In M. Hilsenroth & D. Segal (Eds.), Personality assessment. Volume 2 in M. Hersen (Ed.-in-Chief), *Comprehensive handbook of psychological assessment* (pp. 315-342). Hoboken, NJ: John Wiley & Sons.

CHAPTER 25

The Reliability and Validity of the Rorschach and Thematic Apperception Test (TAT) Compared to Other Psychological and Medical Procedures: An Analysis of Systematically Gathered Evidence

GREGORY J. MEYER

Over the last 6 years, a series of strident criticisms have been directed at the Rorschach (e.g., Garb, 1999; Grove & Barden, 1999; Wood, Lilienfeld, Garb, & Nezworski, 2000; Wood, Nezworski, Garb, & Lilienfeld, 2001; Wood, Nezworski, & Stejskal, 1996) and countered by conceptual and empirical rebuttals (e.g., Acklin, 1999; Exner, 1996, 2001; Garfield, 2000; Hilsenroth, Fowler, Padawer, & Handler, 1997; Meyer, 2001a, 2002; Meyer et al., 2002; Ritzler, Erard, & Pettigrew, 2002; Weiner, 2000). Perhaps most significantly, two special sections in the journal *Psychological Assessment* brought together representatives from these opposing perspectives to debate the Rorschach's evidentiary foundation in a structured and sequential format that allowed for a full consideration of its strengths and limitations (Meyer, 1999, 2001b).

In a legal context, Grove and Barden (1999) argued that because the Rorschach has the most extensive body of research supporting its psychometric properties, any critical deficiencies that could be identified for it should also generalize to the other less studied performance tasks used to assess personality, such as the Thematic Apperception Test (TAT), sentence completion tests (SCTs), or figure drawings. And indeed, substantial criticisms have recently been leveled against all of these procedures (Lilienfeld, Wood, & Garb, 2000). Importantly, these criticisms extended beyond the professional journals to capture public attention. Lilienfeld et al.'s disparaging review of these instruments was discussed in a prominent *New York Times* story (Goode, 2001) and a shorter, repackaged version of the review appeared in the widely circulated popular magazine *Scientific American* (Lilienfeld, Wood, & Garb, 2001).

With this background setting the stage, my central purpose in this chapter is to provide a broad overview of the psychometric evidence for performance-based personality tests, with a primary focus on the Rorschach and TAT. However, any review that focuses exclusively on the reliability and validity evidence for these instruments (e.g., Lilienfeld et al., 2000) encounters a serious limitation. Once the evidence is in hand, one must determine whether it provides reasonable support for a test, evidence of salient deficiencies, or some combi-

Acknowledgments: A grant from the Society for Personality Assessment supported the preparation of this chapter and the new research it reports, including the complete recoding and reexamination of Parker, Hanson, and Hunsley's (1988) meta-analyses on the convergent validity of the WAIS, MMPI, and Rorschach initially reported in Meyer and Archer (2001).

nation, such that the test may be considered reliable and valid for some purposes but not others. Stated differently, the psychometric evidence must be processed and linked to other knowledge in order to be interpreted in a sensible fashion. However, processing and interpreting a complex body of information provides many opportunities for the cognitive biases that afflict all humans to exert their influence (Arkes, 1981; Borum, Otto, & Golding, 1993; Garb, 1998; Hammond, 1996; Spengler, Strohmer, Dixon, & Shivy, 1995).

For instance, does an intraclass correlation of .70 indicate that the interrater reliability for test scoring is good or poor? Does a 5-year test-retest correlation of .50 provide favorable or unfavorable evidence regarding the traitlike stability of a test scale? Does a correlation of .30 between a scale and a relevant criterion provide reasonable evidence for or against the construct validity of a test? When answering each of these questions, the evidence is open to interpretation. And preexisting beliefs, schemas, and affective reactions will shape how one makes sense of the data. Thus, if the coefficients listed previously had been found for the Rorschach, psychologists with a generally unfavorable stance toward this test may be prone to view all of the findings as deficient and as indications the Rorschach should not be used in applied clinical practice. In contrast, psychologists who have a generally favorable preexisting stance toward the test may be prone to view all three examples as indicating reasonable, positive evidence that supports Rorschach reliability and validity. Both sides could find legitimate reasons to argue their positions and they likely would never achieve a resolution, particularly if one or both sides did not consider the magnitude of the reliability and validity coefficients that are typically obtained for other kinds of instruments.

A number of authors have suggested strategies to help minimize or rectify the cognitive biases that affect judgments made under complex or ambiguous circumstances (Arkes, 1981; Borum et al., 1993; Spengler et al., 1995). Invariably, these strategies include some method for listing and systematically attending to information, particularly information that may disconfirm an initial impression or a faulty memory. In the present context, this general de-biasing strategy will be implemented in two ways.

First, data will be drawn from systematically gathered meta-analytic evidence. By statistically summarizing a body of literature, meta-analyses provide a more accurate understanding of existing knowledge than the study-by-study examination of findings found in the traditional narrative review (e.g., Lilienfeld et al., 2000). The value of meta-analysis as a scientific de-biasing procedure was pointedly stated by Chalmers and Lau (1994) when contrasting meta-analysis with conventional reviews in medicine:

> Too often, authors of traditional review articles decide what they would like to establish as the truth either before starting the review process or after reading a few persuasive articles. Then they proceed to defend their conclusions by citing all the evidence they can find. The opportunity for a biased presentation is enormous, and its readers are vulnerable because they have no opportunity to examine the possibilities of biases in the review. (cited in Hunt, 1997, p. 7)

Indeed, meta-analyses have regularly clarified the scientific foundation behind controversial areas of research and have often overturned the accepted wisdom derived from narrative reviews (see Hunt, 1997, for compelling examples).

Second, the psychometric data presented in this chapter will not be limited to tests like the Rorschach and TAT, which evoke strong reactions from many psychologists. Rather, reliability and validity evidence for these instruments will be presented alongside evidence for alternative measures in psychology, psychiatry, and medicine. Doing so will allow readers to compare and weigh the evidence from a relatively broad survey of findings in applied health care. If tests like the Rorschach or TAT have notable psychometric deficiencies, these flaws should be quite obvious when relevant data are compared across assessment procedures.

Space limitations combined with a need to document the steps for systematically sampling the literature result in a chapter that is long on detail and relatively short on commentary. The overview begins with reliability and then examines validity. Reliability is considered in two categories, interrater agreement and retest stability. Interrater reliability indicates how well two people agree on the phenomenon they are observing. This form of reliability is critical to many human endeavors, including the clinical practice of psychology, psychiatry, and medicine. Because it is so important, interrater agreement has been studied frequently and a substantial literature exists from which to systematically cull data. Test-retest reliability documents the extent to which a measured construct has relatively constant, traitlike characteristics. This type of reliability is regularly investigated in psychology but has received less attention in medicine. Thus, the review presented here will be limited in scope.

META-ANALYSES OF INTERRATER RELIABILITY IN PSYCHOLOGY AND MEDICINE

Overview of Procedures

To obtain systematically organized data on interrater reliability, several strategies were used. These included (1) a broad search for existing data summaries, (2) narrow searches

for literature reviews that identified interrater studies and allowed meta-analytic results to be computed, and (3) new meta-analyses derived from a thorough search of relevant literature.

For a study to contribute data, reliability had to be reported as a correlation, intraclass correlation (ICC), or kappa coefficient (κ). ICCs and κ coefficients correct observed agreement for chance agreement and are often lower than Pearson correlations. Common interpretive guidelines for κ and the ICC are as follows: Values less than .40 indicate *poor* agreement, values between .40 and .59 indicate *fair* agreement, values between .60 and .74 indicate *good* agreement, and values greater than .74 indicate *excellent* agreement (Cicchetti, 1994).

Interrater reliability studies vary along several dimensions, including the number of objects rated *(n),* the number of judges *(k)* that examine and rate each object, and the number of qualities that are rated *(q).* Because of these differences, there are several weights that can be applied to samples when computing meta-analytic summary statistics. Meta-analyses traditionally ignore the number of qualities considered in a sample (e.g., in treatment outcome research, the results from multiple outcomes are simply averaged within a sample), so this variable was not considered further. The possible weights that remain include: $n(k)$, which indicates the total number of independent judgments in a sample; $n(k(k - 1)/2)$, which indicates the total number of paired rater judgments in a sample; and $n(k - 1)$, which indicates the maximum number of independently paired judgments in a sample.[1] Because $n(k)$ does not reduce to n in a traditional two-judge design and because $n(k(k - 1)/2)$ counts nonindependent judgments and increases quadratically as more judges are used, $n(k - 1)$ was used to weigh sample results whenever new computations were undertaken. This weight reduces to n in a two-judge design, gives appropriate added emphasis to multirater studies, provides a summary value that can be meaningfully interpreted, and ensures that excessively large weights are not assigned to designs containing multiple, nonindependent judgments. For the studies reported here, the different weights almost always produced similar results. I note the two instances when alternative weighing led to results that differed by more than .03 from the value obtained using $n(k - 1)$.

Some studies report reliability coefficients after they have been corrected according to the Spearman-Brown formula, which estimates the reliability for a multirater composite (e.g., for an average score computed across three raters). Because these adjusted coefficients are higher than traditional findings (i.e., the association between one rater and another), whenever Spearman-Brown estimates were reported, the findings were back-transformed to indicate reliability for a single rater using Table 3.10 in Rosenthal (1991).

Table 25.1 presents a summary of the interrater reliability research to be described more fully later. For each construct, the table indicates how many independent pairs of observations produced the findings, as well as the average reliability coefficient. Reliability coefficients were computed separately for scales and items. Items are single units of observation or single judgments, while scales are derived from the numerical aggregation of item-level judgments. Because aggregation allows random measurement errors associated with the lower-level items to cancel out, scales should be more reliable than items. Reliability coefficients were also computed separately for correlations and for chance-corrected coefficients (i.e., κ/ICC).

The findings in Table 25.1 are roughly organized by magnitude. This ordering attempts to take into account item-scale differences and differences between correlations and chance-corrected coefficients. However, because the ordering is imprecise and because some samples are small, minor differences in rankings should be ignored.

Broad Search for Existing Meta-Analyses or Systematic Reviews

To identify existing data summaries, I searched PsycINFO and PubMed in August 2001 for meta-analyses that had examined interrater reliability. Using the overly inclusive terms *reliability* or *agreement* combined with *meta-analy**, 304 nonduplicate, English-language articles were identified. Subsequently, the search on PubMed was expanded by combining the medical subject heading (MeSH) categories *meta-analysis* or *review literature* or the text phrase *systematic-review* with the MeSH terms *reproducibility of results* or *observer variation.* After excluding articles without abstracts, the new search identified 214 studies, 189 of which did not overlap with the prior search. After reviewing these 493 abstracts and deleting those that clearly were not designed to summarize data on interrater reliability, 63 studies remained. To these I added two relevant reviews that did not appear in the searches (Etchells, Bell, & Robb, 1997; Gosling, 2001). All 65 articles were obtained and 25 contained either summary data on interrater reliability or a complete list of studies that provided relevant results. For the latter, all of the original citations were obtained and their findings were aggregated. A brief description of each review follows, starting with those addressing psychological topics.

Two nonoverlapping meta-analyses (Achenbach, McConaughy, & Howell, 1987; Duhig, Renk, Epstein, & Phares, 2000) examined the reliability of ratings made on child or

TABLE 25.1 Meta-Analyses of Interrater Reliability in the Psychological and Medical Literature

Target reliability construct	$n(k-1)$ = Number of independent pairs of judgments	r		κ/ICC	
		Scale	Item	Scale	Item
1. Bladder Volume by Real-Time Ultrasound	40		.95		
	320				.92
2. Count of Decayed, Filled, or Missing Teeth (or Surfaces) in Early Childhood	113	.97			
	237				.79
3. Rorschach Oral Dependence Scale Scoring	934	.91			
	40			.91	
	6,430				.84
4. Measured Size of the Spinal Canal and Spinal Cord on MRI, CT, or X-Ray	200	.90			
	86		.88		
5. Scoring the Rorschach Comprehensive System[a]:					
Summary scores	219	.90			
	565			.91	
Response segments	11,518				.86
Scores for each response	11,572				.83
6. Neuropsychologists' Test-Based Judgments of Cognitive Impairment	901				.80
7. Hamilton Depression Rating Scale Scoring From Joint Interviews[b]	1,773	.93			
	334		.68		
	2,074			.80	
	161				.77
8. Hamilton Anxiety Rating Scale Scoring From Joint Interviews[b]	512	.91			
	240			.87	
	214				.72
9. Level of Drug Sedation by ICU Physicians or Nurses	327	.91			
	789			.84	
	165				.71
10. Functional Independence Measure	1,365			.91	
	1,345				.62
11. TAT Personal Problem-Solving Scale Scoring	282	.86			
	103			.83	
12. Borderline Personality Disorder:					
Diagnosis	402			.82	
Specific symptoms	198				.64
13. Signs and Symptoms of Temporomandibular Disorder	192			.86	
	562				.56
14. Hamilton Anxiety Rating Scale Scoring From Separate Interviews	60	.65			
	208			.79	
	208				.58
15. Axis I Psychiatric Diagnosis by Semi-Structured SCID in Joint Interviews	216			.75	
16. Axis II Psychiatric Diagnosis by Semi-Structured Joint Interviews	740			.73	
17. Hamilton Depression Rating Scale Scoring From Separate Interviews	649	.88			
	367		.55		
	363			.72	
	230				.46
18. TAT Social Cognition and Object Relations Scale Scoring	653	.86			
	281			.72	
19. Rorschach Prognostic Rating Scale Scoring	472	.84			
20. TAT Defense Mechanism Manual Scoring	713	.80			
	30			.79	
21. Therapist or Observer Ratings of Therapeutic Alliance in Treatment	(S = 31)	.78			
22. Type A Behavior Pattern by Structured Interview	(S = 3)	.74			
23. Job Selection Interview Ratings	12,549	.70			
24. Personality or Temperament of Mammals	151	.71			
	637		.49		
25. Editors' Ratings of the Quality of Manuscript Reviews or Reviewers	113		.66		
	3,608				.54[c]
26. Visual Analysis of Plotted Behavior Change in Single-Case Research	410		.61		
	867				.55
27. Presence of Clubbing in Fingers or Toes	630				.52[d]
28. Stroke Classification by Neurologists	1,362				.51
29. Axis I Psychiatric Diagnosis by Semi-Structured SCID in Separate Interviews	693			.56	

TABLE 25.1 *(Continued)*

Target reliability construct	$n(k-1)$ = Number of independent pairs of judgments	r		κ/ICC	
		Scale	Item	Scale	Item
30. Axis II Psychiatric Diagnosis by Semi-Structured Separate Interviews	358			.52	
31. Child or Adolescent Problems:					
Teacher ratings	2,100	.64			
Parent ratings	4,666	.59			
Externalizing problems	7,710	.60			
Internalizing problems	5,178	.54			
Direct observers	231	.57			
Clinicians	729	.54			
32. Job Performance Ratings by Supervisors	1,603	.57			
	10,119		.48		
33. Self and Partner Ratings of Conflict:					
Men's aggression	616	.55			
Women's aggression	616	.51			
34. Determination of Systolic Heart Murmur by Cardiologists	500				.45
35. Abnormalities on Clinical Breast Examination by Surgeons or Nurses	1,720				.42
36. Mean Quality Scores From Two Grant Panels:					
Dimensional ratings	1,290		.41		
	1,177				.46
Dichotomous decision	398				.39
37. Job Performance Ratings by Peers	1,215	.43			
	6,049		.37		
38. Number of Dimensions to Extract From a Correlation Matrix by Scree Plots	2,300				.35
39. Medical Quality of Care as Determined by Physician Peers	9,841				.31
40. Definitions of Invasive Fungal Infection in the Clinical Research Literature	21,653				.25
41. Job Performance Ratings by Subordinates	533	.29			
	4,500		.31		
42. Research Quality by Peer Reviewers:					
Dimensional ratings	6,129		.30		
	24,939				.24
Dichotomous decisions	4,807				.21

Note. See text for complete description of sources contributing data to this table. CT = computed tomography, Dx = diagnosis, ICC = intraclass correlation, ICU = intensive care unit, κ = kappa, MRI = magnetic resonance imaging, r = correlation, S = number of studies contributing data, SCID = Structured Clinical Interview for the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*), and TAT = Thematic Apperception Test.

