Psychological Testing and Psychological Assessment

A Review of Evidence and Issues

Gregory J. Meyer	*University of Alaska Anchorage*
Stephen E. Finn	*Center for Therapeutic Assessment*
Lorraine D. Eyde	*U.S. Office of Personnel Management*
Gary G. Kay	*Georgetown University Medical Center*
Kevin L. Moreland	*Fort Walton Beach, FL*
Robert R. Dies	*New Port Richey, FL*
Elena J. Eisman	*Massachusetts Psychological Association*
Tom W. Kubiszyn and Geoffrey M. Reed	*American Psychological Association*

This article summarizes evidence and issues associated with psychological assessment. Data from more than 125 meta-analyses on test validity and 800 samples examining multimethod assessment suggest 4 general conclusions: (a) Psychological test validity is strong and compelling, (b) psychological test validity is comparable to medical test validity, (c) distinct assessment methods provide unique sources of information, and (d) clinicians who rely exclusively on interviews are prone to incomplete understandings. Following principles for optimal nomothetic research, the authors suggest that a multimethod assessment battery provides a structured means for skilled clinicians to maximize the validity of individualized assessments. Future investigations should move beyond an examination of test scales to focus more on the role of psychologists who use tests as helpful tools to furnish patients and referral sources with professional consultation.

For clinical psychologists, assessment is second only to psychotherapy in terms of its professional importance (Greenberg, Smith, & Muenzen, 1995; Norcross, Karg, & Prochaska, 1997; Phelps, Eisman, & Kohout, 1998). However, unlike psychotherapy, formal assessment is a distinctive and unique aspect of psychological practice relative to the activities performed by other health care providers. Unfortunately, with dramatic health care changes over the past decade, the utility of psychological assessment has been increasingly challenged (Eisman et al., 1998, 2000), and there has been declining use of the time-intensive, clinician-administered instruments that have historically defined professional practice (Piotrowski, 1999; Piotrowski, Belter, & Keller, 1998).

In response, the American Psychological Association's (APA) Board of Professional Affairs (BPA) established a Psychological Assessment Work Group (PAWG) in 1996 and commissioned it (a) to evaluate contemporary threats to psychological and neuropsychological assessment services and (b) to assemble evidence on the efficacy of assessment in clinical practice. The PAWG's findings and recommendations were released in two reports to the BPA (Eisman et al., 1998; Meyer et al., 1998; also see Eisman et al., 2000; Kubiszyn et al., 2000). This article extends Meyer et al. (1998) by providing a large and systematic summary of evidence on testing and assessment.[1]

Our goals are sixfold. First, we briefly describe the purposes and appropriate applications of psychological assessment. Second, we provide a broad overview of testing and assessment validity. Although we present a great deal of data, by necessity, we paint in broad strokes and rely heavily on evidence gathered through meta-analytic reviews. Third, to help readers understand the strength of the assessment evidence, we highlight findings in two comparative contexts. To ensure a general understanding of what constitutes a small or large correlation (our effect size measure), we review a variety of nontest correlations culled from psychology, medicine, and everyday life. Next, to more specifically appreciate the test findings, we consider

Gregory J. Meyer, Department of Psychology, University of Alaska Anchorage; Stephen E. Finn, Center for Therapeutic Assessment, Austin, TX; Lorraine D. Eyde, U.S. Office of Personnel Management, Washington, DC; Gary G. Kay, Georgetown University Medical Center; Kevin L. Moreland, independent practice, Fort Walton Beach, FL; Robert R. Dies, independent practice, New Port Richey, FL; Elena J. Eisman, Massachusetts Psychological Association, Boston, MA; Tom W. Kubiszyn and Geoffrey M. Reed, Practice Directorate, American Psychological Association, Washington, DC.

Tom W. Kubiszyn is now at the Department of Educational Psychology, University of Texas at Austin.

Kevin L. Moreland passed away in 1999.

We thank the Society for Personality Assessment for supporting Gregory J. Meyer's organization of the literature summarized in this article.

Correspondence concerning this article should be addressed to Gregory J. Meyer, Department of Psychology, University of Alaska Anchorage, 3211 Providence Drive, Anchorage, AK 99508. Electronic mail may be sent to afgjm@uaa.alaska.edu.

[1] The PAWG reports can be obtained free of charge from Christopher J. McLaughlin, Assistant Director, Practice Directorate, American Psychological Association, 750 First Street NE, Washington, DC 20002-4242; e-mail: cmclaughlin@apa.org. Because of space limitations, this article does not cover some important issues detailed in Meyer et al. (1998).

0003-066X/01/$5.00
Vol. 56, No. 2, 128–165 DOI: 10.1037//0003-066X.56.2.128

psychological test validity alongside medical test validity. On the basis of these data, we conclude that there is substantial evidence to support psychological testing and assessment. Fourth, we describe features that make testing a valuable source of clinical information and present an extensive overview of evidence that documents how distinct methods of assessment provide unique perspectives. We use the latter to illustrate the clinical value of a multimethod test battery and to highlight the limitations that emerge when using an interview as the sole basis for understanding patients. Fifth, we discuss the distinction between testing and assessment and highlight vital issues that are often overlooked in the research literature. Finally, we identify productive avenues for future research.

The Purposes and Appropriate Uses of Psychological Assessment

Some of the primary purposes of assessment are to (a) describe current functioning, including cognitive abilities, severity of disturbance, and capacity for independent living; (b) confirm, refute, or modify the impressions formed by clinicians through their less structured interactions with patients; (c) identify therapeutic needs, highlight issues likely to emerge in treatment, recommend forms of intervention, and offer guidance about likely outcomes; (d) aid in the differential diagnosis of emotional, behavioral, and cognitive disorders; (e) monitor treatment over time to evaluate the success of interventions or to identify new issues that may require attention as original concerns are resolved; (f) manage risk, including minimization of potential legal liabilities and identification of untoward treatment reactions; and (g) provide skilled, empathic assessment feedback as a therapeutic intervention in itself.

APA ethical principles dictate that psychologists provide services that are in the best interests of their patients (American Psychological Association, 1992). Thus, all assessors should be able to furnish a sound rationale for their work and explain the expected benefits of an assessment, as well as the anticipated costs. Although it is valuable to understand the benefits of a test relative to its general costs, it is important to realize how cost-benefit ratios ultimately can be determined only for individual patients when working in a clinical context (Cronbach & Gleser, 1965; Finn, 1982). Tests expected to have more benefits than costs for one patient may have different or even reversed cost-benefit ratios for another. For instance, memory tests may have an excellent cost-benefit ratio for an elderly patient with memory complaints but a decidedly unfavorable ratio for a young adult for whom there is no reason to suspect memory problems. This implies that general bureaucratic rules about appropriate test protocols are highly suspect. A test that is too long or costly for general use may be essential for clarifying the clinical picture with particular patients. In addition, certain assessment practices that may have been common in some settings can now be seen as questionable, including (a) mandated testing of patients on a fixed schedule regardless of whether the repeat assessment is clinically indicated, (b) administrative guidelines specifying that all patients or no patients are to receive psychological evaluations, and (c) habitual testing of all patients using large fixed batteries (Griffith, 1997; Meier, 1994).

Finally, although specific rules cannot be developed, provisional guidelines for when assessments are likely to have the greatest utility in general clinical practice can be offered (Finn & Tonsager, 1997; Haynes, Leisen, & Blaine, 1997).[2] In *pretreatment evaluation,* when the goal is to describe current functioning, confirm or refute clinical impressions, identify treatment needs, suggest appropriate interventions, or aid in differential diagnosis, assessment is likely to yield the greatest overall utility when (a) the treating clinician or patient has salient questions, (b) there are a variety of treatment approaches from which to choose and a body of knowledge linking treatment methods to patient characteristics, (c) the patient has had little success in prior treatment, or (d) the patient has complex problems and treatment goals must be prioritized. The *therapeutic impact* of assessment on patients and their interpersonal systems (i.e., family, teachers, and involved health service providers) is likely to be greatest when (a) initial treatment efforts have failed, (b) patients are curious about themselves and motivated to participate, (c) collaborative procedures are used to engage the patient, (d) family and allied health service providers are invited to furnish input, and (e) patients and relevant others are given detailed feedback about results.

Identifying several circumstances when assessments are likely to be particularly useful does not mean that assessments under other circumstances are questionable. Rather, the key that determines when assessment is appropriate is the rationale for using specific instruments with a particular patient under a unique set of circumstances to address a distinctive set of referral questions. An assessment should not be performed if this information cannot be offered to patients, referring clinicians, and third-party payers.

A Foundation for Understanding Testing and Assessment Validity Evidence

To summarize the validity literature on psychological testing and assessment, we use the correlation coefficient as our effect size index. In this context, the effect size quantifies the strength of association between a predictor test scale and a relevant criterion variable. To judge whether the test validity findings are poor, moderate, or substantial, it helps to be clear on the circumstances when one is likely to see a correlation of .10, .20, .30, and so on. Therefore, before delving into the literature on testing and assessment,

(*text continues on page 132*)

[2] Different issues are likely to come to the forefront during forensic evaluations, although they are not considered here.

