which largely replicates Cramer's (1987) findings (although Lilienfeld et al. erroneously attributed Porcerelli et al.'s replication to Hibbard et al., 1994). Furthermore, Lilienfeld et al. ignored two other studies that confirm this developmental difference: Cramer (1997) and Cramer and Gaul (1988). Moreover, Lilienfeld et al. erroneously challenged the interpretation of the data. In Porcerelli et al., three defenses were coded for five grade- or age-level groups, spanning from second grade (about 7 years old) to college freshman (about 18 years old). Cramer's (1991b) theory is that these defenses become prominent and then fade at different stages of development. The highest use of Denial is in early childhood and declines prior to the second grade, Projection then rising but declining by late adolescence. Projection is then superceded by Identification. Porcerelli et al.'s findings confirmed this, excepting only that the youngest group was too old to test the hypothesis concerning Denial. Lilienfeld et al. (2000, p. 34) claimed that Porcerelli et al. failed to confirm Cramer (1991b), which Lilienfeld et al. said predicts an increase in relative Projection scores from ages 7 to 10. However, when one consults Cramer (1991b, p. 34), there is no such specific prediction that Projection ought to rise from the ages of 7 to 10; this prediction is not a constituent of her theory or model, and the Porcerelli et al. empirical findings map nicely onto Cramer's (1987) results, although the data were collected a decade apart. It is unclear how Lilienfeld et al. came to this incorrect interpretation of Cramer's (1991b) views.

Lilienfeld et al. (2000) also were mistaken about statistical findings in Porcerelli et al. (1998). Porcerelli et al. noted their 11th graders' Denial percent scores increased slightly but not significantly (Tukey's honestly significant difference) over that for the 8th graders. Providing no statistical analysis and apparently overlooking printed findings in Porcerelli et al. Table 2, Lilienfeld et al. (2000) stated that this difference was significant. Just to make sure, I went back and hand computed the *t* for independent samples (Hays, 1988, Formula 8.9.1) for the *n, M,* and *SD* values printed in Porcerelli et al. Table 2, and it was not significant, $t(58) = .85$, *ns.* Moreover, even if it had been significant, this would not have diminished Porcerelli et al.'s confirmation of the Cramer (1991b) model because that model is indifferent to minor fluctuations in Denial percent in later ages. Again, I have not been able to identify any source of Lilienfeld et al.'s mistaken claim regarding this statistic.

Lilienfeld et al. (2000) criticized Hibbard et al. (1994) for using relative (percentage) defense scores, claiming that they are not used in Cramer's studies. Contrary to this claim, however, Cramer (1987; Cramer & Gaul, 1988) indeed has used percentage scores precisely because they are the most appropriate form of score to use in these developmental studies. Indeed, Cramer's (1991b) model that Lilienfeld et al. cited clearly labels the *y* axis as "relative frequency of occurrence" (p .34). Hence, not only were Lilienfeld et al. wrong about whether Cramer used percentage scores, they also seemed to drastically misunderstand the developmental aspect of Cramer's (1991b) theory: The theory itself is about proportions of defense, not raw scores. Porcerelli et al. (1998), of course, also used percentage scores because the entire analysis there depends on the use of such relative scores. Lilienfeld et al. seemed to not fully understand the basis of this well-replicated developmental theory.

Lilienfeld et al. (2000, p. 45) also apparently failed to understand the experimental design of Cramer and Gaul (1988). Lilienfeld et al. (p. 45) stated that participants were randomly assigned to treatments when in fact participants in this study were first matched on total defense to create two groups. Cramer and Gaul clearly reported that they did not randomly assign participants but first administered two TAT cards to groups of children to create the matched groups. The two groups were then given different manipulations and then retested with the TAT. It is no wonder then that Lilienfeld et al. were subsequently confused about why the children were given different numbers of TAT cards on different occasions. Lilienfeld et al. apparently misunderstood or at any rate misreported the experimental design. (This is the same study for which it was seen that Lilienfeld et al. inappropriately reported alternate forms reliability.) Lilienfeld et al. also misrepresented the results of the study. Lilienfeld et al. said that the use of projection decreased after failure, basing this on a .04 change in raw scores. Lilienfeld et al. ignored Cramer's (Cramer & Gaul, 1988) own report of scores adjusted for story length, according to which the failure intervention produced the predicted effect. Lilienfeld et al. rendered this (mis)report, even after admonishing others for not using length-adjusted scores.[4] Lilienfeld et al. also misrepresented Cramer and Block (1998) insofar as they claimed that Cramer and Block "made few explicit predictions"

[4]Increased verbal productivity is expected with age increases. The theory is that changes in use of preferred defenses are independent of normal increases in the length of verbal responses. On the other hand, it seems likely that increases in verbal facility and also increases in the use of identification are both aspects of cognitive complexity.

(Lilienfeld, 2000, p. 45). There may be no strong consensus on the meaning of *few,* but Cramer and Block did predict their major finding, that the amount of immature defenses obtained from males in their early 20s would correlate with observers' ratings of behavior dysfunction recorded when the boys were preschoolers. This was the appropriate prediction to make, given that the prevalence of immature defenses should lead to greater dysfunction later.