[a]Results reported here differ slightly from those in Meyer et al. (2002) because Meyer et al. used $n(k(k-1)/2)$ to weigh samples. For response segments, the number of independent pairs of judgments is the average across segments rather than the maximum per segment (17,540) or the total across all samples (18,040).

[b]This category included ratings of videotaped interviews and instances when the patient's report fully determined both sets of ratings (e.g., the patient completed the scale in writing and then was asked the same questions in a highly structured oral format; the patient was given a highly structured personal interview and then immediately thereafter was asked the same highly structured questions via the telephone).

[c]For agreement among editors on the quality of peer reviewers, results varied noticeably depending on the weighting scheme selected and thus should be considered tentative. Reliability was .58 when weighing by the total number of independent ratings and it was .48 when weighing by the total number of objects judged by all possible combinations of rater pairs.

[d]One study produced outlier results (κ = .90) relative to the others (κ range from .36 to .45). When weighing by the total number of independent ratings κ remained at .52, but when weighing by the total number of objects judged by all possible rater pair combinations κ was .46.

adolescent behavioral and emotional problems (Table 25.1, Entry 31). Duhig et al. only examined interparent agreement and their findings were combined with those reported by Achenbach et al. Using the Conflict Tactics Scales, Archer (1999; Table 25.1, Entry 33) examined the extent to which relationship partners agreed on each other's aggressiveness. Conway, Jako, and Goodman (1995; Table 25.1, Entry 23) examined agreement among interviewers conducting job selection interviews.

The search identified three meta-analyses on the interrater reliability of job performance ratings (Conway & Huffcutt, 1997; Salgado & Moscoso, 1996; Viswesvaran, Ones, & Schmidt, 1996), as well as one large-scale study that combined data on this topic from 79 settings (Rothstein, 1990). Because Conway and Huffcutt provided the most differentiated analysis of rater types (i.e., supervisors, peers, and subordinates; see Table 25.1, Entries 32, 37, and 41) and also presented results for summary scales and for items, their findings were used.

Garb and Schramke (1996; Table 25.1, Entry 6) examined judgments about cognitive impairment derived from neuropsychological test findings. Results were computed from all

the interrater data in their review. Gosling (2001; Table 25.1, Entry 24) examined temperament in mammals. He evaluated the extent to which researchers or others familiar with the target animals could agree on their personality traits. Gosling's results were retabulated to allow ratings made on items to be separated from those made using scales. Because the within and between subject correlations (i.e., *Q*- and *R*-type correlations) he reported were similar, results from both designs were combined. Grueneich (1992; Table 25.1, Entry 12) examined the reliability of clinicians for diagnosing borderline personality disorder. The original studies he cited were checked to determine the *N*s used in the reliability analyses. Martin, Garske, and Davis (2000; Table 25.1, Entry 21) examined ratings on the extent to which patients had a strong and positive alliance with their therapists. The results reported here excluded one study that used patients as the raters. A meta-analysis on scoring the Rorschach Comprehensive System was identified in the literature search (Meyer, 1997). However, the results in Table 25.1 are from a more recent meta-analysis on this topic (Meyer et al., 2002; Table 25.1, Entry 5).

Ottenbacher (1993) examined the interpretation of visually plotted behavior change, as used in behavior therapy or single-case research designs. He reported an average interrater reliability coefficient of .58 but his results included intrarater reliability and percent agreement coefficients. Consequently, the journal studies in his meta-analysis were obtained and intrarater or percent agreement findings were omitted (Table 25.1, Entry 26). Intraclass correlations were used in nine samples (Gibson & Ottenbacher, 1988; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1986; Ottenbacher & Cusik, 1991), while correlations were used in two additional samples (DeProspero & Cohen, 1979; Park, Marascuilo, & Gaylord-Ross, 1990).

Ottenbacher, Hsu, Granger, and Fiedler (1996; Table 25.1, Entry 10) examined the Functional Independence Measure, which is an index of basic self-care skills. Studies in their Table 2 were obtained to determine *N* and to differentiate item- and scale-level reliability. One study was excluded because it did not report the reliability *N*. Finally, Yarnold and Mueser (1989; Table 25.1, Entry 22) conducted a small meta-analysis of the Type A behavior pattern as determined by the Structured Interview, which is an instrument developed for this purpose.

With respect to medically related studies, Barton, Harris, and Fletcher (1999; Table 25.1, Entry 35) examined the extent to which surgeons and/or nurses agreed on the presence of breast abnormalities based on a clinical examination. Results were computed from all the data in their review and one primary study was obtained to determine *N*. De Jonghe et al. (2000; Table 25.1, Entry 9) examined ratings from intensive care physicians or nurses on the extent to which patients were sedated from medications designed for this purpose (e.g., to facilitate time on a ventilator). De Kanter et al. (1993) conducted a meta-analysis on the prevalence of temporomandibular disorder symptoms, not the reliability of assessing those symptoms. However, they provided citations for all the interrater reliability analyses that had been conducted in the context of prevalence research. Those six studies were obtained. Two presented intrarater reliability coefficients, one presented no specific reliability results, and two others presented statistics that could not be cumulated for the current review (i.e., percent agreement; mean differences). Data from the remaining study (Dworkin, LeResche, DeRouen, & Von Korff, 1990) were combined with the new data reported by De Kanter et al. (Table 25.1, Entry 13).

The original studies in D'Olhaberriague, Litvan, Mitsias, and Mansbach's (1996; Table 25.1, Entry 28) systematic review on stroke classification were obtained to determine the number of patients and neurologists participating in each sample. An independent subsample of stroke classifications in Gross et al. (1986) was initially overlooked by D'Olhaberriague et al. but was included here. Etchells et al. (1997; Table 25.1, Entry 34) examined the extent to which cardiologists agreed on the presence of a systolic heart murmur. The authors reported the *N* for each study in their review, but not the number of cardiologist examiners. Because the primary sources were obscure, it was assumed that each study used two cardiologists. Goldman (1994; Table 25.1, Entry 39) examined the extent to which physician reviewers agreed on the quality of medical care received by patients.

Ismail and Sohn's (1999) systematic review identified 20 reliability studies on early childhood dental caries. These studies examined the extent to which dentists or hygienists agreed on the number of teeth or tooth surfaces that were decayed, had fillings, or were missing (Table 25.1, Entry 2). Two of their 20 studies contained intrarater reliability, 7 presented statistics that could not be cumulated for the current review, and 3 did not report *N* for the reliability analysis. A final excluded study computed reliability for student practice examinations that were all conducted under the direct supervision of a single trainer. Of the remaining seven studies, scale-level data were provided by Yagot, Nazhat, and Kuder (1990), while item-level classifications were provided by Dini, Holt, and Bedi (1988); Jones, Schlife, and Phipps (1992); Katz, Ripa, and Petersen (1992; only results for the dentists and hygienists); Marino and Onetto (1995; the number of examiners was not specified, so two was assumed; reliability was reported to be $> .90$, so the value .91 was used); O'Sullivan and Tinanoff (1996); and Paunio, Rautava, Helenius, Alanen, and Sillanpää (1993). The *N*s for this meta-

analysis refer to the number of children included in the reliability studies, not the number of teeth or tooth surfaces examined.

Myers and Farquhar (2001; Table 25.1, Entry 27) reviewed the extent to which physicians agreed that patients had clubbing in their fingers or toes (enlargement at the tips) based on a visual clinical examination. Although one study in their review produced outlier results (κ = .90) relative to the others (κ range from .36–.45), it was retained in the summary data. Nwosu, Khan, Chien, and Honest (1998) did not conduct a meta-analysis of reliability for measuring bladder volume by ultrasound, but they indicated the three citations in the literature that provided available data (Table 25.1, Entry 1). One study (Ramsay, Mathers, Hood, & Torbet, 1991) reported an imprecise Spearman correlation that had been rounded to one decimal place. However, they also presented a figure of raw data for their reliability analysis, so both a Pearson correlation and ICC were computed from their findings.

In a review examining the physical measurement of spinal canal and spinal cord size from MRI, CT, or X-ray scans, Rao and Fehlings (1999) found seven reliability studies in the literature and provided new data from their own study (Fehlings et al., 1999). Of the studies identified in their literature search, one reported intrarater reliability, three presented mean differences, and one provided findings for just 2 of the 20 variables examined. Thus, summary results (Table 25.1, Entry 4) came from Fehlings et al. (1999), Matsuura et al. (1989), and Okada, Ikata, and Katoh (1994).

Finally, Ascioglu, de Pauw, Donnelly, and Collette (2001; Table 25.1, Entry 40) used a novel approach to assess interrater agreement. Rather than reviewing the extent to which clinicians agreed a patient should be diagnosed with an invasive fungal infection, they examined the extent to which the definitions of invasive fungal infection in the research literature agreed with each other. Specifically, they found 60 studies that defined infection from the mold *Aspergillus*. They then classified 367 patients according to each definition to determine "how the same case would be evaluated if the patient had been entered into these different studies on the basis of the same diagnostic information" (Ascioglu et al., p. 36). Thus, the reliability coefficient indicates the extent to which findings in the research literature are based on the same definition of pathology.

Although the studies described in this section examine a wide range of phenomena, a number of notable holes remain in the evidence base. In particular, just one review addressed the Rorschach and none addressed the TAT. In addition, although one study examined borderline personality disorder, no summary evidence was available on the general reliability of assigning psychiatric diagnoses, which are probably the most frequent determinations made in psychology and psychiatry. Similarly, no evidence was available on commonly used psychiatric rating scales, such as the Hamilton Depression Rating Scale (HDRS) or Hamilton Anxiety Rating Scale (HARS). Finally, virtually all of the reviews addressed clinical issues and applications. With the exception of Gosling's (2001) overview of mammal personality traits, no studies addressed the reliability of determinations that are more commonly or uniquely made as part of academic enterprises, such as the peer review of manuscripts and grants for scientific quality or the determination of how many factors should be extracted from a correlation matrix. To fill in these gaps, I conducted focused searches for existing literature reviews that identified reliability studies and completed several new searches for primary studies.

Focused Searches for Literature Reviews

To obtain data on the reliability of contemporary psychiatric diagnoses according to the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM;* American Psychiatric Association, 1994), I searched for existing reviews of this literature. For *DSM* Axis I disorders, PubMed and PsycINFO were searched by combining the term *review* with the terms *DSM, Diagnostic-and-Statistical, Axis-I, SCID, Structured-Clinical-Interview, Semi-Structured,* or *Structured-Interview* and the terms *reliability, agreement, interrater,* or *inter-rater.* Non-English-language articles were excluded from PsycINFO. After combining 305 PubMed citations with 143 PsycINFO citations and deleting duplicates, 394 citations remained. Abstracts were reviewed and relevant articles were obtained. Several reviews identified in the search addressed interrater reliability for narrow diagnostic constructs (e.g., Steinberg, 2000; Szatmari, 2000; Weathers, Keane, & Davidson, 2001). Because it would be most valuable to have data organized across the wide range of conditions diagnosed by the *DSM,* these studies were not used. Instead, the data summarized here came from the studies included in Segal, Hersen, and Van Hasselt's (1994) review of reliability for the Structured Clinical Interview for *DSM-III-R* (SCID). Original studies were obtained to determine the number of raters in each sample. In the process, it was discovered that one study (Stukenberg, Dura, & Kielcolt-Glaser, 1990) did not report reliability for any specific diagnoses but rather just for the presence or absence of any *DSM* disorder. This finding was excluded. Summary results were computed for both joint interview designs (Table 25.1, Entry 15) and for two independently conducted interviews (Table 25.1, Entry 29).

To find reviews covering the reliability of *DSM* Axis II personality disorders, PubMed and PsycINFO were searched by combining the term *review* with the terms *personality-disorder* or *Axis-II* and the terms *reliability, agreement, interrater,* or *inter-rater.* After deleting duplicates, 101 citations remained. These abstracts and two recent special sections in the *Journal of Personality Disorders* (Livesley, 2000a, 2000b) were examined for relevant review articles.

Two reviews were found that summarized reliability across Axis II diagnoses. Segal et al. (1994) focused exclusively on the SCID, while Zimmerman (1994) cast a broader net and examined data derived from all the semistructured interviews in use for this purpose. Because Zimmerman's overview suggested there was little difference in the reliability of interview schedules, his results were tabulated here. Zimmerman reported unweighted averages across studies for each Axis II diagnosis (his Table 5). To be consistent with the other findings summarized here, overall reliability was recomputed by averaging results within a study and then weighing by sample size. Zimmerman also provided preliminary data from an unpublished sample of Loranger's. This information was supplemented with the complete findings described in the final published report (Loranger et al., 1994). As with Axis I disorders, summary results were computed for joint interviews (Table 25.1, Entry 16) and for independent interviews conducted over a short interval (Table 25.1, Entry 30).

New Meta-Analytic Studies of Interrater Reliability

By searching the literature for narrative reviews it was possible to generate meta-analytic summaries for the reliability of Axis I and Axis II diagnoses obtained through commonly used semistructured interviews. Although valuable, to further the knowledge base concerning interrater agreement a series of new meta-analyses was undertaken addressing psychological tests (i.e., Rorschach, TAT, and Hamilton scales) and traditional academic pursuits (i.e., the ubiquitous peer-review process and the interpretation of scree plots).

Rorschach Scales

Two new meta-analyses were conducted to summarize interrater reliability of Rorschach scoring. One examined the Rorschach Oral Dependency (ROD) Scale and the other Klopfer's Rorschach Prognostic Rating Scale (RPRS). The ROD was selected because it appears to be the most commonly studied Rorschach scale in the literature (see Bornstein, 1996, for an overview). The RPRS was selected because its predictive validity was examined in two recent meta-analyses (Meyer, 2000; Meyer & Handler, 1997).

To identify ROD interrater reliability studies, Bornstein's (1996) review was consulted and three PsycINFO searches were undertaken. After the first two were completed, Robert Bornstein recommended the third search and identified a relevant study that was not otherwise identified (personal communication, September 22–25, 2001). The search strategies included (1) the term *Masling* combined with *orality* or *oral-depen**, (2) Robert F. Bornstein as an author of any study, and (3) the term *Rorschach* combined with Samuel Juni as an author. In total, these searches identified 168 nonduplicate articles, 89 of which appeared likely to contain relevant data. These studies were obtained and *r*, kappa, or ICC results were obtained from 37. However, to ensure that only independent samples contributed data, when different studies reported the same reliability *N*, the same reliability value, and the same type of participants, only one sample was allowed into the analyses. In the end, data were obtained from 31 studies and 40 samples (Table 25.1, Entry 3). A complete list of citations is in the Appendix.