Table 1
Examples of the Strength of Relationship Between Two Variables in Terms of the Correlation Coefficient (r)

Predictor and criterion (study and notes)	*r*	*N*
1. Effect of sugar consumption on the behavior and cognitive processes of children (Wolraich, Wilson, & White, 1995; the sample-size weighted effect across the 14 measurement categories reported in their Table 2 was $r = .01$. However, none of the individual outcomes produced effect sizes that were significantly different from zero. Thus, $r = 0.0$ is reported as the most accurate estimate of the true effect).	.00	560
2. Aspirin and reduced risk of death by heart attack (Steering Committee of the Physicians' Health Study Research Group, 1988).	.02	22,071
3. Antihypertensive medication and reduced risk of stroke (Psaty et al., 1997; the effect of treatment was actually smaller for all other disease end points studied [i.e., coronary heart disease, congestive heart failure, cardiovascular mortality, and total mortality]).	.03	59,086
4. Chemotherapy and surviving breast cancer (Early Breast Cancer Trialists' Collaborative Group, 1988).	.03	9,069
5. Post-MI cardiac rehabilitation and reduced death from cardiovascular complications (Oldridge, Guyatt, Fischer, & Rimm, 1988; weighted effect calculated from data in their Table 3. Cardiac rehabilitation was not effective in reducing the risk for a second nonfatal MI [$r = -.03$; effect in direction opposite of expectation]).	.04	4,044
6. Alendronate and reduction in fractures in postmenopausal women with osteoporosis (Karpf et al., 1997; weighted effect calculated from data in their Table 3).	.05	1,602
7. General batting skill as a Major League baseball player and hit success on a given instance at bat (Abelson, 1985; results were mathematically estimated by the author, and thus, no *N* is given).	.06	—
8. Aspirin and heparin (vs. aspirin alone) for unstable angina and reduced MI or death (Oler, Whooley, Oler, & Grady, 1996; weighted effect calculated from data in their Table 2).	.07	1,353
9. Antibiotic treatment of acute middle ear pain in children and improvement at 2–7 days (Del Mar, Glasziou, & Hayem, 1997; coefficient derived from *z* value reported in their Figure 1. All other outcomes were smaller).	.08	1,843
10. Calcium intake and bone mass in premenopausal women (Welten, Kemper, Post, & Van Staveren, 1995).	.08	2,493
11. Coronary artery bypass surgery for stable heart disease and survival at 5 years (Yusuf et al., 1994).	.08	2,649
12. Ever smoking and subsequent incidence of lung cancer within 25 years (Islam & Schottenfeld, 1994).	.08	3,956
13. Gender and observed risk-taking behavior (males are higher; Byrnes, Miller, & Schafer, 1999).	.09	($k = 94$)
14. Impact of parental divorce on problems with child well-being and functioning (Amato & Keith, 1991).	.09	($k = 238$)
15. Alcohol use during pregnancy and subsequent premature birth (data combined from Kliegman, Madura, Kiwi, Eisenberg, & Yamashita, 1994, and Jacobson et al., 1994).	.09	741
16. Antihistamine use and reduced runny nose and sneezing (D'Agostino et al., 1998; these results were averaged across criteria and days of assessment. The largest independent *N* is reported).	.11	1,023
17. Combat exposure in Vietnam and subsequent PTSD within 18 years (Centers for Disease Disease Control Vietnam Experience Study, 1988).	.11	2,490
18. Extent of low-level lead exposure and reduced childhood IQ (Needleman & Gatsonis, 1990; effect size reflects a partial correlation correcting for other baseline characteristics that affect IQ scores [e.g., parental IQ], derived as the weighted effect across blood and tooth lead measurements reported in their Table 5).	.12	3,210
19. Extent of familial social support and lower blood pressure (Uchino, Cacioppo, & Kiecolt-Glaser, 1996).	.12	($K = 12$)
20. Impact of media violence on subsequent naturally occurring interpersonal aggression (Wood, Wong, & Chachere, 1991).	.13	($k = 12$)
21. Effect of relapse prevention on improvement in substance abusers (Irvin, Bowers, Dunn, & Wang, 1999).	.14	($K = 26$)
22. Effect of nonsteroidal anti-inflammatory drugs (e.g., ibuprofen) on pain reduction (results were combined from Ahmad et al., 1997; Eisenberg, Berkey, Carr, Mosteller, & Chalmers, 1994; and Po & Zhang, 1998; effect sizes were obtained from mean differences in the treatment vs. control conditions in conjunction with the standard error of the difference and the appropriate *n*s. The meta-analyses by Po and Zhang [$N = 3,390$] and by Ahmad et al. [$N = 4,302$] appeared to use the same data for up to 458 patients. Thus, the total *N* reported here was reduced by this number. Across meta-analyses, multiple outcomes were averaged, and, because *n*s fluctuated across dependent variables, the largest value was used to represent the study. Finally, Po and Zhang reported that codeine added to ibuprofen enhanced pain reduction, though results from the other two studies did not support this conclusion).	.14	8,488

Table 1 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
23. Self-disclosure and likability (Collins & Miller, 1994).	.14	(*k* = 94)
24. Post-high school grades and job performance (Roth, BeVier, Switzer, & Schippmann, 1996).	.16	13,984
25. Prominent movie critics' ratings of 1998 films and U.S. box office success (data combined from Lewin, 1999, and the Movie Times, 1999; the reported result is the average correlation computed across the ratings given by 15 movie critics. For each critic, ratings for up to 100 movies were correlated with the adjusted box office total gross income [adjusted gross = gross income/maximum number of theaters that showed the film]).	.17	(*k* = 15)
26. Relating material to oneself (vs. general "others") and improved memory (Symons & Johnson, 1997; coefficient derived from their Table 3).	.17	(*k* = 69)
27. Extent of brain tissue destruction on impaired learning behavior in monkeys (Irle, 1990; the average effect was derived from Spearman correlations and combined results across all eight dependent variables analyzed. As indicated by the author, similar findings have been obtained for humans).	.17	(*K* = 283)
28. Nicotine patch (vs. placebo) and smoking abstinence at outcome (Fiore, Smith, Jorenby, & Baker, 1994; sample weighted effect calculated from data in their Table 4. Effect was equivalent for abstinence at end of treatment and at 6-month follow-up).	.18	5,098
29. Adult criminal history and subsequent recidivism among mentally disordered offenders (Bonta, Law, & Hanson, 1998; data from their Table 8 were combined for criminal and violent recidivism and the average *Zr* [mean effect size] was transformed to *r*).	.18	6,475
30. Clozapine (vs. conventional neuroleptics) and clinical improvement in schizophrenia (Wahlbeck, Cheine, Essali, & Adams, 1999).	.20	1,850
31. Validity of employment interviews for predicting job success (McDaniel, Whetzel, Schmidt, & Maurer, 1994).	.20	25,244
32. Extent of social support and enhanced immune functioning (Uchino, Cacioppo, & Kiecolt-Glaser, 1996).	.21	(*K* = 9)
33. Quality of parents' marital relationship and quality of parent–child relationship (Erel & Burman, 1995).	.22	(*k* = 253)
34. Family/couples therapy vs. alternative interventions and outcome of drug abuse treatment (Stanton & Shadish, 1997; data drawn from their Table 3).	.23	(*K* = 13)
35. General effectiveness of psychological, educational, and behavioral treatments (Lipsey & Wilson, 1993).	.23	(*K* ≈ 9,400)
36. Effect of alcohol on aggressive behavior (Ito, Miller, & Pollock, 1996; data drawn from their p. 67).	.23	(*K* = 47)
37. Positive parenting behavior and lower rates of child externalizing behavior problems (Rothbaum & Weisz, 1995).	.24	(*K* = 47)
38. Viagra (oral sildenafil) and side effects of headache and flushing (Goldstein et al., 1998; coefficient is the weighted effect from their Table 3 comparing Viagra with placebo in both the DR and DE trials).	.25	861
39. Gender and weight for U.S. adults (men are heavier; U.S. Department of Health and Human Services National Center for Health Statistics, 1996[a]; analysis used only weights that were actually measured).	.26	16,950
40. General validity of screening procedures for selecting job personnel: 1964–1992 (Russell et al., 1994; coefficient reflects the unweighted average validity coefficient from studies published in *Personnel Psychology* and *Journal of Applied Psychology*).	.27	(*K* = 138)
41. Effect of psychological therapy under clinically representative conditions (Shadish et al., 1997).[b]	.27	(*K* = 56)
42. ECT for depression (vs. simulated ECT) and subsequent improvement (Janick et al., 1985).	.29	205
43. Sleeping pills (benzodiazapines or zolpidem) and short-term improvement in chronic insomnia (Nowell et al., 1997; effect size of treatment relative to placebo, averaged across outcomes of sleep-onset latency, total sleep time, number of awakenings, and sleep quality, as reported in their Table 5. *N* derived from their text, not from their Table 1).	.30	680
44. Clinical depression and suppressed immune functioning (Herbert & Cohen, 1993; weighted effect derived from all parameters in their Table 1 using the "restricted" methodologically superior studies. Average *N* is reported).	.32	438
45. Psychotherapy and subsequent well-being (M. L. Smith & Glass, 1977).	.32	(*K* = 375)
46. Gender and self-reported assertiveness (males are higher; Feingold, 1994; coefficient derived from the "general adult" row of Feingold's Table 6).	.32	19,546
47. Test reliability and the magnitude of construct validity coefficients (Peter & Churchill, 1986; the authors used the term *nomological validity* rather than construct validity).	.33	(*k* = 129)

(*table continues*)

