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-
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 pre-existing 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 test-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 INTRERRATER 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 weight 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 weighting 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-analysis, 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

<table>
<thead>
<tr>
<th>Target reliability construct</th>
<th>Number of independent pairs of judgments</th>
<th>$n(k - 1)$</th>
<th>$r$</th>
<th>$\kappa$/ ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bladder Volume by Real-Time Ultrasound</td>
<td>40</td>
<td></td>
<td>.95</td>
<td>.92</td>
</tr>
<tr>
<td>2. Count of Decayed, Filled, or Missing Teeth (or Surfaces) in Early Childhood</td>
<td>237</td>
<td>113</td>
<td>.97</td>
<td>.79</td>
</tr>
<tr>
<td>3. Rorschach Oral Dependence Scale Scoring</td>
<td>934</td>
<td>40</td>
<td>.91</td>
<td>.91</td>
</tr>
<tr>
<td>4. Measured Size of the Spinal Canal and Spinal Cord on MRI, CT, or X-Ray</td>
<td>6,430</td>
<td>200</td>
<td>.90</td>
<td>.84</td>
</tr>
<tr>
<td>5. Scoring the Rorschach Comprehensive System*:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary scores</td>
<td>219</td>
<td>565</td>
<td>.90</td>
<td>.88</td>
</tr>
<tr>
<td>Response segments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scores for each response</td>
<td>11,518</td>
<td>11,572</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Neuropsychologists' Test-Based Judgments of Cognitive Impairment</td>
<td>901</td>
<td></td>
<td>.86</td>
<td>.83</td>
</tr>
<tr>
<td>7. Hamilton Depression Rating Scale Scoring From Joint Interviewsb</td>
<td>1,773</td>
<td>334</td>
<td>.93</td>
<td>.80</td>
</tr>
<tr>
<td>8. Hamilton Anxiety Rating Scale Scoring From Joint Interviewsb</td>
<td>1,365</td>
<td>1,365</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Level of Drug Sedation by ICU Physicians or Nurses</td>
<td></td>
<td>161</td>
<td>.91</td>
<td>.77</td>
</tr>
<tr>
<td>10. Functional Independence Measure</td>
<td></td>
<td>165</td>
<td>.87</td>
<td>.72</td>
</tr>
<tr>
<td>11. TAT Personal Problem-Solving Scale Scoring</td>
<td></td>
<td></td>
<td>.86</td>
<td>.62</td>
</tr>
<tr>
<td>12. Borderline Personality Disorder:</td>
<td></td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>402</td>
<td></td>
<td>.83</td>
<td></td>
</tr>
<tr>
<td>Specific symptoms</td>
<td>198</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Signs and Symptoms of Temporomandibular Disorder</td>
<td>192</td>
<td></td>
<td>.86</td>
<td>.64</td>
</tr>
<tr>
<td>14. Hamilton Anxiety Rating Scale Scoring From Separate Interviews</td>
<td>60</td>
<td>562</td>
<td>.65</td>
<td>.56</td>
</tr>
<tr>
<td>15. Axis I Psychiatric Diagnosis by Semi-Structured SCID in Joint Interviews</td>
<td>208</td>
<td></td>
<td>.79</td>
<td>.58</td>
</tr>
<tr>
<td>16. Axis II Psychiatric Diagnosis by Semi-Structured Joint Interviews</td>
<td>208</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Hamilton Depression Rating Scale Scoring From Separate Interviews</td>
<td>216</td>
<td></td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>18. TAT Social Cognition and Object Relations Scale Scoring</td>
<td></td>
<td></td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>19. Rorschach Prognostic Rating Scale Scoring</td>
<td></td>
<td></td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>20. TAT Defense Mechanism Manual Scoring</td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>21. Therapist or Observer Ratings of Therapeutic Alliance in Treatment</td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>22. Type A Behavior Pattern by Structured Interview</td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>23. Job Selection Interview Ratings</td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>24. Personality or Temperament of Mammals</td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>25. Editors' Ratings of the Quality of Manuscript Reviews or Reviewers</td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>26. Visual Analysis of Plotted Behavior Change in Single-Case Research</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. Presence of Clubbing in Fingers or Toes</td>
<td></td>
<td></td>
<td>.54</td>
<td>.55</td>
</tr>
<tr>
<td>28. Stroke Classification by Neurologists</td>
<td></td>
<td></td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>29. Axis I Psychiatric Diagnosis by Semi-Structured SCID in Separate Interviews</td>
<td></td>
<td></td>
<td>.54</td>
<td>.55</td>
</tr>
</tbody>
</table>
TABLE 25.1 (Continued)