To identify studies on RPRS interrater reliability, I searched PsycINFO for English-language articles that contained *prognostic-rating-scale* combined with *Rorschach* or *Klopfer.* Thirty-five studies were identified. Seven additional studies were obtained from Goldfried, Stricker, and Weiner's (1971) early review of this literature or Meyer and Handler's (1997) meta-analysis. Out of these 42 studies, 8 provided interrater coefficients that could be cumulated (Table 25.1, Entry 19). Data came from Cooper, Adams, and Gibby, (1969; estimated from the range reported); Edinger and Weiss (1974); Endicott and Endicott (1963; pre- and posttreatment samples); Hathaway (1982); Newmark, Finkelstein, and Frerking (1974; pre- and posttreatment samples); Newmark, Hetzel, Walker, Holstein, and Finkelstein (1973; pre- and posttreatment samples); Newmark, Konanc, Simpson, Boren, and Prillaman (1979); and Williams, Monder, and Rychlak (1967).

TAT Scales

Three new meta-analyses were conducted to summarize the interrater reliability of TAT scoring. Although there are many TAT scales described in the literature, several of which have been studied extensively (e.g., the Achievement, Intimacy, and Power motives), the scales examined here have been actively researched in clinical samples over the past several years. They include the Defense Mechanism Manual (DMM), the Social Cognition and Object Relations Scale (SCORS), and the Personal Problem-Solving Scale (PPSS).

To identify relevant data, I initially searched PsycINFO for all English-language studies published after 1989 that in-

cluded the term *thematic-apperception-test.* This search identified 172 articles; 11 had reliability data on the SCORS, 7 contained data for the DMM, and 2 had data for the PPSS. Next, these articles were used to identify additional, nonredundant reliability studies. Nine additional sources were found for the DMM (Table 25.1, Entry 20), four more were found for the SCORS (Table 25.1, Entry 18), and one new study was found for the PPSS (Table 25.1, Entry 11). A list of the 16 DMM studies and 15 SCORS citations can be found in the Appendix.

For the PPSS, Pearson correlations were obtained from the three samples in Ronan, Colavito, and Hammontree (1993) and the two samples in Ronan et al. (1996). ICC results were obtained from the two studies reported in Ronan, Date, and Weisbrod (1995).

The Hamilton Rating Scales

To identify studies examining the interrater reliability of the HDRS and HARS, the English-language literature on PsycINFO and PubMed was searched by combining the scale names with the terms *reliability* or *agreement.* A total of 168 unique citations were identified. Based on the article abstracts, 52 of these appeared likely to contain relevant data. In addition, the text and bibliographies in these articles were searched for other relevant publications, which produced 25 new citations. In total, 77 articles were examined for relevant HDRS and HARS data, including Hedlund and Vieweg's (1979) review of HDRS research. Studies were excluded if they were missing information required for aggregation (e.g., *N;* type of reliability design; *r,* kappa, or ICC as an effect size). One additional study was omitted. Gottlieb, Gur, and Gur (1988) reported an intraclass correlation of .998. However, their mean HDRS score was approximately 3.0 ($SD \approx 4.5$) and the exact agreement rate was approximately 60%, with raters differing by up to 5 raw score points. The latter findings appear incompatible with an ICC of .998. In the end, 51 citations provided relevant information. A complete list organized by test can be found in the Appendix.

To summarize HDRS and HARS reliability, results were partitioned into designs in which clinicians based their ratings on separate interviews and those in which ratings were based on the same fixed source of information. The latter category included instances when both raters were present for the same interview, when one or more raters viewed a videotape, and when the patient's self-report fully determined both sets of ratings being compared (e.g., the patient completed the scale in writing and then was asked the same questions in a highly structured oral format; the patient was given a highly structured personal interview and then immediately thereafter was asked the same highly structured questions via the telephone).

Designs that rely on separate interviews can be confounded by the time lapse between interviews. As this delay increases, greater changes in the patient's emotional state should be expected, which would reduce reliability for reasons unrelated to rater agreement. Thus, separate interview studies were excluded if the retest interval exceeded the time frame of the ratings. For instance, although joint interview results were used from Potts, Daniels, Burnam, and Wells (1990), their separate interview results were excluded. In this study, subjects reported on their affect over the past month. However, the average interval between the first and second interviews was 15 days with an *SD* of 11 days, suggesting that in a salient proportion of interviews the patient would have been describing his or her mood over distinct time frames.

Peer Review of Research

To identify studies examining the interrater reliability of the scientific peer-review process, I searched the English-language literature using PsycINFO and PubMed. For PsycINFO, the terms *peer-review* or *manuscript-review* were combined with the terms *reliability* or *agreement.* For PubMed, the MeSH category of *peer review* was combined with the MeSH categories of *reproducibility of results* or *observer variation* and results were limited to citations containing an abstract. A total of 159 unique citations were found, 34 of which appeared likely to contain relevant data based on the abstract. In addition, 34 new citations were found by reviewing these articles for other relevant studies. In total, 68 articles were examined in detail, including Cicchetti's (1991) major review, which contained results from 26 samples of manuscript or grant submissions. When it was necessary to obtain information not included in Cicchetti's review (e.g., *N*), the original citations were obtained.

My initial goal was to update Cicchetti's (1991) review of agreement between two peer reviewers. However, studies have examined other aspects of reliability related to scientific peer review. For instance, Justice, Berlin, Fletcher, Fletcher, and Goodman (1994) examined the extent to which different types of physicians agreed on the quality of manuscripts slated to be published in the *Annals of Internal Medicine.* Ratings of quality made by journal readers who expressed interest in the article's topic were compared to ratings from (1) other journal readers who expressed the same interest (weighted kappa = .05; $N = 159$), (2) standard peer reviewers for the journal (weighted kappa = −.02; $N = 352$), and (3) experts in the relevant area of research (weighted kappa

= −.01; N = 371). Although these findings suggest that different consumers of the research literature have very different views on the quality of published studies, because this was the only investigation of its kind, the results were not included in Table 25.1. However, two other types of designs occurred more frequently and their results were aggregated. In one design, used in five samples, the mean ratings obtained from one panel of grant reviewers were compared to the mean ratings obtained from a second, independent panel of grant reviewers. When interpreting these results, it should be kept in mind that ratings averaged across multiple raters sharply curtail the random error that would be associated with any single rater. Thus, results from these averaged ratings will produce markedly higher reliability coefficients than would be obtained from any two individual grant reviewers (e.g., Scullen, 1997). The second type of reliability design, used in five studies, investigated the extent to which editors agreed with each other on the quality of the reviews submitted by manuscript reviewers.

Ultimately, reliability data addressing aspects of the peer-review process were obtained from 34 citations. Virtually all of the excluded citations did not contain relevant reliability data, though Das and Froehlich's (1985) findings on reviewer agreement were excluded because they used an ICC formula that differed from standard formulas (McGraw & Wong, 1996), and Gottfredson's (1978) findings were not used because the design relied on previously published articles as well as raters who were nominated by the original authors.

To summarize the reliability of peer reviewers (Table 25.1, Entry 42), various forms of scientific review were combined, including journal or conference submissions, grant applications, and submissions to research ethics committees or institutional review boards. Results were partitioned into dimensional ratings and dichotomous decisions. If a study did not report a relevant coefficient but provided raw data, results were computed from that information. Many studies addressing manuscript review examined publication recommendations as well as more specific questions, such as the quality of the literature review or appropriateness of the analyses. Within a study, the most direct measure of overall merit was used here. Doing so produced slightly higher estimates of reliability ($p <$ .0005). For studies reporting both types of results, direct ratings of quality had a mean reliability of .26 (N = 10,514), while ratings for more specific items had a reliability of .23 (N = 9,695). Finally, if a study presented both unweighted kappa and weighted kappa or an ICC, the weighted kappa or ICC was used in the analyses. A list of final studies providing peer-reviewer agreement data is provided in the Appendix.

With respect to agreement among editors on the quality of peer reviewers or their submitted reviews (Table 25.1, Entry 25), one study used the Pearson correlation (van Rooyen, Godlee, Evans, Black & Smith, 1999). The four studies that contributed chance-corrected reliability statistics were Callaham, Baxt, Waeckerle, and Wears (1998); Feurer et al. (1994); van Rooyen, Black, and Godlee (1999); and Walsh, Rooney, Appleby, and Wilkinson (2000).

Studies that examined agreement between the mean ratings of one panel of grant reviewers and the mean ratings from a separate panel of reviewers (Table 25.1, Entry 36) were partitioned into dimensional ratings and dichotomous determinations (i.e., fund vs. not). The four studies that contributed correlations for dimensional ratings were Hodgson (1995, 1997); Plous and Herzog (2001; from raw data); and Russell, Thorn, and Grace (1983; omitting the confounded panel A with panel D data). The three studies that contributed chance-corrected statistics for dimensional ratings were Hodgson (1995, 1997) and Plous and Herzog (2001). The studies that contributed chance-corrected statistics for dichotomous funding decisions were Cicchetti (1991) and Hodgson (1997).

Cattell's Scree Plots

To identify articles containing interrater reliability data for using Cattell's scree plot to determine the proper number of factors to extract in a factor analysis or cluster analysis, the English-language literature on PsycINFO was searched using the terms *scree-plot* or *scree-test* and *reliability* or *agreement.* Eleven articles were identified. Three of these provided relevant data and their bibliographies identified a fourth study (Table 25.1, Entry 38). ICCs were obtained from Cattell and Vogelmann (1977; from raw data), Crawford and Koopman (1979), Lathrop and Williams (1987), and Streiner (1998). Two of the studies investigated the influence of factor analytic expertise and found no differences between experts and novices who had been provided with some training. Nonetheless, the summary results should be treated cautiously because reliability varied widely across studies, ranging from a low of .00 to a high of .89.

Summary of Meta-Analytic Studies of Interrater Reliability

The interrater reliability meta-analyses examine a diverse set of topics. The targeted constructs vary substantially in their complexity and in the methods used to obtain measurements (e.g., physical measurements with a ruler vs. application of complex theoretical constructs to narrative productions). The constructs also range from scoring tasks that code very discrete or circumscribed events (e.g., Entries 1 through 5) to

interpretive tasks that code more abstract or higher level inferences (e.g., Entries 24, 32, 37, 39, and 42). Nonetheless, in the context of this chapter, the main point embedded in Table 25.1 appears clear. The interrater reliability observed when coding Rorschach or TAT protocols falls in the range between .80 and .91. This level of agreement compares favorably with the reliability seen for a wide range of other determinations made in psychology and medicine.[2]

Also, from a purely psychometric perspective, the findings in Table 25.1 indicate that scales generally are more reliable than items. Although it is a gross comparison, in the 17 instances when these levels of measurement could be directly compared, scales had an average reliability of .77, while items had an average reliability of .62. Thus, aggregated determinations that mathematically combine lower level judgments are more reliable than single observations. Surprisingly, while chance-corrected statistics did produce somewhat lower estimates of reliability than correlations, the difference was not large. Across the 16 topics that provided both types of statistics, the average kappa/ICC was .70 and the average correlation was .74.

Overall, the data in Table 25.1 do not support the notion that interrater reliability coefficients for the Rorschach or TAT are deficient relative to other tests or applied judgments in psychology and medicine. However, it remains possible that psychometric deficiencies will be evident when considering other types of systematically collected data. To address this prospect, evidence on test-retest reliability will be examined and then evidence on test validity.

META-ANALYSES OF TEST-RETEST RELIABILITY

For any variable that is thought to measure a stable characteristic, evidence on test-retest reliability is very pertinent because it indicates whether the test scale validly measures a traitlike feature of the person rather than a transient statelike quality. One could wonder if tests like the Rorschach or TAT show comparatively less traitlike consistency than other personality tests.

To identify meta-analyses examining stability over time, I searched the English-language literature on PsycINFO and PubMed. In both databases, the terms *test-retest, retest-reliability, retest-stability, stability and reliability,* or *stability and consistency* were combined with terms restricting the search to meta-analyses or systematic literature reviews. In total, 101 unique studies were identified. Based on the abstracts, 19 articles seemed likely to contain aggregated retest coefficients and these were obtained. In addition, I summarized any systematically gathered retest data that had been reported in the interrater reliability meta-analyses discussed above. Three of those studies presented relevant data (Gosling, 2001; Yarnold & Mueser, 1989; Zimmerman, 1994), though one had also been identified in the search for stability studies.[3]

Of the 21 relevant studies, several were excluded. Two examined statelike variables (i.e., depression and happiness), three others were superseded by more recent and comprehensive meta-analyses, and one did not provide a quantitative summary of the literature. From the final group of 15 meta-analyses, 2 provided stability coefficients for performance tests of personality (Parker, Hanson, & Hunsley, 1988; Roberts & DelVecchio, 2000). Results from all the meta-analyses are presented in Table 25.2 and each will be described briefly.

Roberts and DelVecchio (2000) completed a comprehensive meta-analysis on the long-term stability of personality traits. They relied on data from nonclinical samples and excluded studies with a retest interval of less than 1 year, leaving them with a final sample of 152 longitudinal studies that had an average retest interval of 6.7 years. The authors subsequently categorized personality tests into three types of methods using data from 135 samples. Rorschach, TAT, and sentence completion scales were treated as one method category, self-report scales comprised a second category, and observer rating scales formed the final category. As indicated in Table 25.2 (Entry 1), after statistically adjusting for sample age and fixing the estimated retest interval to 6.7 years, they found retest correlations of .45 for the Rorschach, TAT, or SCT; .50 for self-reports; and .51 for observer ratings. While encouraging more research, the authors concluded "Given the small magnitude of the difference, we feel the most impressive feature of these analyses is the lack of substantive differences between the three primary methods of assessing traits" (p. 16).

Parker, Hanson, and Hunsley (1988) reached a similar conclusion. They examined the 1970 to 1981 journal literature and compared stability for the Rorschach, Wechsler Adult Intelligence Scale (WAIS), and Minnesota Multiphasic Personality Inventory (MMPI). Although they did not report average retest intervals, reliability was .74, .82, and .85 for the MMPI, WAIS, and Rorschach, respectively (Table 25.2, Entry 16).

Of the remaining meta-analyses, Ashton (2000) examined the retest reliability of professional judgments in medicine, psychology, meteorology, human resources, and business (mainly accounting). These studies examine intrarater reliability and quantify the extent to which the same professional evaluates the same information equivalently at two different points in time. Following Ashton, studies that obtained both judgments in the same experimental session were omitted, relying instead on just those findings he classified as indic-

TABLE 25.2 Meta-Analyses of Test-Retest Reliability From the Psychological Literature

Study/test or method	Mean retest interval in months	*N* of		Test-retest *r*	
		Samples	Subjects	TAT, SCT, or Rorschach	Other tests
1. Roberts and DelVecchio (2000); Personality Traits, Controlling for Age and Retest Duration[a]					
Observer rating	80.4	54	11,662		.51
Self-report	80.4	73	46,196		.50
Rorschach, TAT, or SCT	80.4	8	1,083	.45	
2. Schuerger and Witt (1989); Individually Tested IQ[b]					
Adults aged 18–24	72.0	79	—		.79
Children aged 6–9	72.0	79	—		.72
3. Holden and Miller (1999); Parental Child Rearing[c]					
Self-reported beliefs	82.7	10	875		.50
Observed and self-reported	36.4	47	3,786		.45
Observed behaviors	18.4	33	2,587		.41
Observed behaviors	≤1	11	—		.59
4. van Ijzendoorn, Schuengel, and Bakermans-Kranenburg (1999); Disorganized Child-Parent Attachment	25.2	14	840		.34
5. Viswesvaran and Ones (2000); "Big Five" Personality Scales (*M*s across 5 scales)[d]	19.5	170	41,074		.73
6. Gosling (2001); Mammal Personality Traits					
Items	18.6	4	64		.46
Scales	12.0	1	12		.92
7. Schuerger and Witt (1989); Individually Tested IQ[b]					
Adults aged 18–24	12.0	79	—		.85
Children aged 6–9	12.0	79	—		.80
8. Zimmerman (1994); DSM Personality Disorders[e]	7.1	7	457		.44
9. Holland, Johnston, and Asama (1993); Self-Report Vocational Identity Scale[f]	5.4	13	1,708		.68
10. Swain and Suls (1996); Physiological Reactivity to Stressors[g]					
Heart rate	3.0	95	—		.55
Blood pressure	3.8	73	—		.38
11. Ashton (2000); Professional Judgments (Intrarater Stability)[h]					
Medicine and psychology	4.1	6	1,997		.73
Human resources	2.7	11	36,712		.73
Business	1.7	3	2,576		.82
Meteorology	1.6	2	300		.87
All studies	2.9	22	41,585		.76
12. McKelvie (1995); Self-Reported Vividness of Visual Imagery[i]	1.3	7	—		.74
13. Schuerger and Witt (1989); Individually Tested IQ[b]					
Adults aged 18–24	1.0	79	—		.92
Children aged 6–9	1.0	79	—		.83
14. Yarnold and Mueser (1989); Type A Behavior					
Structured Interview	—	1	—		.68
Jenkins Activity Survey	—	7	—		.71
15. Shulman (2000); Clock Drawings for Cognitive Impairment	—	4	277		.76
16. Parker, Hanson, and Hunsley (1988); Common Psychological Tests					
Rorschach	—	2	125	.85	
WAIS	—	4	93		.82
MMPI	—	5	171		.74
17. Capraro, Capraro, and Henson (2001); Mathematics Anxiety Rating Scale (Self-Report)	—	7	—		.84

[a]The number of subjects was obtained from the data reported in their Table 1. However, the values reported here are slight overestimates because they are based on 65 observer rating samples; 80 self-report samples; and 9 Rorschach, TAT, or SCT samples.