Table 1 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
48. Elevation above sea level and lower daily temperatures in the U.S.A. (National Oceanic and Atmospheric Administration, 1999; data reflect the average of the daily correlations of altitude with maximum temperature and altitude with minimum temperature across 187 U.S. recording stations for the time period from January 1, 1970, to December 31, 1996).	.34	(*k* = 19,724)
49. Viagra (oral sildenafil) and improved male sexual functioning (Goldstein et al., 1998; coefficient is the weighted effect comparing Viagra with placebo from both the DR and DE trials. The authors did not report univariate effect size statistics, so effects were derived from all outcomes that allowed for these calculations: (a) frequency of penetration [DR, DE], (b) maintenance after penetration [DR, DE], (c) percentage of men reporting global improvement [DR, DE], and (d) percentage of men with Grade 3 or 4 erections [DR]. For (a) and (b) in the DE trial, the pooled SD was estimated from the more differentiated subgroup standard errors presented in their Table 2. *N* varied across analyses, and the average is reported).	.38	779
50. Observer ratings of attractiveness for each member of a romantic partnership (Feingold, 1988).	.39	1,299
51. Past behavior as a predictor of future behavior (Ouellette & Wood, 1998; data drawn from their Table 1).	.39	(*k* = 16)
52. Loss in habitat size and population decline for interior-dwelling species[c] (Bender, Contreras, & Fahrig, 1998; the *N* in this analysis refers to the number of landscape patches examined).	.40	2,406
53. Social conformity under the Asch line judgment task (Bond & Smith, 1996).	.42	4,627
54. Gender and self-reported empathy and nurturance (females are higher; Feingold, 1994; coefficient is derived from the "general adult" row of Feingold's Table 6).	.42	19,546
55. Weight and height for U.S. adults (U.S. Department of Health and Human Services National Center for Health Statistics, 1996; analysis used only weights and heights that were actually measured).	.44	16,948
56. Parental reports of attachment to their parents and quality of their child's attachment (Van IJzendoorn, 1995).	.47	854
57. Increasing age and declining speed of information processing in adults (Verhaeghen & Salthouse, 1997).	.52	11,044
58. Gender and arm strength for adults (men are stronger; Blakley, Quiñones, & Crawford, 1994[a]; effect size was computed from the means and standard deviations for arm lift strength reported in their Table 6).	.55	12,392
59. Nearness to the equator and daily temperature in the U.S.A. (National Oceanic and Atmospheric Administration, 1999; data reflect the average of the daily correlations for latitude with maximum temperature and latitude with minimum temperature across 187 U.S. recording stations for the time period from January 1, 1970, to December 31, 1996).	.60	(*k* = 19,724)
60. Gender and height for U.S. adults (men are taller; U.S. Department of Health and Human Services National Center for Health Statistics, 1996[a]; analysis used only heights that were actually measured).	.67	16,962

Note. DE = dose-escalation; DR = dose-response; ECT = electroconvulsive therapy; IQ = intelligence quotient; *k* = number of effect sizes contributing to the mean estimate; *K* = number of studies contributing to the mean estimate; MI = myocardial infarction; PTSD = posttraumatic stress disorder.
[a] These values differ from those reported by Meyer and Handler (1997) and Meyer et al. (1998) because they are based on larger samples. [b] Treatment was conducted outside a university, patients were referred through usual clinical channels, and treatment was conducted by experienced therapists with regular caseloads. For a subgroup of 15 studies in which therapists also did not use a treatment manual and did not have their treatment techniques monitored, the average *r* was .25. [c] Interior-dwelling species are those that are live within the central portion of a habitat as opposed to its border.

we present an overview of some non-test-related correlational values.[3] We believe this is important for several reasons. Because psychology has historically emphasized statistical significance over effect size magnitudes and because it is very hard to recognize effect magnitudes from many univariate statistics (e.g., *t*, *F*, χ^2) or multivariate analyses, it is often difficult to appreciate the size of the associations that are studied in psychology or encountered in daily life.

In addition, three readily accessible but inappropriate benchmarks can lead to unrealistically high expectations about effect magnitudes. First, it is easy to recall a perfect association (i.e., $r = 1.00$). However, perfect associations are never encountered in applied psychological research, making this benchmark unrealistic. Second, it is easy to implicitly compare validity correlations with reliability coefficients because the latter are frequently reported in the literature. However, reliability coefficients (which are often

[3] J. Cohen (1988) suggested helpful rules of thumb to characterize the size of correlations (wherein $r \approx \pm .10$ is *small*, $r \approx \pm .30$ is *medium*, and $r \approx \pm .50$ is *large*). However, following Rosenthal (1990, 1995), we believe it is most optimal to let actual relationships serve as mental benchmarks.

in the range of $r = .70$ or higher) evaluate only the correspondence between a variable and itself. As a result, they cannot provide a reasonable standard for evaluating the association between two distinct real-world variables.

A final class of coefficients may often come to mind, though again they do not provide a reasonable standard of comparison. These are monomethod validity coefficients. Such coefficients (often in the range of $r \geq .50$) are ubiquitous in the psychological literature. They are obtained whenever numerical values on a predictor and criterion are completely or largely derived from the same source of information. Examples include (a) a self-report scale (e.g., of depression) that is validated by correlating it with a conceptually similar scale that is also derived from self-report (i.e., another questionnaire or a structured interview) or (b) an individually administered performance task (e.g., of verbal intelligence) that is correlated with a second performance task thought to measure the same construct. Because the systematic error of method variance is aligned in such studies, the results are artificially inflated and do not provide a reasonable benchmark for considering the real-world associations between two independently measured variables.

With the foregoing in mind, Table 1 presents a range of illustrative correlations. When considering these results (and those in the next table), several points should be noted. First, all examples make use of coefficients that have not been corrected for unreliability, range restriction, or the imperfect construct validity of criterion measures. Second, the coefficients do not all come from equivalent designs. Some studies select extreme groups of participants (e.g., patients with severe Alzheimer's disease vs. nonpatients with normal cognitive functioning); examine rare, low base-rate events; artificially dichotomize truly continuous variables; use relatively small samples; or use procedures not typically found in clinical practice (e.g., consensus reading of electrocardiograms by two physicians). All of these methodological factors can influence validity coefficients and make them fluctuate or systematically differ in magnitude (Hunter & Schmidt, 1990). Consequently, even though table entries are organized by their magnitude, differences between one entry and another should be interpreted cautiously.

In terms of the data in Table 1, one of the first examples indicates how taking aspirin on a regular basis helps to reduce the risk of dying from a heart attack ($r = .02$; Table 1, Entry 2), even though the effect would be considered quite small. Other small effects include the impact of chemotherapy on breast cancer survival ($r = .03$; Table 1, Entry 4), the association between a major league baseball player's batting average and his success in obtaining a hit in a particular instance at bat ($r = .06$; Table 1, Entry 7), and the value of antihistamines for reducing sneezes and a runny nose ($r = .11$; Table 1, Entry 16). Correlations are somewhat higher for the extent of damaged brain tissue and impaired learning in nonhuman primates ($r = .17$; Table 1, Entry 27), the link between prominent movie critics' reviews and box office success ($r = .17$; Table 1, Entry 25), and the ability of employment interviews to predict job success ($r = .20$; Table 1, Entry 31). In the middle range of the values listed in Table 1 are the association of gender and weight ($r = .26$; Table 1, Entry 39), the effect of psychotherapy under clinically representative conditions ($r = .27$; Table 1, Entry 41), the effect of sleeping pills for short-term treatment of insomnia ($r = .30$; Table 1, Entry 43), the impact of elevation on daily temperatures in the United States ($r = .34$; Table 1, Entry 48), and the effect of contiguous natural environments on the population density of species that prefer the center of those habitats ($r = .40$; Table 1, Entry 52). Recently, the medication Viagra has received extensive media attention. As Table 1 indicates, the initial large-scale clinical trial on this drug found that its impact on improved sexual functioning was $r = .38$ (Table 1, Entry 49), whereas its influence on unwanted side effects was $r = .25$ (Table 1, Entry 38). At the high end of the spectrum is the relationship between gender and arm strength ($r = .55$; Table 1, Entry 58) or height ($r = .67$; Table 1, Entry 60), with male adults being stronger and taller than female adults. One also sees a strong connection between physical distance from the equator (and thus the sun) and daily temperature recordings in the United States ($r = .60$; Table 1, Entry 59), such that in the northern hemisphere, more northern locations have cooler temperatures than southern ones.

By and large, the examples in Table 1 illustrate how many medical and psychological interventions (e.g., antihypertensive medication, nicotine patches, sleeping pills, psychotherapy), as well as many constructs that interest psychologists (e.g., the impact of divorce, parenting strategies, memorization techniques, alcohol, psychometric reliability), produce correlations in the range of approximately .15 to .30. Even the axiom that past behavior is the best predictor of future behavior produces a correlation of only $r = .39$ (Table 1, Entry 51; see Ouellette & Wood, 1998, for moderators).

In many respects, these findings highlight how challenging it is to consistently achieve uncorrected univariate correlations that are much above .30. Given psychologists' frequent desire to square correlational values and discuss findings using proportion of variance terminology, some may feel disappointed by the magnitudes in Table 1 because many variables account for only about 2% to 9% of the variance in a criterion.[4] Indeed, even the extent of brain damaged tissue accounts for only 3% of the variance in primate learning behavior, the degree of landscape fragmentation accounts for only 16% of the variance in the population density of central habitat species, and the distance from the sun accounts for only 37% of the variance in daily U.S. temperature. For those who may be inclined to square the values in table 1 and feel discouraged, we recommend an alternative, which is to reconceptualize effect size magnitudes.

[4] For a general criticism of squared correlations and reasons to avoid them, see D'Andrade and Dart (1990) and Ozer (1985). For a discussion of why r should be preferred to r^2 as an effect size measure, see J. Cohen (1988), Hunter and Schmidt (1990), and Rosenthal (1991).