<table>
<thead>
<tr>
<th>Target reliability construct</th>
<th>n(k - 1) = Number of independent pairs of judgments</th>
<th>r</th>
<th>( \kappa / \text{ICC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>30. Axis II Psychiatric Diagnosis by Semi-Structured Separate Interviews</td>
<td>358</td>
<td>( \text{Scale Item} )</td>
<td>( \text{Item} )</td>
</tr>
<tr>
<td>31. Child or Adolescent Problems:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher ratings</td>
<td>2,100</td>
<td>.64</td>
<td></td>
</tr>
<tr>
<td>Parent ratings</td>
<td>4,666</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>Externalizing problems</td>
<td>7,710</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Internalizing problems</td>
<td>5,178</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>Direct observers</td>
<td>231</td>
<td>.57</td>
<td></td>
</tr>
<tr>
<td>Clinicians</td>
<td>729</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>32. Job Performance Ratings by Supervisors</td>
<td>1,803</td>
<td>.57</td>
<td>.48</td>
</tr>
<tr>
<td>33. Self and Partner Ratings of Conflict:</td>
<td>10,119</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men’s aggression</td>
<td>616</td>
<td>.55</td>
<td></td>
</tr>
<tr>
<td>Women’s aggression</td>
<td>616</td>
<td>.51</td>
<td></td>
</tr>
<tr>
<td>34. Determination of Systolic Heart Murmur by Cardiologists</td>
<td>500</td>
<td>.45</td>
<td></td>
</tr>
<tr>
<td>35. Abnormalities on Clinical Breast Examination by Surgeons or Nurses</td>
<td>1,720</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>36. Mean Quality Scores From Two Grant Panels:</td>
<td></td>
<td>( \text{Scale Item} )</td>
<td>( \text{Item} )</td>
</tr>
<tr>
<td>Dimensional ratings</td>
<td>1,290</td>
<td>.41</td>
<td></td>
</tr>
<tr>
<td>Dichotomous decision</td>
<td>1,177</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td>37. Job Performance Ratings by Peers</td>
<td>398</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>38. Number of Dimensions to Extract From a Correlation Matrix by Sree Plot</td>
<td>1,215</td>
<td>.43</td>
<td></td>
</tr>
<tr>
<td>39. Medical Quality of Care as Determined by Physician Peers</td>
<td>6,049</td>
<td>.37</td>
<td></td>
</tr>
<tr>
<td>40. Definitions of Invasive Fungal Infection in the Clinical Research Literature</td>
<td>2,300</td>
<td>.35</td>
<td></td>
</tr>
<tr>
<td>41. Job Performance Ratings by Subordinates</td>
<td>9,841</td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>42. Research Quality by Peer Reviewers:</td>
<td>21,659</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>Dimensional ratings</td>
<td>533</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>Dichotomous decisions</td>
<td>4,500</td>
<td>.31</td>
<td></td>
</tr>
<tr>
<td>43. Job Performance Ratings by Subordinates</td>
<td>6,129</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>44. Research Quality by Peer Reviewers:</td>
<td>24,939</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>Dimensional ratings</td>
<td>3,807</td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>Dichotomous decisions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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, \( \kappa \) = 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.

Results reported here differ slightly from those in Meyer et al. (2002) because Meyer et al. used \( n(k-1)/2 \) 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).

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).

For agreement among editors on the quality of peer reviewers, results varied noticeably depending on the weighting scheme selected and 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.

One study produced outlier results (\( \kappa = .90 \)) relative to the others (\( \kappa \) range from .36 to .45). When weighing by the total number of independent ratings \( \kappa \) remained at .52, but when weighing by the total number of objects judged by all possible rater pair combinations \( \kappa \) 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 reanalyzed 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 Ns 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 39) 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’Olaberriere et al., 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’Olaberriere 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, Schlige, 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 Pauno, Rautava, Helenius, Alanen, and Sillampää (1993). The Ns 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 ($\kappa = .90$) relative to the others ($\kappa$ 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, Ascioiu, 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” (Ascioiu 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 1 disorders, PubMed and PsycINFO were searched by combining the terms 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, & Kielcick-Glas, 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, inter-rater, 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 Klopfers'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 Mazilng 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 Klopf. 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 Frewing (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 researches 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 Weistrad (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 = 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
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 Frohlich’s (1985) findings on reviewer agreement were excluded because they used an ICC formula that differed from standard formulas (McGrath & 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 Callahan, Bax, Waeckerle, and Wears (1998); Feuer 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 Kooiman (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
interpretable 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 intrarater reliability observed when coding Rorschach or TAT protocols falls in the range between .80 and .91. This level of agreement compares favorably with the reliabilities 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 intrarater 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 traitlike 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 intrarater 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 & DeVecchio, 2000). Results from all the meta-analyses are presented in Table 25.2 and each will be described briefly.