[b]The reported values are from regression estimates for specified ages and retest intervals using data from 79 samples.

[c]Retest intervals for the first three coefficients and *N*s for the first and third were obtained from raw data in their Table 4.

[d]The estimated retest interval was computed as a weighted average across the five scales. Total sample size was estimated from the mean sample size across the 158 studies that provided this information.

[e]Reliability is reported using the kappa coefficient, not *r*.

[f]Nonindependent samples were excluded.

[g]Average retest duration was estimated from data reported in their Table 3, assigning an estimate of 2 years to studies that were simply described as being longer than 1 year.

[h]The average retest duration and *N* were obtained from data reported in his Table 1. Virtually all studies used a multirater design, so the reported *N* was computed as $n(k - 1)$.

[i]The average retest interval was estimated as the midpoint of the range across studies (i.e., 3 to 7 weeks).

ative of stability (Table 25.2, Entry 11). Capraro, Capraro, and Henson (2001) examined stability for the self-report Mathematics Anxiety Rating Scale. Few details were provided about the seven retest samples in this review (Table 25.2, Entry 17). Gosling's (2001) review of animal behavior traits provided stability data for one small study using scale-level ratings and four small samples using item-level ratings (Table 25.2, Entry 6).

Holden and Miller (1999) examined the stability of parental child-rearing beliefs and behaviors. Beliefs were assessed with self-report inventories and behaviors were observed and coded by others. Although their primary interest was in studies with retest intervals greater than 1 month, they also provided the median reliability from 11 short-term observational studies (Table 25.2, Entry 3). Holland, Johnston, and Asama (1993) presented a table of quantitative findings on the stability of the Vocational Identity Scale, which measures one's propensity to have clear and stable goals, interests, and talents (Table 25.2, Entry 9). McKelvie (1995) summarized the short-term stability findings for the self-rated Vividness of Visual Imagery Questionnaire (Table 25.2, Entry 12). Schuerger and Witt (1989) examined the stability of measured intelligence using the Wechsler scales and the Stanford-Binet. Data from 79 child and adult samples retested over various intervals were used to develop regression equations that predicted stability as a function of age and retest duration. For the findings in Table 25.2, two sets of ages and three time intervals were selected to illustrate stability (Table 25.2, Entries 2, 7, and 13).

Swain and Suls (1996) examined the stability of physiological changes in response to laboratory stressors. They presented findings separately for heart rate, systolic blood pressure, and diastolic blood pressure, though the blood pressure results were averaged for Table 25.2 (Entry 10). Shulman (2000) presented a table of quantitative findings regarding the stability of clock drawings as a neuropsychological procedure to quantify cognitive impairment (Table 25.2, Entry 15). When aggregating information from this review, it was assumed that the reported *N*s indicated the number of subjects included in the reliability analyses. Van Ijzendoorn, Schuengel, and Bakermans-Kranenburg (1999) conducted a meta-analysis on the stability of the disorganized pattern of child-parent attachment (Table 25.2, Entry 4).

Viswesvaran and Ones (2000) provided a comprehensive examination of self-report personality inventories that are currently used in personnel selection. Findings for nonpatient samples were obtained from the technical manuals for 24 tests and results were organized in the framework of the Five-Factor Model (Table 25.2, Entry 5). Yarnold and Mueser's (1989) review of Type A measures provided summary data from seven samples that used the self-report Jenkins Activity Survey and one sample that used the Structured Interview for Type A behavior (Table 25.2, Entry 14). Finally, Zimmerman (1994; supplemented by the additional data reported in Loranger et al., 1994) provided retest findings from seven samples in which structured interviews were used to assess *DSM* personality disorders over moderately long intervals (Table 25.2, Entry 8).

Findings in Table 25.2 have been roughly organized by retest duration, with longer intervals at the top of the table. The three studies that did not report retest durations are listed at the bottom of the table. Based on the information available, when looking across studies, the length of the retest interval tends to be negatively related to stability ($r = -.34$, $N = 27$; $p = .078$), which is consistent with findings from most of the larger meta-analyses (e.g., Roberts & DelVecchio, 2000; Schuerger & Witt, 1989). However, in Table 25.2 the magnitude of the association between longer retest intervals and lower retest reliability is not as strong as might be anticipated. This may be due to the fact that some constructs are less stable than others (e.g., interview-based personality disorder diagnoses may be less stable than test-derived personality traits) or to methodological variables. For instance, items should be less stable than scales (e.g., Gosling's item-level findings vs. Roberts and DelVecchio's scale-level findings), dichotomous determinations may be less stable than scores on a graduated continuum (e.g., Zimmerman's dichotomous diagnoses vs. Roberts and DelVecchio's continuous traits), and chance-corrected reliability coefficients (i.e., Zimmerman's findings) likely produce lower estimates of stability than correlations. Furthermore, many of the findings in Table 25.2 are based upon a small number of observations. This is particularly true for the findings reported in Gosling (2001) and for the Structured Interview coefficient reported by Yarnold and Mueser (1989). Caution is also warranted with Parker et al.'s (1988) findings, particularly for the Rorschach and WAIS.

Keeping these limitations in mind, it remains the case that Table 25.2 provides no support for the idea that tests like the Rorschach, TAT, or SCT are less stable than other instruments. In fact, the two systematic reviews that examined test-retest reliability for both performance tests of personality and other testing methods found roughly comparable results across methods. Thus, the overall pattern of findings in Table 25.2 is quite similar to that observed in Table 25.1. The existing systematically gathered evidence does not suggest that performance tests of personality have deficient stability relative to other types of personality assessment instruments.

Nonetheless, it is possible that the genuine limitations associated with these tests will be evident when validity data is considered. Even though different kinds of reliability findings

may be comparable across tests, performance measures of personality like the Rorschach and TAT may be notably deficient when considering the ultimate question of test validity.

META-ANALYSES OF TEST VALIDITY

Recently, Meyer et al. (2001) presented 133 meta-analytically derived validity coefficients for psychological and medical tests. Meyer and Archer (2001) extended that overview by computing 20 new or refined meta-analytic validity coefficients for the Rorschach, MMPI, and Wechsler intelligence tests. Table 25.3 provides findings from these sources. The table includes the 38 Rorschach, MMPI, and IQ validity coefficients that were reported in Meyer and Archer's summary table (their Table 4), all the thematic story and self-report validity coefficients from Bornstein's (1998c, 1999) meta-analyses on dependency, and all the other thematic story and sentence completion validity coefficients that were reported in Table 2 of Meyer et al. (2001). Because this selection process provided a wide range of validity data for the MMPI and intelligence tests, Meyer et al.'s results for other psychological tests were not reproduced here. However, to provide a representative sampling of medical test validity, every third finding was selected from the 63 medical validity coefficients listed in Table 2 of Meyer et al.

When considering Table 25.3, it is important to keep several caveats in mind. First, all examples make use of coefficients that were not corrected for unreliability, range restriction, or the imperfect construct validity of criterion measures. Second, all the coefficients do not come from equivalent designs. Some studies selected extreme groups of participants; examined rare, low base rate events; artificially dichotomized truly continuous variables; employed relatively small samples; or used procedures not typically found in applied clinical practice. These methodological factors can influence validity coefficients and make them systematically differ in size (Hunter & Schmidt, 1990). Thus, even though table entries are organized by magnitude, differences in ranking cannot be taken to mean that one test is globally better than another.

With these points in mind, the data in Table 25.3 lead to several observations (Meyer & Archer, 2001; Meyer et al., 2001). First, both psychological and medical tests have varying degrees of validity, ranging from tests that are essentially unrelated to a criterion, to tests that are strongly associated with relevant criteria. Second, it is difficult to distinguish the validity coefficients for psychological and medical tests. These tests do not cluster at different ends of the validity continuum but instead produce coefficients that are interspersed throughout its range. Third, test validity is conditional. A given test produces stronger validity for some criteria and weaker validity for other criteria. Finally, as with the interrater and retest reliability data in Tables 25.1 and 25.2, the broad review of systematically collected validity data in Table 25.3 does not reveal uniformly superior or uniformly inferior methods of psychological assessment.

Despite the arguments and criticisms that have been leveled against performance-based tests of personality, the Rorschach, TAT, and SCT do not produce noticeably lower validity coefficients than alternative personality tests. Instead, performance tests of cognitive ability, performance tests of personality, and self-report tests of personality all produce a range of validity coefficients that vary largely as a function of the criterion under consideration.

SUMMARY CONCLUSIONS

The purpose of this chapter was to provide a sound and thorough examination of systematically gathered evidence addressing the psychometric properties of tests like the Rorschach and TAT. More than any other personality assessment instruments, these kinds of measures have been harshly criticized. In fact, recent arguments about their psychometric defects have been accompanied by rallying cries for attorneys to attack their use whenever possible and for psychologists to banish them from applied clinical practice (e.g., Garb, 1999; Grove & Barden, 1999; Lilienfeld et al., 2000; Wood et al., 2000). The latter are dramatic calls for action. Given that these proposals are presented as recommendations emerging from the scientific literature, one should anticipate that the systematically gathered evidence on reliability and validity would unambiguously demonstrate how inferior these methods of personality assessment really are. However, the meta-analytic evidence does not provide such a demonstration. If anything, the findings reveal the opposite. The Rorschach, TAT, and SCT produce reliability and validity coefficients that appear indistinguishable from those found for alternative personality tests, for tests of cognitive ability, and for many medical assessment procedures.

At the same time, there are limitations to the data presented in this chapter that should be appreciated. Considering Table 25.1, it is likely that some approaches to scoring tests like the Rorschach or TAT are less reliable than those included in this table. (Of course, some are likely to be more reliable as well.) However, the Rorschach and TAT scales in the table were not selected because of the reliability results they produced. They were selected because they have all been part of recent publications addressing clinical topics. As such,

TABLE 25.3 Summary Effect Sizes (r) From Meta-Analyses Examining Test Validity

Predictor test and criterion	Rorschach, TAT, or SCT	Other psych test	Medical test	*N*
1. Dexamethasone Suppression Test Scores and Response to Depression Treatment			.00	2,068
2. Routine Ultrasound Examinations and Successful Pregnancy Outcomes			.01	16,227
3. MMPI Ego Strength Scores and Subsequent Psychotherapy Outcome		.02		280
4. Unique Contribution of an MMPI High Point Code (vs. Other Codes) to Relevant Criteria[a]		.07		8,614
5. MMPI Scores and Subsequent Prison Misconduct		.07		17,636
6. MMPI Elevations on Scales F, 6, or 8 and Criminal Defendant Incompetency		.08		1,461
7. In Cervical Cancer, Lack of Glandular Differentiation on Tissue Biopsy and 5+ Year Survival			.11	685
8. MMPI Scale 8 and Differentiation of Schizophrenic versus Depressed Disorders		.12		2,435
9. Lower General Cognitive Ability and Involvement in Automobile Accidents		.12		1,020
10. General Intelligence and Success in Military Pilot Training		.13		15,403
11. Rorschach DEPI and Detection of Depressive Diagnosis	.14			994
12. MMPI Scale 2 and Differentiation of Neurotic versus Psychotic Disorders		.14		6,156
13. MMPI Scale 8 and Differentiation of Neurotic versus Psychotic Disorders		.14		6,156
14. Baseline IQ and Subsequent Psychotherapy Outcome		.15		246
15. Low Serotonin Metabolites in Cerebrospinal Fluid (5-HIAA) and Subsequent Suicide Attempts			.16	140
16. MMPI Cook-Medley Hostility Scale Elevations and Subsequent Death From All Causes		.16		4,747
17. Motivation to Manage from the Miner Sentence Completion Test and Managerial Effectiveness	.17			2,151
18. MMPI Validity Scales and Detection of Known or Suspected Under-Reported Psychopathology		.18		328
19. Dexamethasone Suppression Test Scores and Subsequent Suicide			.19	626
20. Self-Reported Dependency Test Scores and Physical Illness		.21		1,034
21. Rorschach to Detect Thought Disturbance in Relatives of Schizophrenic Patients	.22			230
22. MMPI Dependency Scale and Dependent Behavior		.22		320
23. TAT Scores of Achievement Motivation and Spontaneous Achievement Behavior	.22			(k = 82)
24. Traditional Electrocardiogram Stress Test Results and Coronary Artery Disease			.22	5,431
25. WISC Distractibility Subscales and Learning Disability Diagnoses		.24		(K = 54)
26. Decreased Bone Mineral Density and Lifetime Risk of Hip Fracture in Women			.25	20,849
27. General Intelligence Test Scores and Functional Effectiveness Across Jobs		.25		40,230
28. Self-Reported Dependency Test Scores and Dependent Behavior		.26		3,013
29. Thematic Story Test Dependency Scores and Recalled Physical Illness	.29			269
30. C-Reactive Protein Test Results and Diagnosis of Acute Appendicitis			.28	3,338
31. General Validity of Rorschach Studies Without Method Confounds	.29			6,520
32. General Validity of MMPI Studies Without Method Confounds		.29		15,985
33. For Women, Electrocardiogram Stress Test Results and Detection of Coronary Artery Disease			.30	3,872
34. MMPI Scale 2 and Differentiation of Schizophrenic versus Depressed Disorders		.31		2,435
35. General Validity of Rorschach Hypotheses Without Method Confounds	.32			(k = 523)
36. General Validity of MMPI Hypotheses (Includes Some Method Confounds)		.32		(k = 533)
37. Screening Mammogram Results and Detection of Breast Cancer Within 1 Year			.32	263,359
38. General Validity of WAIS Studies Without Method Confounds		.33		3,593
39. MMPI Scale 2 or DEP and Detection of Depressive Diagnosis		.35		2,905
40. Incremental Contribution of Rorschach PRS Scores Over IQ to Predict Treatment Outcome	.36			290
41. General Validity of WAIS Hypotheses Without Method Confounds		.36		(k = 104)
42. Papanicolaou Test (Pap Smear) and Detection of Cervical Abnormalities			.36	17,421
43. Competency Screening Sentence Completion Test Scores and Defendant Competency	.37			627
44. Rorschach Oral Dependence Scale and Dependent Behavior	.37			1,320
45. Sperm Penetration Assay Results and Success With in Vitro Fertilization			.39	1,335
46. MMPI Validity Scales to Detect Under-Reported Psychopathology (Primarily Analog Studies)		.39		2,297
47. Computed Tomography Results and Detection of Lymph Node Metastases in Cervical Cancer			.41	1,022
48. MMPI Scale 8 and Differentiation of Psychiatric Patients versus Controls		.42		23,747
49. Conventional Dental X-rays and Diagnosis of Between-Tooth Cavities (Approximal Caries)			.43	(K = 8)
50. Rorschach SCZI and Detection of Psychotic Diagnosis	.44			717
51. MMPI Scale 2 and Differentiation of Psychiatric Patients versus Controls		.44		23,747
52. WAIS IQ and Obtained Level of Education		.44		(k = 9)
53. Digitally Enhanced Dental X-rays and Diagnosis of Biting Surfaces Cavities			.44	2,870
54. Rorschach PRS Scores and Subsequent Psychotherapy Outcome	.45			624
55. MMPI Validity Scales and Detection of Known or Suspected Malingered Psychopathology		.45		771
56. Thematic Story Test Dependency Scores and Dependent Behavior	.46			448
57. Rorschach X+% and Differentiation of Clinical or Target Groups From Controls	.46			1,517
58. Antineutrophil Cytoplasmic Antibody Testing and Detection of Wegener Granulomatosis			.47	13,562
59. Lecithin/Sphingomyelin Ratio and Prediction of Neonatal Respiratory Distress Syndrome			.50	1,170