Instead of relying on unrealistic benchmarks to evaluate the findings in Table 1, it seems that psychologists studying highly complex human behavior should be rather satisfied when they can identify replicated univariate correlations among independently measured constructs that are of the magnitude observed for antihistamine effectiveness (r = .11; Table 1, Entry 16), college grades and job performance (r = .16; Table 1, Entry 24), or criminal history and recidivism (r = .18; Table 1, Entry 29). Furthermore, it appears that psychologists generally should be pleased when they can attain replicated univariate correlations among independently measured constructs that approximate the magnitude seen for gender and weight (r = .26; Table 1, Entry 39), reliability and validity (r = .33; Table 1, Entry 47), or elevation above sea level and daily temperature (r = .34; Table 1, Entry 48). Finally, psychologists probably should rejoice when they find replicated evidence that uncorrected univariate correlations are of the same magnitude as those observed for gender and arm strength (r = .55; Table 1, Entry 58) or for latitude and daily temperature (r = .60; Table 1, Entry 59).

Examples of Evidence Supporting the Goals of Psychological Testing and Assessment

The PAWG report provided a narrative review of data on the utility of testing for various clinical purposes (Meyer et al., 1998; also see Kubiszyn et al., 2000), including (a) the description of clinical symptomatology and differential diagnosis, (b) the description and prediction of functional capacities in everyday behavior, (c) the prediction of subsequent functioning and differential treatment needs for medical and mental health conditions, (d) the monitoring of treatment over time, and (e) the use of psychological assessment as a treatment in itself. Our current goal is to provide a more systematic overview of the psychological testing and assessment evidence.

To provide a reasonable overview of the evidence, we present data from meta-analytic reviews and several large-scale studies (the latter are noted in our table). To identify relevant meta-analyses, we searched PsycINFO for English language articles using the term *meta-analy** combined with the terms *test* or *validity* or *neuropsych** or *personality* or *cognitive*.[5] When the search was last run (December 1999), it produced 1,352 articles, to which we added 5 studies uncovered during a search of the medical literature (see below) and 5 that were known to us but had not been indexed. After deleting irrelevant articles, 241 studies remained. From these, we selected examples that either reviewed commonly used instruments or illustrated a wide range of testing and assessment applications. Specifically, from the pool of 241 meta-analyses, we obtained and reviewed 107 articles and present results from 69.[6] No studies were excluded because of the results they obtained.

To provide a reasonable overview of the evidence on medical testing, we used PubMed to search the English language MEDLINE literature with three strategies. The first search combined the MeSH terms *meta-analysis* and *Diagnostic Techniques and Procedures*. The second strategy was an unrestricted field search that combined the term *meta-analysis* with *MRI* or *CT* or *ultrasound* or *x-ray* or *sensitivity* or *specificity*. These searches produced 776 unique references, which were combined with 12 medical test citations found in our PsycINFO search and 3 additional citations from a recent review (Lijmer et al., 1999). After deleting irrelevant articles, we were left with a final pool of 203 articles. From these, we again selected examples that reviewed commonly used instruments or illustrated a wide range of applications. From the pool of 203 meta-analyses, we obtained and reviewed 99 and present results for 57.[7] No studies were excluded due to the results they obtained. Our final search examined medically focused, multidisciplinary geriatric assessment teams. Because many controlled trials have examined the value of these teams on subsequent survival, we extended a 1991 meta-analysis on this topic through July 1999. Post-1989 studies were identified by combining the following text words: (*assessment* or *evaluation* or *consultation*) and *geriatric* and (*control** or *random**) and (*mortality* or *survival*). This search produced 109 studies, for which 18 provided relevant data. In conjunction with the earlier meta-analysis, results from a total of 32 samples were summarized.

Table 2 presents the findings from our review, with validity coefficients for psychological tests interspersed

[5] A complete list of all search results and decisions can be obtained from Gregory J. Meyer.

[6] Irrelevant articles included comments or letters and meta-analyses that dealt with (a) psychotherapy, (b) medical tests or procedures, (c) the reliability or internal structure of a test, (d) methodological issues, (e) gender differences in personality or cognitive functioning, (f) nonapplied topics (e.g., extrasensory perception), and (e) instances when meta-analysis was used only to summarize several samples gathered by the author(s). The 38 studies that we obtained but did not use were excluded because they did not allow us to calculate a univariate correlational effect size (n = 13), presented results without clear hypotheses or that were difficult to characterize as validity coefficients (e.g., sensitivity to change from various treatments; lack of ethnic differences; n = 7), did not use traditional psychological tests or mixed test and nontest predictors (n = 7), overlapped with results from a larger or more recent meta-analysis (n = 4), presented clearly confounded predictors and criteria (n = 4), examined a literature that the original authors believed was unsuitable for meta-analysis (n = 1), were not genuine meta-analyses (n = 1), or summarized only statistically significant findings from the primary studies (n = 1). When necessary, we translated original research findings into a correlation using standard formulas (see, e.g., Rosenthal, 1991). Because some studies included variables with unequal variances, skewed distributions, or very high specificity rates, we did not use the procedures detailed by Hasselblad and Hedges (1995).

[7] Irrelevant articles included comments and letters as well as meta-analyses that (a) dealt with treatment, (b) addressed methodology, (c) focused on incidence or prevalence, (d) did not have an abstract, (e) dealt with psychological tests, (f) focused solely on estimating cost effectiveness, or (g) dealt with animals. The 42 studies that we obtained but did not use were excluded because they did not allow us to calculate a univariate correlational effect size (n = 29), overlapped with results reported elsewhere or from a more recent meta-analysis (n = 6), were not a genuine meta-analysis or estimated only normative test values (n = 3), did not use traditional definitions for statistics or the accepted gold standard criterion (n = 2), relied heavily on data from abstracts rather than complete reports (n = 1), or were considered by the original authors to be a tentative pilot investigation (n = 1).

with validity coefficients for medical tests. Because this table contains a large amount of information, we urge readers to closely examine the results before reading further.

A thorough inspection of Table 2 suggests four observations. First, both psychological and medical tests have varying degrees of validity, ranging from tests that are essentially uninformative for a given criterion (e.g., the Minnesota Multiphasic Personality Inventory [MMPI] Ego Strength scale [Table 2, Entry 5] or the dexamethasone suppression test [Table 2, Entry 1] for predicting response to treatment) to tests that are strongly predictive of appropriate criteria (e.g., neuropsychological tests for differentiating dementia from normal cognitive functioning [Table 2, Entry 137], computed tomography [CT] for detecting metastases from head and neck cancer [Table 2, Entry 136]).

Second, validity coefficients for many psychological tests are indistinguishable from those observed for many medical tests. For instance, when considering validity coefficients in the .30–.50 range, one finds results from the MMPI (Table 2, Entries 94, 99, 100, & 114), Millon Clinical Multiaxial Inventory (Table 2, Entry 93), Thematic Apperception Test (TAT; Table 2, Entries 90 & 95), Rorschach (Table 2, Entries 86, 89, 90, 95, & 111), Hare Psychopathy Checklist (Table 2, Entry 84), various neuropsychological and cognitive tests (Table 2, Entries 75, 76, 81, 83, 101, 103, 113, & 122), and the impact of psychological assessment feedback on the subsequent well-being of patients (Table 2, Entry 77). One also finds results from electrocardiograms (Table 2, Entry 72), CT (Table 2, Entries 79, 82, & 104), mammography (Table 2, Entry 80), magnetic resonance imaging (MRI; Table 2, Entry 107), ultrasound (Table 2, Entry 98), dental radiographs (Table 2, Entries 88, 108, & 112), Papanicolaou (Pap) smears (Table 2, Entry 87), cardiac fluoroscopy (Table 2, Entry 109), single photon emission computed tomography (Table 2, Entry 116), technetium bone scanning (Table 2, Entry 118), and serum cholesterol levels (Table 2, Entry 121).

At the upper end of Table 2, one generally sees results from studies in which the experimental design helped to increase effect size magnitudes. Of the 22 coefficients above .50, 19 are larger than the effects likely to be found in applied clinical practice. Most often (in 17 cases), this was because the condition to be detected by the test (e.g., peripheral artery disease, impaired kidney function, malingering) occurred much more often in the research studies than it would in actual practice (Finn & Kamphuis, 1995; Lijmer et al., 1999). In another instance, tests from the same method family as the predictor were used occasionally as validation criteria (Table 2, Entry 131), and in a final instance, it appears the author may have excluded studies when results were not as expected (Table 2, Entry 141). Despite these factors, what is most salient for our purpose is the difficulty one has in distinguishing psychological test validity from medical test validity. For instance, the ability to detect dementia is at least as good with neuropsychological tests (r = .68; Table 2, Entry 137) as it is with MRI (r = .57; Table 2, Entry 130).

At the low end of the validity range, one generally sees results from studies that should produce low associations. These include studies that (a) evaluate the impact of testing on a subsequent outcome variable (e.g., ultrasound on pregnancy outcome, Table 2, Entries 3, 4, & 6; geriatric medical assessment on reduced deaths, Table 2, Entry 9), (b) use tests to screen for rare conditions (e.g., triple marker screening for Trisomy 18, Table 2, Entry 8), or (c) use tests to predict rare outcome events (e.g., hopelessness for predicting suicide, Table 2, Entry 15). Once again, however, even at these lower values, psychological test validity is difficult to distinguish from medical test validity. For instance, the MMPI, Rorschach, and ventilatory lung function test all have roughly equal validity coefficients (rs = .05–.07; Table 2, Entries 10–12) for the difficult task of predicting cancer 2 to 3 decades later.