Lilienfeld et al. (2000) failed to cite relevant vaidity studies forthe DMM. These include Cramer's (1991a) study of increased defenses after the manipulation of anger in college students, Hibbard et al.'s (1994) support for the DMM's construct validity through factor analysis, Cramer's (1997) cross-lagged study showing relevant within-subjects changes, Cramer's (1999b) study relating defenses in theoretically predicted ways to Loevinger's (Hy & Loevinger, 1996) ego development scales, and Cramer's (1999a) study relating defenses to personality pathology in predicted ways. A complete account of the validity of the DMM would consider all the relevant validity studies.

DISCUSSION AND CONCLUSIONS

Herein, I have tried to document a large number of errors of omission and commission in the Lilienfeld et al. (2000) review of the CS and of three TAT scoring systems. I have tried to show how and when these errors have resulted in a more negative appraisal of the validity of projectives than is warranted. Specifically, they have resulted in a misdepiction of the views of Rorschach and TAT researchers (the projective rationale); in inappropriate statistical analysis (dichotomous vs. continuous decisions); in unjustified suggestions of cultural bias in the Rorschach; in large numbers of validity studies going either misreported or unreported (Westen and Cramer scales); in passing over longstanding and contemporary arguments (or simple extensions thereof) relevant to the scientific consideration of reliability; in instances of test validation being confused with scientific discovery (Westen); in cases of statistical tests (Porcerelli et al., 1998), experimental designs, and theories being inaccurately reported (Cramer); in the misconstrual of trait stability as test–retest or alternate form reliability; and in alternate forms being cited in the absence of evidence for parallel tests (Cramer).

I have tried to document the extent to which these errors have accumulated within Lilienfeld et al.'s (2000) largely negative account of projectives. It is important to recall that, on the basis of their report, Lilienfeld et al. made recommendations for strong constraints on the use of projectives, for constraints and modifications in clinical training in assessment, and on how research on projectives ought to proceed. If as is argued herein, there is a large aggregation of errors in their report, then the reasonableness of their recommendations may require some rethinking. I contend that it would not be reasonable to implement their recommendations in the light of the errors in their criticisms. This said, it needs also be acknowledged that Lilienfeld et al. made some reasonable criticisms of the Rorschach and TAT and that more research needs to be done. The validity work on these measures is an ongoing process. Constructive scientific inquiry in this field will advance the interests of research and, it is likely, advance the interests of clinical service as well.

REFERENCES

Acklin, M. W., McDowell, C. J., II, 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.

Allen, J. G., Coyne, L., & David, E. (1986). Relation of intelligence to ego functioning in an adult psychiatric population. *Journal of Personality Assessment, 50,* 212–221.

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

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

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

Browning, D. L., & Quinlan, D. M. (1985). Ego development and intelligence in a psychiatric population: Wechsler subtest scores. *Journal of Personality Assessment, 49,* 260–263.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in psychological assessment research. *Psychological Assessment, 7,* 309–319.

Cramer, P. (1987). The development of defense mechanisms. *Journal of Personality, 55,* 597–614.

Cramer, P. (1991a). Anger and the use of defense mechanisms in college students. *Journal of Personality, 59,* 39–55.

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

Cramer, P. (1995). Identity, narcissism, and defense mechanisms in late adolescence. *Journal of Research in Personality, 29,* 341–361.

Cramer, P. (1997). Evidence of change in children's use of defense mechanisms. *Journal of Personality, 65,* 233–247.

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

Cramer, P. (1999b). Personality, personality disorders, and defense mechanisms. *Journal of Personality, 67,* 535–554.

Cramer, P., Blatt, S. J., & Ford, R. Q. (1988). Defense mechanisms in the anaclitic and introjective personality configuration. *Journal of Consulting and Clinical Psychology, 56,* 610–616.

Cramer, P., & Block, J. (1998). Pre-school antecedents of defense mechanism use in young adults. *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.

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

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

Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (3rd ed.). New York: Wiley.

Exner, J. E., Jr. (Ed.). (1995). *Issues and methods in Rorschach research.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

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

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

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory.* Minneapolis: University of Minnesota Press.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Hibbard, S., Farmer, L., Wells, C., Difillipo, E., Barry, W., Korman, R., et al. (1994). Validation of Cramer's defense mechanism manual for the TAT. *Journal of Personality Assessment, 63,* 197–210.

Hibbard, S., Hilsenroth, M. J., Hibbard, J. K., & Nash, M. R. (1995). A validity study of two projective object representations measures. *Psychological Assessment, 7,* 432–439.

Hibbard, S., Mitchell, D., & Porcerelli, J. (2001). Internal consistency of the Object Relations and Social Cognition scales for the Thematic Apperception Test. *Journal of Personality Assessment, 77,* 408–419.

Hibbard, S., & Porcerelli, J. (1998). Further validation for Cramer's defense mechanism manual. *Journal of Personality Assessment, 70,* 460–483.