(continued)

TABLE 25.3 *(Continued)*

Predictor test and criterion	Rorschach, TAT, or SCT	Other psych test	Medical test	*N*
60. WAIS IQ Subtests and Differentiation of Dementia From Normal Controls		.52		516
61. MRI Results and Detection of Ruptured Silicone Gel Breast Implants			.53	382
62. MRI Results and Differentiation of Dementia From Normal Controls			.57	374
63. Computed Tomography Results and Detection of Metastases From Head and Neck Cancer			.64	517
64. MMPI Validity Scales and Detection of Malingered Psychopathology (Primarily Analog Studies)		.74		11,204
65. MMPI Basic Scales: Booklet versus Computerized Form		.78		732
66. Creatinine Clearance Test Results and Kidney Function (Glomerular Filtration Rate)			.83	2,459

Note. Original citations for most table entries can be found in Meyer et al. (2001). However, source information for Entries 8, 11–14, 21, 22, 31, 32, 34–36, 38, 39, 41, 44, 48, 50, 51, and 57 can be found in Meyer and Archer (2001) and Entries 29 and 56 were obtained from Bornstein (1998c, 1999). K = number of samples; k = number of effects.

[a]The design in this research should produce results more akin to incremental validity than univariate validity.

there is no reason to believe their meta-analytic findings provide biased estimates of interrater reliability. It is also the case that Table 25.1 does not provide a comprehensive inventory of meta-analytic findings on interrater reliability. This is because the evidence has yet to be summarized for the vast majority of assessment procedures in psychology, psychiatry, and medicine. For instance, the table contains no summary data on sentence completion tests or alternative performance tasks of personality (e.g., figure drawings), and it contains no results for cognitive ability measures, global functioning scales, MRI readings, or tissue sample classifications.

The Table 25.2 findings on retest reliability also are not definitive. The Rorschach, TAT, and SCT results emerge from a relatively small pool of retest studies, none of which examined all the various scales that can be derived from these instruments. Furthermore, the evidence for these instruments is too limited to address potential moderators of stability, such as age, gender, scale distributions, engagement with the task, and so on. However, the same limitations are present for other psychological and medical measures, the vast majority of which have never been examined in a meta-analysis or perhaps ever studied in a retest design.

Finally, the validity coefficients in Table 25.3 do not encompass all the individual validity findings in the literature, or even all the topics that have been the subject of frequent validation research. In addition, the table presents information on just a limited number of instruments. For each instrument only a limited subset of hypothesized associations are typically included in the meta-analytic findings. Thus, the results in this table are far from exhaustive. These limitations exist not just for the Rorschach or TAT but for all psychological and medical tests.

At present, it seems fair to say that all assessment procedures have an incomplete evidentiary foundation relative to the diverse ways they are used in applied practice. Nevertheless, the findings in all three tables serve a valuable purpose. They emerge from systematic evidence gathered across a range of instruments and therefore provide a compelling snapshot of typical psychometric properties in the research literature.

When taken together, this compilation of meta-analyses on interrater reliability, test-retest reliability, and validity have direct implications for the recent criticisms that have been leveled against performance-based tests of personality. These criticisms often take one of two forms: (1) pointing out how some relevant questions have yet to be studied or conclusively answered and (2) scrutinizing individual studies to identify potential methodological problems. Both types of criticism can be problematic.

In any scientific endeavor, from psychology to physics, one can always identify questions that remain understudied, unanswered, or unresolved. This is true regardless of the scope and quality of the existing evidence base. Thus, when considering that which is not yet known about assessment procedures, one must also appreciate the knowledge base that does exist, as well as the way in which unanswered questions pervade all domains of science.

Second, because criticism comes easier than craftsmanship, flaws or shortcomings can always be identified for an individual study. Indeed, the last entry in Table 25.1 documents that interrater reliability for the scientific peer-review process is very poor. This finding reveals how scientists have markedly different opinions about the merits and strengths of individual research projects. What one scientist considers a sound and valuable study is regularly seen by another scientist to be a weak and defective study. Because this extensive degree of disagreement pervades all areas of science, debates about the merits or limitations of an individual study could often be endless, even when the topic is not controversial.

In light of the foregoing, it is useful to take a step back from many of the criticisms that have been directed at per-

formance tests of personality. Rather than focus on the merits of individual studies or what remains to be known, it is valuable to recognize the broader patterns that are embedded in the research literature. Indeed, it is the purpose of systematically gathered meta-analytic data to clarify these broader patterns (Hunt, 1997).

What patterns can be discerned from the 184 meta-analytic findings in Tables 25.1, 25.2, and 25.3? A logical interpretation of these data is that the Rorschach and TAT have reasonable evidence supporting their reliability and validity. They are not noticeably deficient in their psychometric properties relative to other assessment procedures commonly used in psychology, psychiatry, and medicine.

Given this, it seems necessary for the arguments surrounding these tests to shift focus. Rather than criticizing individual studies or pointing out questions that have yet to be studied or resolved, it is necessary for those who argue these tests are flawed to systematically gather evidence that illustrates how this is so relative to other commonly used tests.

In the absence of such new evidence, the available scientific foundation indicates the Rorschach and TAT have reasonable psychometric properties. As such, while recognizing the limitations of all tests, clinicians should continue to employ these instruments as sources of information to be integrated with other findings in a sophisticated and differentiated psychological assessment. In addition, researchers should continue to focus on ways to refine individual scales for these tests and strive to provide an enhanced understanding of the construct measured by each scale. Efforts should also continue documenting the ways in which distinct testing methods can produce unique information about people. Patients who seek out clinical assessments and trust that psychologists will strive to fully understand their individual qualities and difficulties deserve no less.

APPENDIX: CITATIONS CONTRIBUTING DATA TO THE LARGER META-ANALYSES OF INTERRATER RELIABILITY CONDUCTED FOR THIS CHAPTER

The Rorschach Oral Dependency (ROD) Scale

The studies contributing scale-level intraclass correlations were Bornstein, Hilsenroth, Padawer, and Fowler (2000) and Fowler, Hilsenroth, and Nolan (2000). The studies that reported scale-level correlations were Bornstein (1998a); Bornstein, Bowers, and Bonner (1996a; four samples); Bornstein, Bowers, and Bonner (1996b); Bornstein, Bowers, and Robinson (1995); Bornstein and Greenberg (1991); Bornstein, Greenberg, Leone, and Galley (1990); Bornstein, Hill, Robinson, Calabrese, and Bowers (1996); Bornstein, Leone, and Galley (1988); Bornstein, Manning, Krukonis, Rossner, and Mastrosimone (1993); Bornstein and O'Neill (1997); Bornstein, O'Neill, Galley, Leone, and Castrianno (1988; two samples); Duberstein and Talbot (1992); Fowler, Hilsenroth, and Handler (1996); Gordon and Tegtmeyer (1983); Juni, Masling, and Brannon (1979; Spearman correlation); Levin and Masling (1995; r was estimated from the range reported for six findings); Masling, Weiss, and Rothschild (1968); Ruebush and Waite (1961; an early study of the construct); Russo, Cecero, and Bornstein (2001); and Wixom, Ludolph, and Westen (1993; results for practice scoring and the actual reliability sample were averaged).

The studies that reported both scale-level correlations and response-level kappa coefficients were Bornstein (1998b; two samples); Bornstein, Bonner, Kildow, and McCall (1997; two samples); Bornstein, Galley, and Leone (1986); Bornstein and Masling (1985); Bornstein and O'Neill (2000); Bornstein, Rossner, and Hill (1994; two samples); and Bornstein, Rossner, Hill, and Stepanian (1994; three samples). The studies that contributed just response-level kappa coefficients were Greenberg and Bornstein (1989; this study also reported r but because that result appeared to overlap with data reported in another study it was not used) and Narduzzi and Jackson (2000; the reliability N was confirmed by the first author). The studies that were excluded because they appeared to use overlapping reliability samples were Bornstein, Masling, and Poynton (1987); Bornstein, Poynton, and Masling (1985); Duberstein and Talbot (1993); O'Neill and Bornstein (1990); O'Neill and Bornstein (1991); and O'Neill and Bornstein (1996).

The Defense Mechanism Manual (DMM)

For the DMM, one study contributed both Pearson and intraclass correlations (Cramer, 1998b). Pearson correlations alone were obtained from Cramer (1987; reliability N was provided by the author and averaged across scales); Cramer (1998a); Cramer (2001); Cramer, Blatt, and Ford (1988); Cramer and Block (1998); Cramer and Brilliant (2001); Cramer and Gaul (1988; using total defense scores); Hibbard et al. (1994); Hibbard and Porcerelli (1998); Hibbard et al. (2000); and Porcerelli, Thomas, Hibbard, and Cogan (1998).

Four articles were excluded from the reliability analyses. Cramer (1995) only reported intrarater reliability (mean r = .90, N = 75). Cramer (1997a) did not report the reliability N and it was not readily available from the author's files (r = .91 for total defense use). Finally, the reliability samples in Cramer (1997b, 1999) appeared to be the same as those used in Cramer and Block (1998).

The Social Cognition and Object Relations Scale (SCORS)

For the SCORS, Pearson correlations were obtained from Freedenfeld, Ornduff, and Kelsey (1995); Hibbard, Hilsenroth, Hibbard, and Nash (1995); Leigh, Westen, Barends, Mendel, and Byers (1992); Ornduff, Freedenfeld, Kelsey, and Critelli (1994); Ornduff and Kelsey (1996); Westen, Klepser, et al. (1991); Westen, Lohr, Silk, Gold, and Kerber (1990); Westen, Ludolph, Lerner, et al. (1990); and Westen, Ludolph, Silk, et al. (1990). Findings in Barends, Westen, Leigh, Silbert, and Byers (1990) were not used because of sample overlap with Leigh et al. (1992).

ICCs were obtained from Ackerman, Hilsenroth, Clemence, Weatherill, and Fowler (2000); Westen, Huebner, Lifton, Silverman, and Boekamp (1991); and Westen, Ludolph, Block, Wixom, and Wiss (1990). Results in Ackerman, Clemence, Weatherill, and Hilsenroth (1999) were not used because of sample overlap with Ackerman et al. (2000).

The Hamilton Depression Rating Scale (HDRS)

For HDRS reliability emerging from joint interviews or static sources of information (Table 25.1, Entry 7), scale-level correlations came from Akdemir et al. (2001); Bech, Bolwig, Kramp, and Rafaelsen (1979; Kendall's *W* was treated as equivalent to Spearman *r*); Bech et al. (1975; Spearman *r*); Deluty, Deluty, and Carver (1986); Faravelli, Albanesi, and Poli (1986); Hamilton (1960); Knesevich, Biggs, Clayton, and Ziegler (1977; Spearman *r*); Kobak, Reynolds, Rosenfeld, and Greist (1990); Koenig, Pappas, Holsinger, and Bachar (1995); Montgomery and Asberg (1979; two samples); Potts et al. (1990); Rapp, Smith, and Britt (1990); Rehm and O'Hara (1985; median from two samples); Reynolds and Kobak (1995); Robins (1976); Snaith, Bridge, and Hamilton (1976); Wilson et al. (1966; results from six pairs of clinicians); and Ziegler, Meyer, Rosen, and Biggs (1978; Spearman *r*). Studies that contributed item-level correlations were Bech et al. (1975; Spearman *r*), Koenig et al. (1995), Rapp et al. (1990), Rehm and O'Hara (1985; median from two samples), and Ziegler et al. (1978; Spearman *r*).

Studies that contributed scale-level kappa/ICC values for the HDRS using joint interviews or static sources of information were Baer et al. (1995; two samples); Bech et al. (1997; using data reported for the Danish University Antidepressant Group [1993]); Danish University Antidepressant Group (1990); Demitrack, Faries, Herrera, DeBrota, and Potter (1998; only results from trained raters); Endicott, Cohen, Nee, Fleiss, and Sarantakos (1981); Faravelli et al. (1986; using only weighted kappa); Foster, Sclan, Welkowitz, Boksay, and Seeland (1988; two samples); Fuglum et al. (1996); Korner et al. (1990; using the average number of raters attending each interview); Maier, Philipp, et al. (1988, *M* results for 17- and 21-item scales); Mazure, Nelson, and Price (1986); Miller, Bishop, Norman, and Maddever (1985; two samples); Nair et al. (1995); O'Hara and Rehm (1983; two samples); Potts et al. (1990); Robbins, Alessi, Cook, Poznanski, and Yanchyshyn (1982); and Zheng et al. (1988). Studies that contributed item-level kappa/ICC values using joint interviews or static sources of information were Endicott et al. (1981; *M* results for the HDRS and extracted HDRS), Mazure et al. (1986), and Miller et al. (1985; two samples).

For HDRS separate interview designs (Table 25.1, Entry 17), studies that contributed scale-level correlations were Moras, Di Nardo, and Barlow (1992); Mundt et al. (1998); and Waldron and Bates (1965). The only study that contributed item-level correlations was Mundt et al. (1998). Studies that contributed scale-level kappa/ICC coefficients were Cicchetti and Prusoff (1983; two samples); Feldman-Naim, Myers, Clark, Turner, and Leibenluft (1997); Maier, Philipp, et al. (1988; two samples, *M* for 17- and 21-item scales); Miller et al. (1985); Moberg et al. (2001; two samples); and Williams (1988). Finally, studies that contributed item-level kappa/ICC values were Cicchetti and Prusoff (1983; two samples), Moberg et al. (2001; two samples), and Williams (1988).

The Hamilton Anxiety Rating Scale (HARS)

For HARS studies using a joint interview or static information design (Table 25.1, Entry 8), scale-level correlations came from Bech, Grosby, and Rafaelsen (1984; Spearman *r*); Deluty et al. (1986); Gjerris et al. (1983; Kendall's *W* treated as equivalent to Spearman *r;* only data from experienced raters); Hamilton (1959; *M* of all pairwise reliabilities); Kobak, Reynolds, and Greist (1993; two samples); and Snaith et al. (1976). Studies that contributed scale-level kappa/ICC values were Baer et al. (1995; two samples); Bruss, Gruenberg, Goldstein, and Barber (1994); Clark and Donovan (1994); Maier, Buller, et al. (1988); and Shear et al. (2001; two samples). The studies that contributed item-level kappa/ICC values were Bruss et al. (1994); Clark and Donovan (1994); Maier, Buller, et al. (1988); and Shear et al. (2001; two samples).