As a third general observation, our review does not reveal uniformly superior or uniformly inferior methods of psychological assessment. Despite the perceptions held by some, assessments with the Rorschach and TAT do not produce consistently lower validity coefficients than alternative personality tests. Instead, performance tests of cognitive ability, performance tests of personality (e.g., Rorschach, TAT), and self-report tests of personality all produce a range of validity coefficients that vary largely as a function of the criterion under consideration.[8]

Fourth, the findings indicate that psychological tests often generate substantial effect sizes. In particular, the validity coefficients found for psychological tests frequently exceed the coefficients found for many of the medical and psychological interventions listed in Table 1.

Taken together, the extensive array of findings in Table 2 offers compelling support for the value of psychological testing and assessment. To the extent that health care administrators differentially limit reimbursement for psychological tests relative to medical tests, such actions are not justifiable on the basis of a broad overview of the empirical evidence.

(*text continues on page 143*)

[8] Technically, it is not appropriate to compare validity coefficients across the types of tests presented in Table 2. As our notes to the table indicate, we did not report every coefficient obtained from each meta-analysis, some meta-analyses contributed more than one coefficient to the table, and at times, results from more than one meta-analysis were combined into a single value for the table. Furthermore, we made no effort to correct for design features that may have caused effect sizes to vary, and the table presents a vast array of nonequivalent criterion measures and validation tasks. Nonetheless, we realize that some readers may still wonder if differences exist within Table 2. Keeping in mind how the analysis is not strictly warranted, we used a random effects model and looked for differences across types of tests using the studies that were identified in our meta-analytic search. There were no significant differences at a global level, $F(4, 128) = 1.96$, $p > .05$, or when pairwise differences were examined with post hoc Scheffé tests. The unweighted means rs were as follows: Self-report personality tests = .24 (SD = .18, n = 24), performance personality tests (i.e., Rorschach, apperceptive storytelling tasks, sentence completion) = .33 (SD = .09, n = 8), cognitive or neuropsychological tests = .34 (SD = .17, n = 26), other psychological tests (e.g., observer ratings) = .30 (SD = .08, n = 7), and medical tests = .36 (SD = .21, n = 63).

Table 2
Examples of Testing and Assessment Validity Coefficients With an Emphasis on Meta-Analytic Results

Predictor and criterion (study and notes)	*r*	*N*
1. Dexamethasone suppression test scores and response to depression treatment (Ribeiro, Tandon, Grunhaus, & Greden, 1993).[a]	.00	2,068
2. Fecal occult blood test screening and reduced death from colorectal cancer (Towler et al., 1998).	.01	329,642
3. Routine umbilical artery Doppler ultrasound and reduced perinatal deaths in low-risk women (Goffinet, Paris-Llado, Nisand, & Bréart, 1997; the authors also examined the impact of routine umbilical artery ultrasound on 13 other measures of successful outcome. The average effect size across these other criteria was $r = -.0036$ [*ns* from 6,373 to 11,375], with the largest correlation in the expected direction being .0097 [for Apgar scores at 5 minutes]).	.01	11,375
4. Routine ultrasound examinations and successful pregnancy outcomes (Bucher & Schmidt, 1993; outcomes considered were live births [$r = .0009$], no induced labor [$r = .0176$], no low Apgar scores [$r = -.0067$], no miscarriages [$r = .0054$], and no perinatal mortality [$r = .0168$]).	.01	16,227
5. MMPI Ego Strength scores and subsequent psychotherapy outcome (Meyer & Handler, 1997; this meta-analysis considered only studies in which the Ego Strength scale was used along with the Rorschach PRS).	.02	280
6. Routine umbilical artery Doppler ultrasound and reduced perinatal deaths in high-risk women (Alfirevic & Neilson, 1995; the authors also examined the impact of routine umbilical artery ultrasound on 19 other measures of successful outcome. The average effect size across these other criteria was $r = .018$ [*ns* from 476 to 7,474]).	.03	7,474
7. Denial/repressive coping style and development of breast cancer (McKenna, Zevon, Corn, & Rounds, 1999; weighted effect size computed from the study data in their Table 1).	.03	12,908
8. Triple marker[b] prenatal screening of maternal serum and identification of Trisomy 18 (Yankowitz, Fulton, Williamson, Grant, & Budelier, 1998).[c]	.03	40,748
9. Impact of geriatric medical assessment teams on reduced deaths (data combined from the meta-analysis by Rubenstein, Stuck, Siu, & Wieland, 1991, and the following more recent studies: Boult et al., 1994; Büla et al., 1999; Burns, Nichols, Graney, & Cloar, 1995; Englehardt et al., 1996; Fabacher et al., 1994; Fretwell et al., 1990; Germain, Knoeffel, Wieland, & Rubenstein, 1995; Hansen, Poulsen, & Sørensen, 1995; Harris et el., 1991; Karppi & Tilvis, 1995; Naughton, Moran, Feinglass, Falconer, & Williams, 1994; Reuben et la., 1995; Rubenstein, Josephson, Harker, Miller, & Wieland, 1995; Rubin, Sizemore, Loftis, & de Mola, 1993; Silverman et al., 1995; Siu et al., 1996; Thomas, Brahan, & Haywood, 1993; and Trentini et al., 1995; only the latest available outcome data were used for each sample).	.04	10,065
10. MMPI depression profile scores and subsequent cancer within 20 years (Persky, Kempthorne-Rawson, & Shekelle, 1987).[c]	.05	2,018
11. Ventilatory lung function test scores and subsequent lung cancer within 25 years (Islam & Schottenfeld, 1994).[c]	.06	3,956
12. Rorschach Interaction Scale scores and subsequent cancer within 30 years (Graves, Phil, Mead & Pearson, 1986; scores remained significant predictors after controlling for baseline smoking, serum cholesterol, systolic blood pressure, weight, and age).[c]	.07	1,027
13. Unique contribution of an MMPI high-point code (vs. other codes) to conceptually relevant criteria (McGrath & Ingersoll, 1999a, 1999b).	.07	8,614
14. MMPI scores and subsequent prison misconduct (Gendreau, Goggin, & Law, 1997).	.07	17,636
15. Beck Hopelessness Scale scores and subsequent suicide (data combined from Beck, Brown, Berchick, Stewart, & Steer, 1990; and Beck, Steer, Kovacs, & Garrison, 1985).[c]	.08	2,123
16. MMPI elevations on Scales F, 6, or 8 and criminal defendant incompetency (Nicholson & Kugler, 1991).	.08	1,461
17. Extraversion test scores and success in sales (concurrent and predictive; data combined from Barrick & Mount, 1991, Table 2; Salgado, 1997, Table 3; and Vinchur, Schippman, Switzer, & Roth, 1998 [coefficients from their Tables 2 and 3 were averaged, and the largest *N* was used for the overall sample size]).	.08	6,004
18. Attention and concentration test scores and residual mild head trauma (Binder, Rohling, & Larrabee, 1997).	.09	622
19. In cervical cancer, lack of glandular differentiation on tissue biopsy and survival past 5 years (Heatley, 1999; this study reported two meta-analyses. The other one found that nuclear DNA content was of no value for predicting cancer progression in initially low-grade cervical intraepithelial neoplasia).	.11	685

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
20. Negative emotionality test scores and subsequent heart disease (Booth-Kewley & Friedman, 1987; data were derived from their Table 7, with negative emotionality defined by the weighted effect for anger/hostility/aggression, depression, and anxiety).	.11	(*k* = 11)
21. Triple marker[b] prenatal screening of maternal serum and identification of Down's syndrome (Conde-Agudelo & Kafury-Goeta, 1998; results were reported across all ages).	.11	194,326
22. General cognitive ability and involvement in automobile accidents (Arthur, Barrett, & Alexander, 1991).	.12	1,020
23. Conscientiousness test scores and job proficiency (concurrent and predictive; data combined from Barrick & Mount, 1991, Table 3; Mount, Barrick, & Stewart, 1998; Salgado, 1998, Table 1; and Vinchur et al., 1998 [coefficients from their Tables 2 and 3 were averaged, and the largest *N* was used for the overall sample size]).	.12	21,650
24. Platform posturography and detection of balance deficits due to vestibular impairment (Di Fabio, 1996).	.13	1,477
25. General intelligence and success in military pilot training (Martinussen, 1996).	.13	15,403
26. Self-report scores of achievement motivation and spontaneous achievement behavior (Spangler, 1992; coefficient derived from the weighted average of the semioperant and operant criterion data reported in Spangler's Table 2).	.15	(*k* = 104)
27. Graduate Record Exam Verbal or Quantitative scores and subsequent graduate GPA in psychology (E. L. Goldberg & Alliger, 1992).	.15	963
28. Low serotonin metabolites in cerebrospinal fluid (5-HIAA) and subsequent suicide attempts (Lester, 1995).	.16	140
29. Personality tests and conceptually meaningful job performance criteria (data combined from Robertson & Kinder, 1993; Tett, Jackson, & Rothstein, 1991; and Tett, Jackson, Rothstein, & Reddon, 1994; we used the single scale predictors from Robertson & Kinder [their Table 3] and the confirmatory results from Table 1 in Tett et al., 1994).	.16	11,101
30. Implicit memory tests and differentiation of normal cognitive ability from dementia (Meiran & Jelicic, 1995).	.16	1,156
31. MMPI Cook–Medley hostility scale elevations and subsequent death from all causes (T. Q. Miller, Smith, Turner, Guijarro, & Hallet, 1996; data were drawn from their Table 6).	.16	4,747
32. Motivation to manage from the Miner Sentence Completion Test and managerial effectiveness (Carson & Gilliard, 1993; results were averaged across the three performance criterion measures of managerial success. Because the three criterion measures were not independent across studies, the *N* reported is the largest *N* used for any single criterion).	.17	2,151
33. Extraversion and subjective well-being (DeNeve & Cooper, 1998).	.17	10,364
34. MRI T_2 hyperintensities and differentiation of affective disorder patients from healthy controls (Videbech, 1997; data from Videbech's Tables 1 and 2 were combined, but only those statistics used by the original author are included here).	.17	1,575
35. Test anxiety scales and lower school grades (Hembree, 1988; reported effect is the average effect size for the course grade and GPA data from Hembree's Table 1. Participants were assumed to be independent across studies).	.17	5,750
36. High trait anger assessed in an interpersonal analogue and elevated blood pressure (Jorgensen, Johnson, Kolodziej, & Schreer, 1996; data come from the "Overall" column of their Table 4).	.18	(*k* = 34)
37. Reduced blood flow and subsequent thrombosis or failure of synthetic hemodialysis graft (Paulson, Ram, Birk, & Work, 1999).	.18	4,569
38. MMPI validity scales and detection of known or suspected underreported psychopathology (Baer, Wetter, & Berry, 1992; weighted average effect size was calculated from data reported in their Table 1 for all studies using participants presumed to be underreporting).	.18	328
39. Dexamethasone suppression test scores and subsequent suicide (Lester, 1992).	.19	626
40. Short-term memory tests and subsequent job performance (Verive & McDaniel, 1996).	.19	17,741
41. Depression test scores and subsequent recurrence of herpes simplex virus symptoms (Zorrilla, McKay, Luborsky, & Schmidt, 1996; effect size is for prospective studies).	.20	333
42. Four preoperative cardiac tests and prediction of death or MI within 1 week of vascular surgery (Mantha et al., 1994; the four tests considered were dipyridamole-thallium scintigraphy, ejection fraction estimation by radionuclide ventriculography, ambulatory ECG, and dobutamine stress ECG. The authors concluded no test was conclusively superior to the others).	.20	1,991
43. Scholastic Aptitude Test scores and subsequent college GPA (Baron & Norman, 1992).[c]	.20	3,816