Roberts and DeVecchio (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

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

*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.

*Retest intervals for the first three coefficients and Ns for the first and third were obtained from raw data in their Table 4.

*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.

*Reliability is reported using the kappa coefficient, not κ.

*Nonindependent samples were excluded.

*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.

*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).

*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). Shalman (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 Ns 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 through-out 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 re-test 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>
<thead>
<tr>
<th>Predictor test and criterion</th>
<th>Rorschach, TAT, or SCT</th>
<th>Other psych test</th>
<th>Medical test</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Dexamethasone Suppression Test Scores and Response to Depression Treatment</td>
<td>.90</td>
<td>.01</td>
<td>2.068</td>
<td></td>
</tr>
<tr>
<td>2. Routine Ultrasound Examinations and Successful Pregnancy Outcomes</td>
<td>.02</td>
<td>.07</td>
<td>16.227</td>
<td></td>
</tr>
<tr>
<td>3. MMPI Ego Strength Scores and Subsequent Psychotherapy Outcome</td>
<td>.07</td>
<td>.07</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>4. Unique Contribution of an MMPI High Point Code (vs. Other Codes) to Relevant Criteria*</td>
<td>.07</td>
<td>.07</td>
<td>8.514</td>
<td></td>
</tr>
<tr>
<td>5. MMPI Scores and Subsequent Prison Misconduct</td>
<td>.08</td>
<td>.08</td>
<td>17.636</td>
<td></td>
</tr>
<tr>
<td>6. MMPI Elevations on Scales F, 6, or 8 and Criminal Defendant Incompetency</td>
<td>.11</td>
<td>.11</td>
<td>1,461</td>
<td></td>
</tr>
<tr>
<td>8. MMPI Scale 8 and Differentiation of Schizophrenic versus Depressed Disorders</td>
<td>.12</td>
<td>.12</td>
<td>2,435</td>
<td></td>
</tr>
<tr>
<td>9. Lower Cognitive Ability and Involvement in Automobile Accidents</td>
<td>.12</td>
<td>.12</td>
<td>1,020</td>
<td></td>
</tr>
<tr>
<td>10. General Intelligence and Success in Military Pilot Training</td>
<td>.13</td>
<td>.13</td>
<td>15,403</td>
<td></td>
</tr>
<tr>
<td>11. Rorschach DEPI and Detection of Depressive Diagnosis</td>
<td>.14</td>
<td>.15</td>
<td>994</td>
<td></td>
</tr>
<tr>
<td>12. MMPI Scale 2 and Differentiation of Neurotic versus Psychotic Disorders</td>
<td>.14</td>
<td>.15</td>
<td>6,156</td>
<td></td>
</tr>
<tr>
<td>13. MMPI Scale 8 and Differentiation of Neurotic versus Psychotic Disorders</td>
<td>.14</td>
<td>.15</td>
<td>6,156</td>
<td></td>
</tr>
<tr>
<td>14. Baseline IQ and Subsequent Psychotherapy Outcome</td>
<td>.15</td>
<td>.16</td>
<td>246</td>
<td></td>
</tr>
<tr>
<td>15. Low Serotonin Metabolites in Cerebrospinal Fluid (5-HIAA) and Subsequent Suicide Attempts</td>
<td>.16</td>
<td>.17</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>16. MMPI Cook-Medley Hostility Scale Elevations and Subsequent Death From All Causes</td>
<td>.16</td>
<td>.17</td>
<td>4,747</td>
<td></td>
</tr>
<tr>
<td>17. Motivation to Manage from the Miner Sentence Completion Test and Managerial Effectiveness</td>
<td>.17</td>
<td>.18</td>
<td>2,151</td>
<td></td>
</tr>
<tr>
<td>18. MMPI Validity Scales and Detection of Known or Suspected Under-Reported Psychopathology</td>
<td>.18</td>
<td>.19</td>
<td>328</td>
<td></td>
</tr>
<tr>
<td>19. Dexamethasone Suppression Test Scores and Subsequent Suicide</td>
<td>.21</td>
<td>.22</td>
<td>626</td>
<td></td>