For HARS studies using a separate interview design (Table 25.1, Entry 14), only Moras et al. (1992) contributed scale-level correlations. Scale-level kappa/ICC values came from Bruss et al. (1994) and Shear et al. (2001; two samples). Finally, studies that contributed item-level kappa/ICC coefficients were Bruss et al. (1994) and Shear et al. (2001).

Scientific Peer Review: Agreement Between Two Reviewers

For the reliability of peer reviewers, studies that contributed correlations for dimensional ratings of research quality were Bowen, Perloff, and Jacoby (1972; Kendall's *W* was used for *r*); Eaton (1983; from raw data); Glover and Henkelman (1994; from raw data); Hargens and Herting (1990; three samples, from raw data); Hodgson (1995; two samples); Howard and Wilkinson (1998; from raw data); Kirk and Franke (1997); Marsh and Ball (1981; footnote 1 in Marsh & Ball, 1989, indicated that *r* and the ICC were identical in this sample); Marsh and Ball (1989); Marsh and Bazeley (1999); McReynolds (1971); Munley, Sharkin, and Gelso (1988; from raw data); Petty, Fleming, and Fabrigar (1999); Rothwell and Martyn (2000; two samples, from raw data); and Whitehurst (1984; three samples, from raw data).

The studies that contributed chance-corrected reliability statistics for dimensional ratings were Cicchetti (1991; 26 samples); Eaton (1983; from raw data); Fiske and Fogg (1990); Glover and Henkelman (1994; from raw data); Goodman, Berlin, Fletcher, and Fletcher (1994); Howard and Wilkinson (1998; from raw data); Kemper, McCarthy, and Cicchetti (1996; four samples); Kirk and Franke (1997; only results from two raters); Marsh and Ball (1981); Marsh and Ball (1989; footnote 1 indicated that *r* and the ICC were identical); Marsh and Bazeley (1999); Marusic, Mestrovic, Petrovecki, and Marusic (1998; *M* of two ratings); Munley et al. (1988); Plous and Herzog (2001); Plug (1993); Rothwell and Martyn (2000; four samples, two results from raw data); Rubin, Redelmeier, Wu, and Steinberg (1993; two samples); Scharschmidt, DeAmicis, Bacchetti, and Held (1994; from raw data); Strayhorn, McDermott, and Tanguay (1993; two samples, *M* of two ratings); Whitehurst (1984; using only the sample that did not overlap with Cicchetti, 1991); and Wiener et al. (1977).

The peer-reviewer studies that contributed chance-corrected statistics for dichotomous accept versus reject determinations were Cicchetti (1991; four samples), Eaton (1983), Hargens and Herting (1990; three samples, from raw data), Strayhorn et al. (1993; two samples), Varki (1994; from raw data, estimated $N = 727$), and Whitehurst (1984; using two samples that did not overlap with Cicchetti, 1991; one result from raw data).

NOTES

1. I appreciate the helpful input of S.P. Wong, Kenneth McGraw, and Robert Rosenthal on this issue. Robert Rosenthal also suggested considering weights derived from z scores computed across samples for n and separately for judges (i.e., k, $k(k-1)/2$, or $k-1$). Because these weights would not directly quantify the number of objects rated or judgments rendered when summarizing results across studies, they were not used here.

2. Comparing reliability across disciplines is complicated because one cannot assume that the coefficients assembled here are a representative sample of findings from within these disciplines. However, if one computes mean values, they are consistent with the conclusion just stated in the text. For the Rorschach and TAT, average scale-level reliability was .85 ($n = 11$) and average item-level reliability was .84 ($n = 3$). The corresponding score and item-level averages were .67 ($n = 28$) and .58 ($n = 14$) for clinical determinations in psychology and psychiatry, .90 ($n = 5$) and .61 ($n = 12$) for medical variables, and .71 ($n = 1$) and .41 ($n = 10$) for nonclinical topics. Nonclinical topics consisted of Entries 24, 25, 36, 38, and 42 in Table 25.1. If one collapses across scales and items, the means are .85 for the Rorschach and TAT, .69 for medicine, .64 for clinical determinations in mental health, and .43 for nonclinical variables.

3. Several findings in Table 25.1 also address retest reliability. Specifically, interrater reliability studies that use a separate interview design provide evidence of short-term stability (see Table 25.1, Entries 14, 17, 29, and 30).

REFERENCES

Achenbach, T.M., McConaughy, S.H., & Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232.

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,* 422–442.

Ackerman, S.J., Hilsenroth, M.J., Clemence, A.J., Weatherill, R., & Fowler, J.C. (2000). The effects of social cognition and object representation on psychotherapy continuation. *Bulletin of the Menninger Clinic, 64,* 386–408.

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

Akdemir, A., Tuerkcapar, M.H., Orsel, S.D., Demirergi, N., Dag, I., & Ozbay, M.H. (2001). Reliability and validity of the Turkish version of the Hamilton Depression Rating Scale. *Comprehensive Psychiatry, 42,* 161–165.

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

Archer, J. (1999). Assessment of the reliability of the Conflict Tactics Scales: A meta-analytic review. *Journal of Interpersonal Violence, 14,* 1263–1289.

Arkes, H.R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323–330.

Ascioglu, S., de Pauw, B.E., Donnelly, J.P., & Collette, L. (2001). Reliability of clinical research on invasive fungal infections: A systematic review of the literature. *Medical Mycology, 39,* 35–40.

Ashton, R.H. (2000). A review and analysis of research on the test-retest reliability of professional judgment. *Journal of Behavioral Decision Making, 13,* 277–294.

Baer, L., Cukor, P., Jenike, M.A., Leahy, L., O'Laughlen, J., & Coyle, J.T. (1995). Pilot studies of telemedicine for patients with obsessive-compulsive disorder. *American Journal of Psychiatry, 152,* 1383–1385.

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

Barton, M.B., Harris, R., & Fletcher, S.W. (1999). Does this patient have breast cancer? The screening clinical breast examination: Should it be done? How? *Journal of the American Medical Association, 282,* 1270–1280.

Bech, P., Bolwig, T.G., Kramp, P., & Rafaelsen, O.J. (1979). The Bech-Rafaelsen Mania Scale and the Hamilton Depression Scale: Evaluation of homogeneity and inter-observer reliability. *Acta Psychiatrica Scandinavica, 59,* 420–430.

Bech, P., Gram, L.F., Dein, E., Jacobsen, O., Vitger, J., & Bolwig, T.G. (1975). Quantitative rating of depressive states: Correlation between clinical assessment, Beck's self-rating and Hamilton's scale. *Acta Psychiatrica Scandinavica, 51,* 161–170.

Bech, P., Grosby, H., & Rafaelsen, O.J. (1984). Generalized anxiety or depression measured by the Hamilton Anxiety Scale and the Melancholia Scale in patients before and after cardiac surgery. *Psychopathology, 17,* 253–263.

Bech, P., Stage, K.B., Nair, N.P.V., Larsen, J.K., Kragh-Sorensen, P., & Gjerris, A. (1997). The Major Depression Rating Scale (MDS): Inter-rater reliability and validity across different settings in randomized moclobemide trials. *Journal of Affective Disorders, 42,* 39–48.

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

Bornstein, R.F. (1998a). Implicit and self-attributed dependency needs in dependent and histrionic personality disorders. *Journal of Personality Assessment, 71,* 1–14.

Bornstein, R.F. (1998b). Implicit and self-attributed dependency strivings: Differential relationships to laboratory and field measures of help seeking. *Journal of Personality and Social Psychology, 75,* 778–787.

Bornstein, R.F. (1998c). Interpersonal dependency and physical illness: A meta-analytic review of retrospective and prospective studies. *Journal of Research in Personality, 32,* 480–497.

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

Bornstein, R.F., Bonner, S., Kildow, A.M., & McCall, C.A. (1997). Effects of individual versus group test administration on Rorschach Oral Dependency scores. *Journal of Personality Assessment, 69,* 215–228.

Bornstein, R.F., Bowers, K.S., & Bonner, S. (1996a). Effects of induced mood states on objective and projective dependency scores. *Journal of Personality Assessment, 67,* 324–340.

Bornstein, R.F., Bowers, K.S., & Bonner, S. (1996b). Relationships of objective and projective dependency scores to sex role orientation in college student participants. *Journal of Personality Assessment, 66,* 555–568.

Bornstein, R.F., Bowers, K.S., & Robinson, K.J. (1995). Differential relationship of objective and projective dependency scores to self-reports of interpersonal life events in college student subjects. *Journal of Personality Assessment, 65,* 255–269.

Bornstein, R.F., Galley, D.J., & Leone, D.R. (1986). Parental representations and orality. *Journal of Personality Assessment, 50,* 80–89.

Bornstein, R.F., & Greenberg, R.P. (1991). Dependency and eating disorders in female psychiatric inpatients. *Journal of Nervous and Mental Disease, 179,* 148–152.

Bornstein, R.F., Greenberg, R.P., Leone, D.R., & Galley, D.J. (1990). Defense mechanism correlates of orality. *Journal of the American Academy of Psychoanalysis, 18,* 654–666.

Bornstein, R.F., Hill, E.L., Robinson, K.J., Calabrese, C., & Bowers, K.S. (1996). Internal reliability of Rorschach Oral Dependency Scale scores. *Educational and Psychological Measurement, 56,* 130–138.

Bornstein, R.F., Hilsenroth, M.J., Padawer, J.R., & Fowler, J.C. (2000). Interpersonal dependency and personality pathology: Variations in Rorschach Oral Dependency scores across Axis II disorders. *Journal of Personality Assessment, 75,* 478–491.

Bornstein, R.F., Leone, D.R., & Galley, D.J. (1988). Rorschach measures of oral dependence and the internalized self-representation in normal college students. *Journal of Personality Assessment, 52,* 648–657.

Bornstein, R.F., Manning, K.A., Krukonis, A.B., Rossner, S.C., & Mastrosimone, C.C. (1993). Sex differences in dependency: A comparison of objective and projective measures. *Journal of Personality Assessment, 61,* 169–181.

Bornstein, R.F., & Masling, J. (1985). Orality and latency of volunteering to serve as experimental subjects: A replication. *Journal of Personality Assessment, 49,* 306–310.

Bornstein, R.F., Masling, J., & Poynton, F.G. (1987). Orality as a factor in interpersonal yielding. *Psychoanalytic Psychology, 4,* 161–170.

Bornstein, R.F., & O'Neill, R.M. (1997). Construct validity of the Rorschach Oral Dependency (ROD) Scale: Relationship of ROD

scores to WAIS-R scores in a psychiatric inpatient sample. *Journal of Clinical Psychology, 53,* 99–105.

Bornstein, R.F., & O'Neill, R.M. (2000). Dependency and suicidality in psychiatric inpatients. *Journal of Clinical Psychology, 56,* 463–473.

Bornstein, R.F., O'Neill, R.M., Galley, D.J., Leone, D.R., & Castrianno, L.M. (1988). Body image aberration and orality. *Journal of Personality Disorders, 2,* 315–322.

Bornstein, R.F., Poynton, F.G., & Masling, J. (1985). Orality and depression: An empirical study. *Psychoanalytic Psychology, 2,* 241–249.

Bornstein, R.F., Rossner, S.C., & Hill, E.L. (1994). Retest reliability of scores on objective and projective measures of dependency: Relationship to life events and intertest interval. *Journal of Personality Assessment, 62,* 398–415.

Bornstein, R.F., Rossner, S.C., Hill, E.L., & Stepanian, M.L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63,* 363–386.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry and Law, 21,* 35–76.

Bowen, D.D., Perloff, R., & Jacoby, J. (1972). Improving manuscript evaluation procedures. *American Psychologist, 27,* 221–225.

Bruss, G.S., Gruenberg, A.M., Goldstein, R.D., & Barber, J.P. (1994). Hamilton Anxiety Rating Scale Interview Guide: Joint interview and test-retest methods for interrater reliability. *Psychiatry Research, 53,* 191–202.

Callaham, M.L., Baxt, W.G., Waeckerle, J.F., & Wears, R.L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Journal of the American Medical Association, 280,* 229–231.

Capraro, M.M., Capraro, R.M., & Henson, R.K. (2001). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement, 61,* 373–386.

Cattell, R.B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12,* 289–325.

Chalmers, T.C., & Lau, J. (1994). What is meta-analysis? *Emergency Care Research Institute, 12,* 1–5.

Cicchetti, D.V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14,* 119–186.

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Cicchetti, D.V., & Prusoff, B.A. (1983). Reliability of depression and associated clinical symptoms. *Archives of General Psychiatry, 40,* 987–990.

Clark, D.B., & Donovan, J.E. (1994). Reliability and validity of the Hamilton Anxiety Rating Scale in an adolescent sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 33,* 354–360.

Conway, J.M., & Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331–360.

Conway, J.M., Jako, R.A., & Goodman, D.F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579.

Cooper, G.D., Adams, H.B., & Gibby, R.G. (1969). Ego strength changes following perceptual deprivation: Report on a pilot study. *Archives of General Psychiatry, 7,* 213–217.

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

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

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

Cramer, P. (1997b). Identity, personality and defense mechanisms: An observer-based study. *Journal of Research in Personality, 31,* 58–77.

Cramer, P. (1998a). Freshman to senior year: A follow-up study of identity, narcissism and defense mechanisms. *Journal of Research in Personality, 32,* 156–172.

Cramer, P. (1998b). Threat to gender representation: Identity and identification. *Journal of Personality, 66,* 335–357.

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

Cramer, P. (2001). Identification and its relation to identity development. *Journal of Personality, 69,* 667–688.

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). Preschool antecedents of defense mechanism use in young adults: A longitudinal study. *Journal of Personality and Social Psychology, 74,* 159–169.

Cramer, P., & Brilliant, M.A. (2001). Defense use and defense understanding in children. *Journal of Personality, 69,* 297–322.

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

Crawford, C.B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills, 49,* 223–226.

Danish University Antidepressant Group. (1990). Paroxetine: A selective serotonin reuptake inhibitor showing better tolerance, but weaker antidepressant effect than clomipramine in a controlled multicenter study. *Journal of Affective Disorders, 18,* 289–299.

Danish University Antidepressant Group. (1993). Moclobemide: A reversible MAO-A inhibitor showing weaker antidepressant ef-

fect than clomipramine in a controlled multicenter study. *Journal of Affective Disorders, 28,* 105–116.

Das, N.K., & Froehlich, L.A. (1985). Quantitative evaluation of peer review of program project and center applications in allergy and immunology. *Journal of Clinical Immunology, 5,* 220–227.

De Jonghe, B., Cook, D., Appere-De-Vecchi, C., Guyatt, G., Meade, M., & Outin, H. (2000). Using and understanding sedation scoring systems: A systematic review. *Intensive Care Medicine, 26,* 275–285.

De Kanter, R.J.A.M., Truin, G.J., Burgersdijk, R.C.W., van 'T Hof, M.A., Battistuzzi, P.G.F.C.M., Kalsbeek, H., et al. (1993). Prevalence in the Dutch adult population and a meta-analysis of signs and symptoms of temporomandibular disorder. *Journal of Dental Research, 72,* 1509–1518.

Deluty, B.M., Deluty, R.H., & Carver, C.S. (1986). Concordance between clinicians' and patients' ratings of anxiety and depression as mediated by private self-consciousness. *Journal of Personality Assessment, 50,* 93–106.