(*table continues*)

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
44. Self-reported dependency test scores and physical illness (Bornstein, 1998; weighted effect size was calculated from the retrospective studies reported in Bornstein's Table 1 [Studies 3, 5, 7, 8, 13, and 19] and the prospective studies listed in Bornstein's Table 2 [Studies 1–4]).	.21	1,034
45. Dexamethasone suppression test scores and psychotic vs. nonpsychotic major depression (Nelson & Davis, 1997; effect size calculated from the weighted effects for the individual studies in their Table 1).	.22	984
46. Traditional ECG stress test results and coronary artery disease (Fleischmann, Hunink, Kuntz, & Douglas, 1998; results were estimated from the reported sensitivity and specificity in conjunction with the base rate of coronary artery disease and the total independent *N* across studies).	.22	5,431
47. Graduate Record Exam Quantitative scores and subsequent graduate GPA (Morrison & Morrison, 1995).	.22	5,186
48. TAT scores of achievement motivation and spontaneous achievement behavior (Spangler, 1992; coefficient was derived from the weighted average of the semioperant and operant criterion data in Spangler's Table 2).	.22	(*k* = 82)
49. Isometric strength test scores and job ratings of physical ability (Blakley, Quiñones, & Crawford, 1994).	.23	1,364
50. Single serum progesterone testing and diagnosis of ectopic pregnancy (Mol, Lijmer, Ankum, van der Veen, & Bossuyt, 1998; following the original authors, we used only the 18 prospective or retrospective cohort studies listed in their Table III).	.23	6,742
51. Cognitive multitask performance test scores and subsequent pilot proficiency (Damos, 1993).	.23	6,920
52. WISC distractibility subscales and learning disability diagnoses (Kavale & Forness, 1984; the effect sizes from this meta-analysis are likely to be underestimates because the authors computed the average effect for individual test scales rather than the effect for a composite pattern).	.24	(*K* = 54)
53. Fetal fibronectin testing and prediction of preterm delivery (Faron, Boulvain, Irion, Bernard, & Fraser, 1998; data were aggregated across low- and high-risk populations and across designs with single or repeated testing for all studies using delivery before 37 weeks as the criterion).	.24	7,900
54. Decreased bone mineral density and lifetime risk of hip fracture in women (Marshall, Johnell, & Wedel, 1996; the results were restricted to those from absorptiometry using single or dual energy, photon, or X-ray; quantitative CT; quantitative MRI; or ultrasound scanning. The overall effect was estimated from their Table 3 using a total lifetime incidence of 15%; the effect would be smaller if the lifetime risk incidence was lower [e.g., if the incidence were 3%, the effect would be $r = .13$]. Total *N* was derived from the *n* for each study in their Table 1 reporting the incidence of hip fractures).	.25	20,849
55. General intelligence test scores and functional effectiveness across jobs (Schmitt, Gooding, Noe, & Kirsch, 1984; data were obtained from their Table 4).	.25	40,230
56. Internal locus of control and subjective well-being (DeNeve & Cooper, 1998).	.25	8,481
57. Integrity test scores and subsequent supervisory ratings of job performance (Ones, Viswesvaran, & Schmidt, 1993; effect size was taken from the "predictive-applicant" cell of their Table 8).	.25	7,550
58. Self-reported dependency test scores and dependent behavior (Bornstein, 1999; coefficient was derived from all results listed in Bornstein's Table 1 as reported in his footnote 8).	.26	3,013
59. Self-efficacy appraisals and health-related treatment outcomes (Holden, 1991).	.26	3,527
60. Elevated Jenkins Activity Survey scores and heart rate and blood pressure reactivity (Lyness, 1993; the effect size reflects the average reactivity for heart rate, systolic blood pressure, and diastolic blood pressure as reported in Lyness's Table 6. It was assumed that overlapping studies contributed to each of these criterion estimates, so *k* was estimated as the largest number of effect sizes contributing to a single criterion measure).	.26	(*k* = 44)
61. Combined internal, stable, and global attributions for negative event outcomes and depression (Sweeney, Anderson, & Bailey, 1986; only the finding that dealt with the composite measure of attributions and negative outcome was included. Coefficients were lower for positive outcomes and for single types of attributions [e.g., internal]).	.27	5,788
62. Neuroticism and decreased subjective well-being (DeNeve & Cooper, 1998).	.27	9,777
63. Screening mammogram results and detection of breast cancer within 2 years (Mushlin, Kouides, & Shapiro, 1998).	.27	192,009

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
64. Microbiologic blood culture tests to detect bloodstream infection from vascular catheters (Siegman-Igra et al., 1997; only results from studies without criterion contamination were summarized [see Siegman-Igra et al., 1997, pp. 933–934]).	.28	1,354
65. C-reactive protein test results and diagnosis of acute appendicitis (Hallan & Åsberg, 1997; mean weighted effect size was derived from data in their Table 1, excluding two studies that did not use histology as the validating criteria and one study that did not report the prevalence of appendicitis).	.28	3,338
66. Graduate Record Exam Verbal scores and subsequent graduate GPA (Morrison & Morrison, 1995).	.28	5,186
67. Hare Psychopathy Checklist scores and subsequent criminal recidivism (Salekin, Rogers, & Sewell, 1996; only effects for predictive studies were summarized).	.28	1,605
68. Short-term memory tests and subsequent performance on job training (Verive & McDaniel, 1996).	.28	16,521
69. Cranial ultrasound results in preterm infants and subsequent developmental disabilities (Ng & Dear, 1990).	.29	1,604
70. Serum CA-125 testing and detection of endometriosis (Mol, Bayram, et al., 1998).	.29	2,811
71. Neuropsychological test scores and differentiation of patients with multiple sclerosis (Wishart & Sharpe, 1997).	.29	(*k* = 322)
72. For women, ECG stress test results and detection of coronary artery disease (Kwok, Kim, Grady, Segal, & Redberg, 1999; our *N* was obtained from their Table 1. It differs from the *N* reported by the authors [3,872 vs. 3,721], though it is not clear what would account for this difference. Although the article also examined the thallium stress test and the exercise ECG, there was not sufficient data for us to generate effect sizes for these measures).	.30	3,872
73. YASR total problems and psychiatric referral status (receiving treatment vs. not; Achenbach, 1997; effect size was estimated from data in Part 1 of Achenbach's Table 7.5. Because the percentages listed in this table were too imprecise to accurately generate effect size estimates, all possible 2 × 2 tables that would match the given percentages were generated. Subsequently, the effect size was obtained from those 2 × 2 tables that also produced odds ratios that exactly matched the odds ratios reported in the text. When rounded to two decimal places, all appropriate 2 × 2 tables produced the same effect size. The effect size compares the self-reports of young adults in treatment with the self-reports of demographically matched controls who were not receiving treatment).[c]	.30	1,142
74. Fecal leukocyte results and detection of acute infectious diarrhea (Huicho, Campos, Rivera, & Guerrant, 1996; results are reported for the most studied test [*K* = 19]. For the remaining tests, effect sizes could be generated for only two small studies of fecal lactoferrin, and the average results for occult blood tests were lower [*r* = .26; *K* = 7]).	.30	7,132
75. Neuropsychological test scores and differentiation of learning disabilities (Kavale & Nye, 1985; we report the results for neuropsychological functioning because it was studied most frequently).	.30	(*K* = 394)
76. Continuous performance test scores and differentiation of ADHD and control children (Losier, McGrath, & Klein, 1996; overall sample weighted effect was derived by combining the omission and commission data reported in their Tables 7 and 8).	.31	720
77. Effects of psychological assessment feedback on subsequent patient well-being (coefficient combined the follow-up data reported in Finn & Tonsager, 1992; and Newman & Greenway, 1997).[c]	.31	120
78. Expressed emotion on the CFI and subsequent relapse in schizophrenia and mood disorders (Butzlaff & Hooley, 1998).	.32	1,737
79. CT results and detection of aortic injury (Mirvis, Shanmuganathan, Miller, White, & Turney, 1996; from the information provided, an effect size could not be computed for two studies included in this meta-analysis).	.32	3,579
80. Screening mammogram results and detection of breast cancer within 1 year (Mushlin, Kouides, & Shapiro, 1998; overall effect size includes studies that combined mammography with clinical breast examination).	.32	263,359
81. Halstead–Reitan Neuropsychological Tests and differentiation of impaired vs. control children (Forster & Leckliter, 1994; the reported weighted effect size is slightly inflated because some observations were based on group differences relative to the control group standard deviation [rather than the pooled standard deviation]. When possible, effect sizes were computed directly from the data reported in their Tables 1 and 2. The reported *N* indicates the total number of independent observations across studies).	.33	858