</tr>
<tr>
<td>20. Self-Reported Dependency Test Scores and Physical Illness</td>
<td>.21</td>
<td>.22</td>
<td>1,034</td>
<td></td>
</tr>
<tr>
<td>21. Rorschach to Detect Thought Disturbance in Relatives of Schizophrenic Patients</td>
<td>.22</td>
<td>.22</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>22. MMPI Dependency Scale and Dependent Behavior</td>
<td>.22</td>
<td>.22</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>23. TAT Scores of Achievement Motivation and Spontaneous Achievement Behavior</td>
<td>.22</td>
<td>.22</td>
<td>(k = 82)</td>
<td></td>
</tr>
<tr>
<td>24. Traditional Electrocardiogram Stress Test Results and Coronary Artery Disease</td>
<td>.24</td>
<td>.24</td>
<td>5,431</td>
<td></td>
</tr>
<tr>
<td>25. WISC Distractibility Subscales and Learning Disability Diagnoses</td>
<td>.24</td>
<td>.24</td>
<td>(K = 54)</td>
<td></td>
</tr>
<tr>
<td>26. Decreased Bone Mineral Density and Lifetime Risk of Hip Fracture in Women</td>
<td>.25</td>
<td>.25</td>
<td>20,849</td>
<td></td>
</tr>
<tr>
<td>27. General Intelligence Test Scores and Functional Effectiveness Across Jobs</td>
<td>.25</td>
<td>.26</td>
<td>40,230</td>
<td></td>
</tr>
<tr>
<td>28. Self-Reported Dependency Test Scores and Dependent Behavior</td>
<td>.26</td>
<td>.26</td>
<td>3,013</td>
<td></td>
</tr>
<tr>
<td>29. Thematic Story Test Dependency Scores and Recalled Physical Illness</td>
<td>.29</td>
<td>.28</td>
<td>269</td>
<td></td>
</tr>
<tr>
<td>30. C-Reactive Protein Test Results and Diagnosis of Acute Appendicitis</td>
<td>.29</td>
<td>.29</td>
<td>6,338</td>
<td></td>
</tr>
<tr>
<td>31. General Validity of Rorschach Studies Without Method Confounds</td>
<td>.29</td>
<td>.29</td>
<td>15,985</td>
<td></td>
</tr>
<tr>
<td>32. General Validity of MMPI Studies Without Method Confounds</td>
<td>.29</td>
<td>.30</td>
<td>6,520</td>
<td></td>
</tr>
<tr>
<td>33. For Women, Electrocardiogram Stress Test Results and Detection of Coronary Artery Disease</td>
<td>.30</td>
<td>.30</td>
<td>3,872</td>
<td></td>
</tr>
<tr>
<td>34. MMPI Scale 2 and Differentiation of Schizophrenic versus Depressed Disorders</td>
<td>.31</td>
<td>.31</td>
<td>2,435</td>
<td></td>
</tr>
<tr>
<td>35. General Validity of Rorschach Hypotheses Without Method Confounds</td>
<td>.32</td>
<td>.32</td>
<td>(k = 523)</td>
<td></td>
</tr>
<tr>
<td>36. General Validity of MMPI Hypotheses (Includes Some Method Confounds)</td>
<td>.32</td>
<td>.32</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>37. Screening Mammo gram Results and Detection of Breast Cancer Within 1 Year</td>
<td>.32</td>
<td>.32</td>
<td>263,259</td>
<td></td>
</tr>
<tr>
<td>38. General Validity of WAIS Studies Without Method Confounds</td>
<td>.33</td>
<td>.33</td>
<td>3,593</td>
<td></td>
</tr>
<tr>
<td>39. MMPI Scale 2 or DEP and Detection of Depressive Diagnosis</td>
<td>.35</td>
<td>.35</td>
<td>2,905</td>
<td></td>
</tr>
<tr>
<td>40. Incremental Contribution of Rorschach PRS Scores Over IQ to Predict Treatment Outcome</td>
<td>.36</td>
<td>.36</td>
<td>290</td>
<td></td>
</tr>
<tr>
<td>41. General Validity of WAIS Hypotheses Without Method Confounds</td>
<td>.36</td>
<td>.36</td>
<td>(k = 104)</td>
<td></td>
</tr>
<tr>
<td>42. Papnicolaou Test (Pap Smear) and Detection of Cervical Abnormalities</td>
<td>.36</td>
<td>.36</td>
<td>17,421</td>
<td></td>
</tr>
<tr>
<td>43. Competency Screening Sentence Completion Test Scores and Defender Competency</td>
<td>.37</td>
<td>.37</td>
<td>627</td>
<td></td>
</tr>
<tr>
<td>44. Rorschach Oral Dependence Scale and Dependent Behavior</td>
<td>.37</td>
<td>.37</td>
<td>1,320</td>
<td></td>
</tr>
<tr>
<td>45. Sperm Penetration Assay Results and Success With In Vitro Fertilization</td>