Demitrack, M.A., Faries, D., Herrera, J.M., DeBrota, D.J., & Potter, W.Z. (1998). The problem of measurement error in multisite clinical trials. *Psychopharmacology Bulletin, 34,* 19–24.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Dini, E.L., Holt, R.D., & Bedi, R. (1988). Comparison of two indices of caries patterns in 3–6 year old Brazilian children from areas with different fluoridation histories. *International Dental Journal, 48,* 378–385.

D'Olhaberriague, L., Litvan, I., Mitsias, P., & Mansbach, H.H. (1996). A reappraisal of reliability and validity studies in stroke. *Stroke, 27,* 2331–2336.

Duberstein, P.R., & Talbot, N.L. (1992). Parental idealization and the absence of Rorschach oral imagery. *Journal of Personality Assessment, 59,* 50–58.

Duberstein, P.R., & Talbot, N.L. (1993). Rorschach oral imagery, attachment style, and interpersonal relatedness. *Journal of Personality Assessment, 61,* 294–310.

Duhig, A.M., Renk, K., Epstein, M.K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice, 7,* 435–453.

Dworkin, S.F., LeResche, L., DeRouen, T., & Von Korff, M. (1990). Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. *Journal of Prosthetic Dentistry, 63,* 574–579.

Eaton, W.O. (1983). Reliability of ethics reviews: Some initial empirical findings. *Canadian Psychology, 24,* 14–18.

Edinger, J.D., & Weiss, W.U. (1974). The relation between the altitude quotient and adjustment potential. *Journal of Clinical Psychology, 30,* 510–513.

Endicott, J., Cohen, J., Nee, J., Fleiss, J., & Sarantakos, S. (1981). Hamilton Depression Rating Scale: Extracted from regular and change versions of the Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 38,* 98–103.

Endicott, N.A., & Endicott, J. (1963). "Improvement" in untreated psychiatric patients. *Archives of General Psychiatry, 9,* 575–585.

Etchells, E., Bell, C., & Robb, K. (1997). Does this patient have an abnormal systolic murmur? *Journal of the American Medical Association, 277,* 564–571.

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

Exner, J.E., Jr. (2001). A comment on "The misperception of psychopathology: Problems with norms of the Comprehensive System for the Rorschach." *Clinical Psychology: Science and Practice, 8,* 386–396.

Faravelli, C., Albanesi, G., & Poli, E. (1986). Assessment of depression: A comparison of rating scales. *Journal of Affective Disorders, 11,* 245–253.

Fehlings, M.G., Rao, S.C., Tator, C.H., Skaf, G., Arnold, P., Benzel, E., et al. (1999). The optimal radiologic method for assessing spinal canal compromise and cord compression in patients with cervical spinal cord injury. Part II: Results of a multicenter study. *Spine, 24,* 605–613.

Feldman-Naim, S., Myers, F.S., Clark, C.H., Turner, E.H., & Leibenluft, E. (1997). Agreement between face-to-face and telephone-administered mood ratings in patients with rapid cycling bipolar disorder. *Psychiatry Research, 71,* 129–132.

Feurer, I.D., Becker, G.J., Picus, D., Ramirez, E., Darcy, M.D., & Hicks, M.E. (1994). Evaluating peer reviews: Pilot testing of a grading instrument. *Journal of the American Medical Association, 272,* 98–100.

Fiske, D.W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist, 45,* 591–598.

Foster, J.R., Sclan, S., Welkowitz, J., Boksay, I., & Seeland, I. (1988). Psychiatric assessment in medical long-term care facilities: Reliability of commonly used rating scales. *International Journal of Geriatric Psychiatry, 3,* 229–233.

Fowler, J.C., Hilsenroth, M.J., & Handler, L. (1996). A multimethod approach to assessing dependency: The early memory dependency probe. *Journal of Personality Assessment, 67,* 399–413.

Fowler, J.C., Hilsenroth, M.J., & Nolan, E. (2000). Exploring the inner world of self-mutilating borderline patients: A Rorschach investigation. *Bulletin of the Menninger Clinic, 64,* 365–385.

Freedenfeld, R.N., Ornduff, S.R., & Kelsey, R.M. (1995). Object relations and physical abuse: A TAT analysis. *Journal of Personality Assessment, 64,* 552–568.

Fuglum, E., Rosenberg, C., Damsbo, N., Stage, K., Lauritzen, L., & Bech, P. (1996). Screening and treating depressed patients. A comparison of two controlled citalopram trials across treatment settings: Hospitalized patients vs. patients treated by their family doctors. Danish University Antidepressant Group. *Acta Psychiatrica Scandinavica, 94,* 18–25.

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

Garb, H.N. (1999). Call for a moratorium on the use of the Rorschach Inkblot test in clinical and forensic settings. *Assessment, 6,* 313–317.

Garb, H.N., & Schramke, C.J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin, 120,* 140–153.

Garfield, S.L. (2000). The Rorschach test in clinical diagnosis: A brief commentary. *Journal of Clinical Psychology, 56,* 431–434.

Gibson, G., & Ottenbacher, K.J. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *Journal of Applied Behavioral Science, 24,* 298–314.

Gjerris, A., Bech, P., Bojholm, S., Bolwig, T.G., Kramp, P., Clemmesen, L., et al. (1983). The Hamilton Anxiety Scale: Evaluation of homogeneity and inter-observer reliability in patients with depressive disorders. *Journal of Affective Disorders, 5,* 163–170.

Glover, G.H., & Henkelman, R.M. (1994). Abstract scoring for the annual SMR program: Significance of reviewer score normalization. *Magnetic Resonance Medicine, 32,* 435–439.

Goldfried, M.P., Stricker, G., & Weiner, I.B. (1971). *Rorschach handbook of clinical and research applications.* Englewood Cliffs, NJ: Prentice-Hall.

Goldman, R.L. (1994). The reliability of peer assessments: A meta-analysis. *Evaluation and the Health Professions, 17,* 3–21.

Goode, E. (2001, February 20). What's in an Inkblot? Some say, not much. *The New York Times,* pp. D1, D4.

Goodman, S.N., Berlin, J.A., Fletcher, S.W., & Fletcher, R.H. (1994). Manuscript quality before and after peer review and editing at *Annals of Internal Medicine. Annals of Internal Medicine, 121,* 11–21.

Gordon, M., & Tegtmeyer, P.F. (1983). Oral-dependent content in children's Rorschach protocols. *Perceptual and Motor Skills, 57,* 1163–1168.

Gosling, S.D. (2001). From mice to men: What can we learn about personality from animal research? *Psychological Bulletin, 127,* 45–86.

Gottfredson, S.D. (1978). Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist, 33,* 920–934.

Gottlieb, G.L., Gur, R.E., & Gur, R.C. (1988). Reliability of psychiatric scales in patients with dementia of the Alzheimer type. *American Journal of Psychiatry, 145,* 857–860.

Greenberg, R.P., & Bornstein, R.F. (1989). Length of psychiatric hospitalization and oral dependency. *Journal of Personality Disorders, 3,* 199–204.

Gross, C.R., Shinar, D., Mohr, J.P., Hier, D.B., Caplan, L.R., Price, T.R., et al. (1986). Interobserver agreement in the diagnosis of stroke type. *Archives of Neurology, 43,* 893–898.

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.

Grueneich, R. (1992). The borderline personality disorder diagnosis: Reliability, diagnostic efficiency, and covariation with other personality disorder diagnoses. *Journal of Personality Disorders, 6,* 197–212.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32,* 50–55.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Hammond, K.R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York: Oxford University Press.

Harbst, K.B., Ottenbacher, K.J., & Harris, S.R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71,* 107–115.

Hargens, L.L., & Herting, J.R. (1990). A new approach to referees' assessments of manuscripts. *Social Science Research, 19,* 1–16.

Hathaway, A.P. (1982). Intelligence and non-intelligence factors contributing to scores on the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment, 46,* 8–11.

Hedlund, J.L., & Vieweg, B.W. (1979). The Hamilton Rating Scale for Depression: A comprehensive review. *Journal of Operational Psychiatry, 10,* 149–165.

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., & Porcerelli, J. (1998). Further validation of the Cramer Defense Mechanism Manual. *Journal of Personality Assessment, 70,* 460–483.

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.

Hilsenroth, M.J., Fowler, J.C., Padawer, J.R., & Handler, L. (1997). Narcissism in the Rorschach revisited: Some reflections on empirical data. *Psychological Assessment, 9,* 113–121.

Hodgson, C. (1995). Evaluation of cardiovascular grant-in-aid applications by peer review: Influence of internal and external reviewers and committees. *Canadian Journal of Cardiology, 11,* 864–868.

Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology, 50,* 1189–1195.

Holden, G.W., & Miller, P.C. (1999). Enduring and different: A meta-analysis of the similarity in parents' child rearing. *Psychological Bulletin, 125,* 223–254.

Holland, J.L., Johnston, J.A., & Asama, N.F. (1993). The Vocational Identity Scale: A diagnostic and treatment tool. *Journal of Career Assessment, 1,* 1–12.

Howard, L., & Wilkinson, G. (1998). Peer review and editorial decision-making. *British Journal of Psychiatry, 173,* 110–113.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Ismail, A.I., & Sohn, W. (1999). A systematic review of clinical diagnostic criteria of early childhood caries. *Journal of Public Health Dentistry, 59,* 171–191.

Jones, D.B., Schlife, C.M., & Phipps, K.R. (1992). An oral health survey of Head Start children in Alaska: Oral health status, treatment needs, and cost of treatment. *Journal of Public Health Dentistry, 52,* 86–93.

Juni, S., Masling, J., & Brannon, R. (1979). Interpersonal touching and orality. *Journal of Personality Assessment, 43,* 235–237.

Justice, A.C., Berlin, J.A., Fletcher, S.W., Fletcher, R.H., & Goodman, S.N. (1994). Do readers and peer reviewers agree on manuscript quality? *Journal of the American Medical Association, 272,* 117–119.

Katz, L., Ripa, L.W., & Petersen, M. (1992). Nursing caries in Head Start children, St. Thomas, U.S. Virgin Islands: Assessed by examiners with different dental backgrounds. *Journal of Clinical Pediatric Dentistry, 16,* 124–128.

Kemper, K.J., McCarthy, P.L., & Cicchetti, D.V. (1996). Improving participation and interrater agreement in scoring Ambulatory Pediatric Association abstracts: How well have we succeeded? *Archives of Pediatrics and Adolescent Medicine, 150,* 380–383.

Kirk, S.A., & Franke, T.M. (1997). Agreeing to disagree: A study of the reliability of manuscript reviews. *Social Work Research, 21,* 121–126.

Knesevich, J.W., Biggs, J.T., Clayton, P.J., & Ziegler, V.E. (1977). Validity of the Hamilton Rating Scale for Depression. *British Journal of Psychiatry, 131,* 49–52.

Kobak, K.A., Reynolds, W.M., & Greist, J.H. (1993). Development and validation of a computer-administered version of the Hamilton Rating Scale. *Psychological Assessment, 5,* 487–492.

Kobak, K.A., Reynolds, W.M., Rosenfeld, R., & Greist, J.H. (1990). Development and validation of a computer-administered version of the Hamilton Depression Rating Scale. *Psychological Assessment, 2,* 56–63.

Koenig, H.G., Pappas, P., Holsinger, T., & Bachar, J.R. (1995). Assessing diagnostic approaches to depression in medically ill older adults: How reliably can mental health professionals make judgments about the cause of symptoms? *Journal of the American Geriatrics Society, 43,* 472–478.

Korner, A., Nielsen, B.M., Eschen, F., Moller-Madsen, S., Stender, A., Christensen, E.M., et al. (1990). Quantifying depressive symptomatology: Inter-rater reliability and inter-item correlations. *Journal of Affective Disorders, 20,* 143–149.

Lathrop, R.G., & Williams, J.E. (1987). The reliability of inverse scree tests for cluster analysis. *Educational and Psychological Measurement, 47,* 953–959.

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.

Levin, R., & Masling, J. (1995). Relations of oral imagery to thought disorder in subject with frequent nightmares. *Perceptual and Motor Skills, 80,* 1115–1120.

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.

Lilienfeld, S.O., Wood, J.M., & Garb, H.N. (2001, May). What's wrong with this picture? *Scientific American,* 80–87.

Livesley, J. (2000a). Introduction: Critical issues in the classification of personality disorder Part I. *Journal of Personality Disorders, 14,* 1–2.

Livesley, J. (2000b). Introduction: Critical issues in the classification of personality disorder Part II. *Journal of Personality Disorders, 14,* 97–98.

Loranger, A.W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S.M., et al. (1994). The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration international pilot study of personality disorders. *Archives of General Psychiatry, 51,* 215–224.

Maier, W., Buller, R., Philipp, M., & Heuser, I. (1988). The Hamilton Anxiety Scale: Reliability, validity and sensitivity to change in anxiety and depressive disorders. *Journal of Affective Disorders, 14,* 61–68.

Maier, W., Philipp, M., Heuser, I., Schlegel, S., Buller, R., & Wetzel, H. (1988). Improving depression severity assessment: I. Reliability, internal validity and sensitivity to change of three observer depression scales. *Journal of Psychiatric Research, 22,* 3–12.

Marino, R.J., & Onetto, J.E. (1995). Caries experience in urban and rural Chilean 3-year-olds. *Community Dentistry and Oral Epidemiology, 23,* 60–61.

Marsh, H.W., & Ball, S. (1981). Interjudgmental reliability of reviews for the *Journal of Educational Psychology. Journal of Educational Psychology, 73,* 872–880.

Marsh, H.W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education, 57,* 151–169.

Marsh, H.W., & Bazeley, P. (1999). Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analy-

sis evaluations of reliability, validity, and structure. *Multivariate Behavioral Research, 34,* 1–30.

Martin, D.J., Garske, J.P., & Davis, M.K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68,* 438–450.

Marusic, A., Mestrovic, T., Petrovecki, M., & Marusic, M. (1998). Peer review in the *Croatian Medical Journal* from 1992–1996. *Croatian Medical Journal, 39,* 3–9.

Masling, J., Weiss, L., & Rothschild, B. (1968). Relationships of oral imagery to yielding behavior and birth order. *Journal of Consulting and Clinical Psychology, 32,* 89–91.

Matsuura, P., Waters, R.L., Adkins, R.H., Rothman, S., Gurbani, N., & Sie, I. (1989). Comparison of computerized tomography parameters of the cervical spin in normal control subjects and spinal cord-injured patients. *Journal of Bone and Joint Surgery, 71-A,* 183–188.

Mazure, C., Nelson, J.C., & Price, L.H. (1986). Reliability and validity of the symptoms of major depressive illness. *Archives of General Psychiatry, 43,* 451–456.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

McKelvie, S.J. (1995). The VVIQ as a psychometric test of individual differences in visual imagery vividness: A critical quantitative review and plea for direction. *Journal of Mental Imagery, 19,* 1–106.

McReynolds, P. (1971). Reliability of ratings of research papers. *American Psychologist, 26,* 400–401.

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. (Ed.). (1999). Special Section I: The utility of the Rorschach for clinical assessment. *Psychological Assessment, 11,* 235–302.

Meyer, G.J. (2000). 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. (2001a). Evidence to correct misperceptions about Rorschach norms. *Clinical Psychology: Science and Practice, 8,* 389–396.

Meyer, G.J. (Ed.). (2001b). Special Section II: The utility of the Rorschach for clinical assessment II. *Psychological Assessment, 13,* 419–502.

Meyer, G.J. (2002). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment, 78,* 104–129.

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., Baxter, D., Exner, J.E., Jr., Fowler, J.C., Hilsenroth, M.J., Piers, C.C., et al. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78,* 219–274.