(*table continues*)

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
82. CT results for enlarged ventricular volume and differentiation of schizophrenia from controls (Raz & Raz, 1990).	.33	(*k* = 53)
83. Long-term memory test scores and diagnosis of multiple sclerosis (Thornton & Raz, 1997; effect size was obtained from their Table 2 with the outlier study excluded).	.33	(*K* = 33)
84. Hare Psychopathy Checklist scores and subsequent violent behavior (Salekin, Rogers, & Sewell, 1996; only effects for predictive studies were summarized).	.33	1,567
85. Alanine aminotransferase results and detection of improved liver function in hepatitis C patients (Bonis, Ioannidis, Cappelleri, Kaplan, & Lau, 1997; data reflect the criterion of any histologically identified improvement).	.34	480
86. Rorschach scores and conceptually meaningful criterion measures (data combined from Atkinson, 1986, Table 1 [*K* = 79]; Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999, Table 4 [*K* = 30]; and K. P. Parker, Hanson, & Hunsley, 1988, Table 2 [*K* = 14]. Hiller et al. expressed concern that Atkinson's and K. P. Parker et al.'s effect size estimates may have been inflated by some results derived from unfocused *F* tests [i.e., with >1 *df* in the numerator]. However, Atkinson excluded effects based on *F*, and K. P. Parker et al.'s average effect size actually increased when *F* test results were excluded. Recently, Garb, Florio, & Grove, 1998, conducted reanalyses of K. P. Parker et al.'s data. Although these reanalyses have been criticized [see K. P. Parker, Hunsley, & Hanson, 1999], if the results from Garb et al.'s first, second, or third analysis were used in lieu of those from K. P. Parker et al., the synthesized results reported here would change by −.0096, −.0036, or −.0007, respectively, for the Rorschach and by .0203, .0288, or .0288, respectively, for the MMPI [see Entry 100, this table]).	.35	(*K* = 122)
87. Papanicolaou Test (Pap smear) and detection of cervical abnormalities (Fahey, Irwig, & Macaskill, 1995; overall weighted effect calculated from data reported in their Appendix 1).	.36	17,421
88. Conventional dental X-rays and diagnosis of biting surface cavities (occlusal caries; Ie & Verdonschot, 1994; the overall weighted effect was derived from all the studies listed in their Table 1. In each case, the original citations were obtained, and raw effect sizes were calculated from the initial study).	.36	5,466
89. Incremental contribution of Rorschach PRS scores over IQ to predict psychotherapy outcome (Meyer, 2000).	.36	290
90. Rorschach or Apperceptive Test Dependency scores and physical illness (Bornstein, 1998; weighted effect size was calculated from the retrospective studies reported in Bornstein's Table 1 [Studies 1, 11, 14–16, and 18]. No prospective studies used these types of scales as predictors).	.36	325
91. Assessment center evaluations and job success (data combined from Schmitt, Gooding, Noe, & Kirsch, 1984; and Gaugler, Rosenthal, Thornton, & Bentson, 1987; the overall effect size was derived from the sample weighted average reported in each study. Although Schmitt et al.'s study was conducted earlier than Gaugler et al.'s, they relied on a larger *N*. Because each meta-analysis undoubtedly relied on some common studies, the *N* reported here is from Schmitt et al.).	.37	15,345
92. Competency screening sentence-completion test scores and defendant competency (Nicholson & Kugler, 1991).	.37	627
93. MCMI–II scale score and average ability to detect depressive or psychotic disorders (Ganellen, 1996; each individual study contributed one effect size averaged across diagnostic criteria and type of predictor scales [single vs. multiple scales]. Results were averaged across analyses reported in different publications using the same sample. Although Ganellen reported larger effect sizes for studies that used multiscale predictors, these studies relied on unreplicated multivariate predictor equations. As such, multiscale predictors were averaged with hypothesized, single-scale predictors).[c]	.37	575
94. MMPI scale scores and average ability to detect depressive or psychotic disorders (Ganellen, 1996; see Entry 93, this table).[c]	.37	927
95. Rorschach Apperceptive Test Dependency scores and dependent behavior (Bornstein, 1999; coefficient was derived from all results listed in Bornstein's Table 1 as reported in his footnote 8).	.37	1,808
96. Accuracy of home pregnancy test kits in patients conducting testing at home (Bastian, Nanda, Hasselblad, & Simel, 1998; results derived from the pooled "effectiveness score," which was described and thus treated as equivalent to Cohen's *d*. Also, findings were very different when tests were evaluated using researcher-assisted volunteers rather than actual patients [*r* = .81; *N* = 465]).	.38	155

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
97. Sperm penetration assay results and success with in vitro fertilization (Mol, Meijer, et al., 1998).	.39	1,335
98. Endovaginal ultrasound in postmenopausal women and detection of endometrial cancer (Smith-Bindman et al., 1998; effect size was derived from the authors' pooled results [their Table 2] using their recommended cutoff of 5 mm to define endometrial thickening).	.39	3,443
99. MMPI Validity scales and detection of underreported psychopathology (primarily analogue studies; Baer, Wetter, & Berry, 1992; weighted average effect size calculated from data in their Table 1).	.39	2,297
100. MMPI scores and conceptually meaningful criterion measures (data combined from Atkinson, 1986, Table 1; Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999, Table 4; and K. P. Parker, Hanson, & Hunsley, 1988, Table 2. See also Entry 86, this table).	.39	(*K* = 138)
101. Neuropsychologists' test-based judgments and presence/absence of impairment (Garb & Schramke, 1996; coefficient was calculated from the accuracy of judgments relative to base rates [see Garb & Schramke, 1996, pp. 143, 144–145]).	.40	2,235
102. Prostate-specific antigen and estimated detection of prostate cancer for men aged 60–70 (Aziz & Barathur, 1993).	.40	4,200
103. Short-term verbal learning and differentiation of major depression from controls (Veiel, 1997; although the author reported many effect sizes, we report the variable that was studied most often).	.41	(*K* = 10)
104. CT results and detection of lymph node metastases in cervical cancer (Scheidler, Hricak, Yu, Subak, & Segal, 1997; an effect size could not be computed for one study included in this meta-analysis).	.41	1,022
105. Dissociative Experiences Scale scores and detection of MPD or PTSD vs. controls (Van IJzendoorn & Schuengel, 1996; we assumed the *N*s for both criterion diagnoses were not independent, so the reported *N* is that for the largest analysis).	.41	1,705
106. Colposcopy and detection of normal/low-grade SIL vs. high-grade SIL/cancer of the cervix (Mitchell, Schottenfeld, Tortolero-Luna, Cantor, & Richards-Kortum, 1998; effect sizes were calculated from data reported in their Table 3).	.42	2,249
107. Cortical tuber count on MRI and degree of impaired cognitive development in tuberous sclerosis (M. Goodman et al., 1997).	.43	157
108. Conventional dental X-rays and diagnosis of between-tooth cavities (approximal caries; Van Rijkom & Verdonschot, 1995; this is an unweighted effect size for all studies that used a "strong" validity criterion [i.e., microradiography, histology, or cavity preparation]).	.43	(*K* = 8)
109. Cardiac fluoroscopy and diagnosis of coronary artery disease (Gianrossi, Detrano, Colombo, & Froelicher, 1990).	.43	3,765
110. Serum chlamydia antibody levels and detection of fertility problems due to tubal pathology (Mol et al., 1997; only the results for the optimal predictor assays and optimal criterion measures are presented).	.44	2,131
111. Rorschach PRS scores and subsequent psychotherapy outcome (Meyer & Handler, 1997, 2000).	.44	783
112. Digitally enhanced dental X-rays and diagnosis of biting surfaces cavities (Ie & Verdonschot, 1994; the overall weighted effect size was derived from all the studies listed in their Table 1. In each case, the original citations were obtained, and raw effect sizes were calculated from the initial study).	.44	2,870
113. WAIS IQ and obtained level of education (Hanson, Hunsley, & Parker, 1988).	.44	(*k* = 9)
114. MMPI Validity scales and detection of known or suspected malingered psychopathology (data combined from Berry, Baer, & Harris, 1991; and Rogers, Sewell, & Salekin, 1994; the average weighted effect size was calculated from data presented in Tables 1 and 2 of Berry et al. and Table 1 of Rogers et al. for participants presumed or judged to be malingering disturbance).	.45	771
115. D-dimer blood test results and detection of deep vein thrombosis or pulmonary embolism (Becker, Philbrick, Bachhuber, & Humphries, 1996; results are reported for only the 13 [of 29] studies with stronger methodology).	.45	1,652
116. Exercise SPECT imaging and identification of coronary artery disease (Fleischmann, Hunink, Kuntz, & Douglas, 1998; results were estimated from the reported sensitivity and specificity in conjunction with the base rate of coronary artery disease and the total independent *N* across studies).	.46	3,237
117. Antineutrophil cytoplasmic antibody testing and detection of Wegener's granulomatosis (Rao et al., 1995; sensitivity for each study was estimated from their Figure 1).	.47	13,562