Hibbard, S., Porcerelli, J., Kamoo, R., Schwartz, M., Abell, S., Latko, R., et al. (2003). *Cramer's defense mechanism manual and level of personality organization.* Manuscript submitted for publication.

Hibbard, S., Tang, P. C. Y., Latko, R., Park, J. H., Munn, S., Bolz, S., et al. (2000). Differential validity of the Defense Mechanism Manual for the TAT between Asian Americans and whites. *Journal of Personality Assessment, 75,* 351–372.

Hy, L. X., & Loevinger, J. (1996). *Measuring ego development* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Johnson, D. L., & Sikes, M. P. (1965). Rorschach and TAT responses of Negro, Mexican-American, and Anglo psychiatric patients. *Journal of Projective Techniques and Personality Assessment, 29,* 183–188.

Jones, E. E., & Thorne, A. (1987). Rediscovery of the subject: Intercultural approaches to clinical assessment. *Journal of Consulting and Clinical Psychology, 55,* 488–495.

Karon, B. P. (1968). Problems of validities. In A. I. Rabin (Ed.), *Projective techniques in personality assessment* (pp. 85–114). New York: Springer.

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

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

Lasch, K. (1979). *The culture of narcissism.* New York: Norton.

Leigh, J., Westen, D., Barends, A., Mendel, M. J., & Byers, S. (1992). The assessment of complexity of representations of people using TAT and interview data. *Journal of Personality, 60,* 809–837.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Lundy, A. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment, 49,* 141–145.

McNulty, J. L., Graham, J. R., Ben-Porath, Y. S., & Stein, L. A. R. (1997). Comparative validity of MMPI–2 scores of African American and Caucasian mental health center clients. *Psychological Assessment, 9,* 464–470.

Meyer, G. J. (1992). 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. (1997). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment, 9,* 480–489.

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13,* 486–502.

Meyer, G. J., & Richardson, C. (2001, March). An examination of changes in form quality codes in the Rorschach Comprehensive System from 1974 to 1995. Paper presented at the midwinter meeting of the Society for Personality Assessment, Philadelphia, PA.

Millon, T. (1987). *Millon Clinical Multiaxial Inventory–II manual.* Minneapolis, MN: National Computer Systems.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Murstein, B. I. (1963). *Theory and research in projective techniques.* New York: Wiley.

O'Conner, B. P. (2002). A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment, 9,* 188–203.

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.

Peterson, R. A. (1994). A meta-analysis of Cronbach's co-efficient alpha. *Journal of Consumer Research, 21,* 381–391.

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.

Reuman, D. A. (1982). Ipsative behavioral variability and the quality of thematic apperceptive measurement of the achievement motive. *Journal of Personality and Social Psychology, 43,* 1098–1110.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126,* 3–25.

Schmitt, N. (1996). The uses and abuses of co-efficient alpha. *Psychological Assessment, 8,* 350–353.

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

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.

Smith, C. P. (1992). *Handbook of thematic content analysis.* New York: Cambridge University Press.

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

Timbrook, R. E., & Graham, J. R. (1994). Ethnic differences on the MMPI–2? *Psychological Assessment, 6,* 212–217.

United States Census Bureau. (2001). *Statistical abstract of the United States: 2001. The national data book 121st ed.* Washington, DC: U.S. Government Printing Office.

Wallach, M. A., & Wallach, L. (1983). *Psychology's sanction for selfishness: The error of egoism in theory and therapy.* San Francisco: Freeman.

Westen, D. (1991). Social cognition and object relations. *Psychological Bulletin, 109,* 429–455.

Westen, D., Klepser, J., Ruffins, S. A., Silverman, M., Lifton, N., & Boekamp, J. (1991). Object relations in childhood and adolescence: The

development of working representations. *Journal of Consulting and Clinical Psychology, 59,* 400–409.

Westen, D., Lifton, N., Boekamp, J., Huebner, D., & Silverman, M. (1991). Assessing complexity of representations of people and understanding of social causality: A comparison of natural science and clinical psychology graduate students. *Journal of Social and Clinical Psychology, 10,* 448–458.

Westen, D., Ludolph, P., Bock, M. J., Wixom, J., & Wiss, F. C. (1990). Developmental history and object relations in psychiatrically disturbed adolescent girls. *Journal of the Academy of Child and Adolescent Psychiatry, 29,* 338–348.

Westen, D., Silk, K., Lohr, N., & Kerber, K. (1989). *Object relations and social cognition TAT scoring manual* (Rev. ed.). Unpublished manuscript, University of Michigan, Ann Arbor.

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. (1977). Power motive reliability as a function of retest instructions. *Journal of Consulting and Clinical Psychology, 45,* 436–440.

Wood, J. M., Lilienfeld, S. O., Garb, H. N., & Nezworski, M. T. (2000). 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., Nezworski, M. T., & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7,* 3–17.

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.

Stephen R. Hibbard
Department of Psychology
285 South Chrysler Hall
University of Windsor
Windsor, Ontario, N9B 3P4
Canada
E-mail: hibbard@uwindsor.ca

Received January 18, 2002
Revised October 25, 2002