Meyer, G.J., Finn, S.E., Eyde, L., Kay, G.G., Moreland, K.L., Dies, R.R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

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.

Miller, I.W., Bishop, S.B., Norman, W.H., & Maddever, H. (1985). The Modified Hamilton Rating Scale for Depression: Reliability and validity. *Psychiatry Research, 14,* 131–142.

Moberg, P.J., Lazarus, L.W., Mesholam, R.I., Bilker, W., Chuy, I.L., Neyman, I., et al. (2001). Comparison of the standard and structured interview guide for the Hamilton Depression Rating Scale in depressed geriatric inpatients. *American Journal of Geriatric Psychiatry, 9,* 35–40.

Montgomery, S.A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry, 134,* 382–389.

Moras, K., Di Nardo, P.A., & Barlow, D.H. (1992). Distinguishing anxiety and depression: Reexamination of the reconstructed Hamilton scales. *Psychological Assessment, 4,* 224–227.

Mundt, J.C., Kobak, K.A., Taylor, L.V., Mantle, J.M., Jefferson, J.W., Katzelnick, D.J., et al. (1998). Administration of the Hamilton Depression Rating Scale using interactive voice response technology. *MD Computing, 15,* 31–39.

Munley, P.H., Sharkin, B.S., & Gelso, C.J. (1988). Reviewer ratings and agreement on manuscripts reviewed for the *Journal of Counseling Psychology. Journal of Counseling Psychology, 35,* 198–202.

Myers, K.A., & Farquhar, D.R. (2001). Does this patient have clubbing? *Journal of the American Medical Association, 286,* 341–347.

Nair, N.P.V., Amin, M., Holm, P., Katona, C., Klitgaard, N., Ng-Ying-Kin, N.M.K., et al. (1995). Moclobemide and nortriptyline in elderly depressed patients: A randomized, multicenter trial against placebo. *Journal of Affective Disorders, 33,* 1–9.

Narduzzi, K.J., & Jackson, T. (2000). Personality differences between eating-disordered women and a nonclinical comparison sample: A discriminant classification analysis. *Journal of Clinical Psychology, 56,* 699–710.

Newmark, C.S., Finkelstein, M., & Frerking, R.A. (1974). Comparison of the predictive validity of two measures of psychotherapy prognosis. *Journal of Personality Assessment, 38,* 144–148.

Newmark, C.S., Hetzel, W., Walker, L., Holstein, S., & Finkelstein, M. (1973). Predictive validity of the Rorschach Prognostic Rating Scale with behavior modification techniques. *Journal of Clinical Psychology, 29,* 246–248.

Newmark, C.S., Konanc, J.T., Simpson, M., Boren, R.B., & Prillaman, K. (1979). Predictive validity of the Rorschach Prog-

nostic Rating Scale with schizophrenic patients. *Journal of Nervous and Mental Disease, 167,* 135–143.

Nwosu, C.R., Khan, K.S., Chien, P.F.W., & Honest, M.R. (1998). Is real-time ultrasonic bladder volume estimation reliable and valid? A systematic overview. *Scandinavian Journal of Urology and Nephrology, 32,* 325–330.

O'Hara, M.W., & Rehm, L.P. (1983). Hamilton Rating Scale for Depression: Reliability and validity of judgments of novice raters. *Journal of Consulting and Clinical Psychology, 51,* 318–319.

Okada, Y., Ikata, T., & Katoh, S. (1994). Morphologic analysis of the cervical spinal cord, dural tube, and spinal canal by magnetic resonance imaging in normal adults and patients with cervical spondylotic myelopathy. *Spine, 19,* 2331–2335.

O'Neill, R.M., & Bornstein, R.F. (1990). Oral-dependence and gender: Factors in help-seeking response set and self-reported psychopathology in psychiatric inpatients. *Journal of Personality Assessment, 55,* 28–40.

O'Neill, R.M., & Bornstein, R.F. (1991). Orality and depression in psychiatric inpatients. *Journal of Personality Disorders, 5,* 1–7.

O'Neill, R.M., & Bornstein, R.F. (1996). Dependency and alexithymia in psychiatric inpatients. *Journal of Nervous and Mental Disease, 184,* 302–306.

Ornduff, S.R., Freedenfeld, R.N., Kelsey, R.M., & Critelli, J.W. (1994). Object relations of sexually abused female subjects: A TAT analysis. *Journal of Personality Assessment, 63,* 223–238.

Ornduff, 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.

O'Sullivan, D.M., & Tinanoff, N. (1996). The association of early dental caries patterns with caries incidence in preschool children. *Journal of Public Health Dentistry, 56,* 81–83.

Ottenbacher, K.J. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy, 40,* 464–469.

Ottenbacher, K.J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal on Mental Retardation, 98,* 135–142.

Ottenbacher, K.J., & Cusik, A. (1991). An empirical investigation of interrater agreement for single-subject data using graphs with and without trend lines. *Journal of the Association for Persons with Severe Handicaps, 16,* 48–55.

Ottenbacher, K.J., Hsu, Y., Granger, C.V., & Fiedler, R.C. (1996). The reliability of the Functional Independence Measure: A quantitative review. *Archives of Physical Medicine and Rehabilitation, 77,* 1226–1232.

Park, H.-S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *Journal of Experimental Education, 58,* 311–320.

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.

Paunio, P., Rautava, P., Helenius, H., Alanen, P., & Sillanpää, M. (1993). The Finnish Family Competence Study: The relationship between caries, dental health habits and general health in 3-year-old Finnish children. *Caries Research, 27,* 154–160.

Petty, R.E., Fleming, M.A., & Fabrigar, L.R. (1999). The review process at PSPB: Correlates of interreviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin, 25,* 188–203.

Plous, S., & Herzog, H. (2001). Reliability of protocol reviews for animal research. *Science, 293,* 608–609.

Plug, C. (1993). The reliability of manuscript evaluation for the *South African Journal of Psychology. South African Journal of Psychology, 23,* 43–48.

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.

Potts, M.K., Daniels, M., Burnam, M.A., & Wells, K.B. (1990). A structured interview version of the Hamilton Depression Rating Scale: Evidence of reliability and versatility of administration. *Journal of Psychiatric Research, 24,* 335–350.

Ramsay, I.N., Mathers, A.M., Hood, V.D., & Torbet, T.E. (1991). Ultrasonic assessment of residual bladder volume: A quick, simple and accurate bedside assessment. *Obstetrics and Gynecology Today, 2,* 68–70.

Rao, S.C., & Fehlings, M.G. (1999). The optimal radiologic method for assessing spinal canal compromise and cord compression in patients with cervical spinal cord injury. Part I: An evidence-based analysis of the published literature. *Spine, 24,* 598–604.

Rapp, S.R., Smith, S.S., & Britt, M. (1990). Identifying comorbid depression in elderly medical patients: Use of the Extracted Hamilton Depression Rating Scale. *Psychological Assessment, 2,* 243–247.

Rehm, L.P., & O'Hara, M.W. (1985). Item characteristics of the Hamilton Rating Scale for Depression. *Journal of Psychiatric Research, 19,* 31–41.

Reynolds, W.M., & Kobak, K.A. (1995). Reliability and validity of the Hamilton Depression Inventory: A paper-and-pencil version of the Hamilton Depression Rating Scale clinical interview. *Psychological Assessment, 7,* 472–483.

Ritzler, B., Erard, R., & Pettigrew, G. (2002). Protecting the integrity of Rorschach expert witnesses: A reply to Grove and Barden (1999) re: the admissibility of testimony under Daubert/Kumho analyses. *Psychology, Public Policy, and Law, 8,* 201–215.

Robbins, D.R., Alessi, N.E., Cook, S.C., Poznanski, E.O., & Yanchyshyn, G.W. (1982). The use of the Research Diagnostic Criteria (RDC) for depression in adolescent psychiatric inpatients. *Journal of the American Academy of Child Psychiatry, 21,* 251–255.

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.

Robins, A.H. (1976). Depression in patients with Parkinsonism. *British Journal of Psychiatry, 128,* 141–145.

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., et al. (1996). Personal Problem-Solving Scoring of TAT responses: Known-groups validation. *Journal of Personality Assessment, 67,* 641–653.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Rothstein, H.R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Rothwell, P.M., & Martyn, C.N. (2000). Reproducibility of peer review in clinical neuroscience. Is agreement between reviewers any greater than would be expected by chance alone? *Brain, 123 (Pt 9),* 1964–1969.

Rubin, H.R., Redelmeier, D.A., Wu, A.W., & Steinberg, E.P. (1993). How reliable is peer review of scientific abstracts? Looking back at the 1991 Annual Meeting of the Society of General Internal Medicine. *Journal of General Internal Medicine, 8,* 255–258.

Ruebush, B.K., & Waite, R.R. (1961). Oral dependency in anxious and defensive children. *Merrill-Palmer Quarterly, 7,* 181–190.

Russell, A.S., Thorn, B.D., & Grace, M. (1983). Peer review: A simplified approach. *Journal of Rheumatology, 10,* 479–481.

Russo, P.M., Cecero, J.J., & Bornstein, R.F. (2001). Implicit and self-attributed dependency needs in homeless men and women. *Journal of Social Distress and the Homeless, 10,* 269–277.

Salgado, J.F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills, 83,* 1195–1201.

Scharschmidt, B.F., DeAmicis, A., Bacchetti, P., & Held, M.J. (1994). Chance, concurrence, and clustering: Analysis of reviewers' recommendations on 1,000 submissions to *The Journal of Clinical Investigation. Journal of Clinical Investigation, 93,* 1877–1880.

Schuerger, J.M., & Witt, A.C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45,* 294–302.

Scullen, S.E. (1997). When ratings from one source have been averaged, but ratings from another source have not: Problems and solutions. *Journal of Applied Psychology, 82,* 880–888.

Segal, D.L., Hersen, M., & Van Hasselt, V.B. (1994). Reliability of the Structured Clinical Interview for *DSM-III-R:* An evaluative review. *Comprehensive Psychiatry, 35,* 316–327.

Shear, M.K., Vander Bilt, J., Rucci, P., Endicott, J., Lydiard, B., Otto, M.W., et al. (2001). Reliability and validity of a structured interview guide for the Hamilton Anxiety Rating Scale (SIGH-A). *Depression and Anxiety, 13,* 166–178.

Shulman, K.I. (2000). Clock-drawing: Is it the ideal cognitive screening test? *International Journal of Geriatric Psychiatry, 15,* 548–561.

Snaith, R.P., Bridge, G.W.K., & Hamilton, M. (1976). The Leeds Scales for the Self-Assessment of Anxiety and Depression. *British Journal of Psychiatry, 128,* 156–163.

Spengler, P.M., Strohmer, D.C., Dixon, D.N., & Shivy, V.A. (1995). A scientist-practitioner model of psychological assessment: Implications for training, practice and research. *The Counseling Psychologist, 23,* 506–534.

Steinberg, M. (2000). Advances in the clinical assessment of dissociation: The SCID-D-R. *Bulletin of the Menninger Clinic, 64,* 146–163.

Strayhorn, J., McDermott, J.F., & Tanguay, P.E. (1993). An intervention to improve the reliability of manuscript reviews for the *Journal of the American Academy of Child and Adolescent Psychiatry. American Journal of Psychiatry, 150,* 947–952.

Streiner, D.L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports, 83,* 687–694.

Stukenberg, K.W., Dura, J.R., & Kielcolt-Glaser, J.K. (1990). Depression screening scale validation in an elderly, community dwelling population. *Psychological Assessment, 2,* 134–138.

Swain, A., & Suls, J. (1996). Reproducibility of blood pressure and heart rate reactivity: A meta-analysis. *Psychophysiology, 33,* 162–174.

Szatmari, P. (2000). The classification of autism, Asperger's syndrome, and pervasive developmental disorder. *Canadian Journal of Psychiatry, 45,* 731–738.

van Ijzendoorn, M.H., Schuengel, C., & Bakermans-Kranenburg, M.J. (1999). Disorganized attachment in early childhood: Meta-analysis of precursors, concomitants, and sequelae. *Development and Psychopathology, 11,* 225–249.

van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology, 52,* 625–629.

van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: A randomised trial. *British Medical Journal, 318,* 23–27.

Varki, A.P. (1994). The screening review system: Fair or foul? *Journal of Clinical Investigation, 93,* 1871–1874.

Viswesvaran, C., & Ones, D.S. (2000). Measurement error in "Big Five factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60,* 224–235.

Viswesvaran, C., Ones, D.S., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Waldron, J., & Bates, T.J.N. (1965). The management of depression in hospital: A comparative trial of desipramine and imipramine. *British Journal of Psychiatry, 111,* 511–516.

Walsh, E., Rooney, M., Appleby, L., & Wilkinson, G. (2000). Open peer review: A randomised controlled trial. *British Journal of Psychiatry, 176,* 47–51.

Weathers, F.W., Keane, T.M., & Davidson, J.R. (2001). Clinician-administered PTSD scale: A review of the first ten years of research. *Depression and Anxiety, 13,* 132–156.

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

Westen, D., Huebner, D., Lifton, N., Silverman, M., & Boekamp, J. (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., 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., 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, F.C. (1990). Object relations in borderline adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 29,* 338–348.

Westen, D., Ludolph, P., Silk, K.R., Kellam, A., Gold, L., & Lohr, N. (1990). Object relations in borderline adolescents and adults: Developmental differences. *Adolescent Psychiatry, 17,* 360–384.

Whisman, M.A., Strosahl, K., Fruzzetti, A.E., Schmaling, K.B., Jacobson, N.S., & Miller, D.M. (1989). A structured interview version of the Hamilton Rating Scale for Depression: Reliability and validity. *Psychological Assessment, 1,* 238–241.

Whitehurst, G.J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39,* 22–28.

Wiener, S.L., Urivetzky, M., Bregman, D., Cohen, J., Eich, R., Gootman, N., et al. (1977). Peer review: Inter-reviewer agreement during evaluation of research grant applications. *Clinical Research, 25,* 306–311.

Williams, G.J., Monder, R., & Rychlak, J.F. (1967). A one-year concurrent validity study of the Rorschach Prognostic Rating Scale. *Journal of Projective Techniques and Personality Assessment, 31,* 30–33.

Williams, J.B. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry, 45,* 742–747.

Wilson, I.C., Rabon, A.M., Merrick, H.A., Knox, A.E., Taylor, J.P., & Buffaloe, W.J. (1966). Imipramine pamoate in the treatment of depression. *Psychosomatics, 7,* 251–253.

Wixom, J., Ludolph, P., & Westen, D. (1993). The quality of depression in adolescents with borderline personality disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 32,* 1172–1177.

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., Garb, H.N., & Lilienfeld, S.O. (2001). The misperception of psychopathology: Problems with the norms of the Comprehensive System for the Rorschach. *Clinical Psychology: Science and Practice, 8,* 350–373.

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

Yagot, K., Nazhat, N.Y., & Kuder, S.A. (1990). Prolonged nursing-habit caries index. *Journal of the International Association of Dentistry for Children, 20,* 8–10.

Yarnold, P.R., & Mueser, K.T. (1989). Meta-analyses of the reliability of Type A behaviour measures. *British Journal of Medical Psychology, 62,* 43–50.

Zheng, Y., Zhao, J., Phillips, M., Liu, J., Cai, M., Sun, S., et al. (1988). Validity and reliability of the Chinese Hamilton Depression Rating Scale. *British Journal of Psychiatry, 152,* 660–664.

Ziegler, V.E., Meyer, D.A., Rosen, S.H., & Biggs, J.T. (1978). Reliability of video taped Hamilton ratings. *Biological Psychiatry, 13,* 119–122.

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry, 51,* 225–245.