<td>.39</td>
<td>.39</td>
<td>1,335</td>
<td></td>
</tr>
<tr>
<td>46. MMPI Validity Scales to Detect Under-Reported Psychopathology (Primarily Analog Studies)</td>
<td>.39</td>
<td>.41</td>
<td>2,297</td>
<td></td>
</tr>
<tr>
<td>47. Computed Tomography Results and Detection of Lymph Node Metastases in Cervical Cancer</td>
<td>.41</td>
<td>.41</td>
<td>1,022</td>
<td></td>
</tr>
<tr>
<td>48. MMPI Scale 8 and Differentiation of Psychiatric Patients versus Controls</td>
<td>.42</td>
<td>.42</td>
<td>23,747</td>
<td></td>
</tr>
<tr>
<td>49. Conventional Dental X-rays and Diagnosis of Between-Tooth Cavities (Approximal Caries)</td>
<td>.43</td>
<td>.43</td>
<td>(K = 8)</td>
<td></td>
</tr>
<tr>
<td>50. Rorschach SCZD and Detection of Psychotic Diagnosis</td>
<td>.44</td>
<td>.44</td>
<td>717</td>
<td></td>
</tr>
<tr>
<td>51. MMPI Scale 2 and Differentiation of Psychiatric Patients versus Controls</td>
<td>.44</td>
<td>.44</td>
<td>23,747</td>
<td></td>
</tr>
<tr>
<td>52. WAIS IQ and Obtained Level of Education</td>
<td>.44</td>
<td>.44</td>
<td>(k = 9)</td>
<td></td>
</tr>
<tr>
<td>53. Digitally Enhanced Dental X-rays and Diagnosis of Biting Surfaces Cavities</td>
<td>.44</td>
<td>.44</td>
<td>2,870</td>
<td></td>
</tr>
<tr>
<td>54. Rorschach PRS Scores and Subsequent Psychotherapy Outcome</td>
<td>.45</td>
<td>.45</td>
<td>624</td>
<td></td>
</tr>
<tr>
<td>55. MMPI Validity Scales and Detection of Known or Suspected Malingered Psychopathology</td>
<td>.45</td>
<td>.45</td>
<td>771</td>
<td></td>
</tr>
<tr>
<td>56. Thematic Story Test Dependency Scores and Dependent Behavior</td>
<td>.46</td>
<td>.46</td>
<td>448</td>
<td></td>
</tr>
<tr>
<td>57. Rorschach X + % and Differentiation of Clinical or Target Groups From Controls</td>
<td>.46</td>
<td>.46</td>
<td>1,517</td>
<td></td>
</tr>
<tr>
<td>58. Antineutrophil Cytoplasmic Antibody Testing and Detection of Wegener Granulomatosis</td>
<td>.46</td>
<td>.47</td>
<td>13,562</td>
<td></td>
</tr>
<tr>
<td>59. Lecithin/Sphingomyelin Ratio and Prediction of Neonatal Respiratory Distress Syndrome</td>
<td>.50</td>
<td>.50</td>
<td>1,170</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
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-
Appendix: Citations Contributing Data to the Larger Meta-Analyses of Interrater Reliability Conducted for This Chapter

Performance 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 Mustosimone (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); Rubenfeld and Wier (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)


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, DeBrotta, 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, Weikowitz, Boksay, and Seeland (1988; two samples); Fuglism 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 Leibenthal (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); Gjerret 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); Kirks 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); Marusis, Mestrovic, Petrovecki, and Marusic (1998; M of two ratings); Munley et al. (1988); Ploos and Herzog (2001); Ploutz (1993); Rothwell and Martyn (2000; four samples, two results from raw data); Rubin, Redmelmeier, 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 topics.

3. Several findings in Table 25.1 also address retest reliability. Specifically, intrarater 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


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


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


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


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.