(*table continues*)

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
118. Technetium bone scanning results and detection of osteomyelitis (bone infection; Littenberg, Mushlin, & the Diagnostic Technology Assessment Consortium, 1992).	.48	255
119. Clinical examination with routine lab tests and detection of metastatic lung cancer (Silvestri, Littenberg, & Colice, 1995).	.48	1,593
120. Lecithin/sphingomyelin ratio and prediction of neonatal respiratory distress syndrome (Petersen, Smith, Okorodudu, & Bissell, 1996; the most frequently studied predictor test was reported).	.50	1,170
121. Sensitivity of total serum cholesterol levels to changes in dietary cholesterol (Howell, McNamara, Tosca, Smith, & Gaines, 1997).	.50	(*k* = 307)
122. Memory recall tests and differentiation of schizophrenia from controls (Aleman, Hijman, de Haan, & Kahn, 1999; effect size is for studies with demographically matched comparison participants).	.50	2,290
123. CBCL parent report of total problems and psychiatric referral status (receiving treatment vs. not; Achenbach, 1991b; raw data to generate this effect size were obtained from Thomas M. Achenbach [personal communication, February 5, 1999]. Coefficient compares parent ratings of children in treatment to parent ratings of demographically matched control children not receiving treatment).[c]	.51	4,220
124. WAIS IQ subtests and differentiation of dementia from controls (H. Christensen & Mackinnon, 1992; effect computed from data presented in their Tables 1 and 2. The reported *N* is for the largest sample across the individual subtest comparisons).	.52	516
125. Single serum progesterone testing and diagnosis of any nonviable pregnancy (Mol, Lijmer, et al., 1998; following the original authors, we used only the 10 prospective cohort studies listed in their Table II).	.52	3,804
126. MRI results and detection of ruptured silicone gel breast implants (C. M. Goodman, Cohen, Thornby, & Netscher, 1998; these authors found that mammography [r = .21, N = 381] and ultrasound [r = .42, N = 541] were less effective than MRI).	.53	382
127. Association of Hachinski ischemic scores with postmortem classification of dementia type (Moroney et al., 1997; effect size computed from their Figure 1 using continuous scores and the Alzheimer's, mixed, and multiinfarct group classifications on a continuum).	.55	312
128. MRI results and detection of lymph node metastases in cervical cancer (Scheidler, Hricak, Yu, Subak, & Segal, 1997; an effect size could not be computed for one study included in this meta-analysis).	.55	817
129. Cognitive tests of information-processing speed and reasoning ability (Verhaeghen & Salthouse, 1997).	.55	4,026
130. MRI results and differentiation of dementia from controls (Zakzanis, 1998; PET and SPECT findings from this meta-analysis were slightly less valid or based on smaller samples, so are not reported. Neuropsychological findings were not used because D. Christensen, Hadzi-Povlovic, & Jacomb, 1991, reported a more extensive meta-analysis).	.57	374
131. WAIS IQ scores and conceptually meaningful criterion measures (K. P. Parker, Hanson, & Hunsley, 1988, Table 2; Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999, expressed concern about K. P. Parker et al.'s results because some effect sizes came from unfocused F tests [i.e., >1 df in the numerator], though the overall effect increases when these results are excluded).	.57	(*K* = 39)
132. Exercise ECG results and identification of coronary artery disease (Fleischmann, Hunink, Kuntz, & Douglas, 1998; results were estimated from the reported sensitivity and specificity in conjunction with the base rate of coronary artery disease and the total independent *N* across studies).	.58	2,637
133. Ultrasound results and identification of deep venous thrombosis (Wells, Lensing, Davidson, Prins, & Hirsh, 1995).	.60	1,616
134. Neuropsychologists' test-based judgments and presence/localization of impairment (Garb & Schramke, 1996; effect size calculated from the accuracy of judgments relative to base rates [see Garb & Schramke, 1996, pp. 143, 144–145]).	.60	1,606
135. Long-term verbal memory tests and differentiation of dementia from depression (H. Christensen, Griffiths, MacKinnon, & Jacomb, 1997; effect data taken from their Table 4).	.61	(*K* = 32)
136. CT results and detection of metastases from head and neck cancer (Merrit, Williams, James, & Porubsky, 1997; *N* was obtained from the original studies).	.64	517

Table 2 (*continued*)

Predictor and criterion (study and notes)	*r*	*N*
137. Neuropsychological tests and differentiation of dementia from controls (D. Christensen, Hadzi-Pavlovic, & Jacomb, 1991; the effect size was derived from studies explicitly stating that dementia had been diagnosed independent of the neuropsychological test results [see D. Christensen et al., 1991, p. 150]).	.68	(*k* = 94)
138. Immunoglobulin-G antiperinuclear factor scores and detection of rheumatoid arthritis (Berthelot, Garnier, Glémarec, & Flipo, 1998).	.68	2,541
139. MMPI Validity scales and detection of malingered psychopathology (primarily analogue studies; data combined from Berry, Baer, & Harris, 1991; and Rogers, Sewell, & Salekin, 1994; average weighted effect size calculated from Tables 1 and 2 of Berry et al. and Table 1 of Rogers et al.).	.74	11,204
140. MMPI basic scales: booklet vs. computerized form (Finger & Ones, 1999; the alternate forms reliability coefficients for each scale were weighted by sample size [*n*s from 508 to 872], and the average *N* is reported).	.78	732
141. Thoracic impedance scores and criterion measures of cardiac stroke volume and output (Fuller, 1992; only data from methodologically "adequate" studies were included. The mean weighted correlation for each criterion measure was weighted by the number of studies contributing to the mean and then averaged across all criterion measures. Because Fuller [1992, p. 105] cryptically stated that studies were excluded unless there was "concurrence of measurement between the two instruments being compared," it is possible that relevant studies were omitted when the findings did not support the hypothesis).	.81	(*K* = 24)
142. Creatinine clearance test results and kidney function (glomerual filtration rate; Campens & Buntinx, 1997; results for measured and estimated [by the Cockroft–Gault formula] creatinine clearance were pooled. The *N* reported in our table is slightly inflated because it was impossible to identify the specific *n* for two of the studies that used both measures).	.83	2,459
143. Duplex ultrasonography results and identification of peripheral artery disease (de Vries, Hunink, & Polak, 1996; weighted effect size derived from data in their Table 2 using patient samples. The reported *N* refers to the number of observations; some patients were tested multiple times).	.83	4,906
144. Finger or ear pulse oximetry readings in patients and arterial oxygen saturation (L. A. Jensen, Onyskiw, & Prasad, 1998).	.84	4,354

Note. ADHD = attention-deficit hyperactivity disorder; CBCL = Child Behavior Checklist; CFI = Camberwell Family Interview; CT = computed tomography; ECG = electrocardiogram; GPA = grade point average; IQ = intelligence quotient; *k* = number of effect sizes contributing to the mean estimate; *K* = number of studies contributing to the mean estimates; MCMI–II = Millon Clinical Multiaxial Inventory—2nd Edition; MMPI = Minnesota Multiphasic Personality Inventory; MPD = multiple personality disorder; MRI = magnetic resonance imaging; PET = positron emission tomography; PRS = Prognostic Rating Scale; PTSD = posttraumatic stress disorder; SIL = squamous intraepithelial lesions; SPECT = single photon emission computed tomography; TAT = Thematic Apperception Test; WAIS = Wechsler Adult Intelligence Scale; WISC = Wechsler Intelligence Scale for Children; YASR = Young Adult Self-Report.

[a] The actual effect was a statistically nonsignificant value of −.013 (i.e., in the direction of opposite of prediction). [b] Triple marker refers to the joint use of alpha-fetoprotein, human chorionic gonadotropin, and unconjugated estriol. [c] These results are not from meta-analyses and were not identified through our systematic literature search.

Distinctions Between Psychological Testing and Psychological Assessment

Psychological testing is a relatively straightforward process wherein a particular scale is administered to obtain a specific score. Subsequently, a descriptive meaning can be applied to the score on the basis of normative, nomothetic[9] findings. In contrast, psychological assessment is concerned with the clinician who takes a variety of test scores, generally obtained from multiple test methods, and considers the data in the context of history, referral information, and observed behavior to understand the person being evaluated, to answer the referral questions, and then to communicate findings to the patient, his or her significant others, and referral sources.

In psychological testing, the nomothetic meaning associated with a scaled score of 10 on the Arithmetic subtest from the Wechsler Adult Intelligence Scale—Third Edition (Wechsler, 1997) is that a person possesses average skills in mental calculations. In an idiographic assessment, the same score may have very different meanings. After considering all relevant information, this score may mean a patient with a recent head injury has had a precipitous decline in auditory attention span and the capacity to men-

[9] *Nomothetic* refers to general laws or principles. Nomothetic research typically studies the relationship among a limited number of characteristics across a large number of people. *Idiographic* refers to the intensive study of a single individual. Here, the focus is on how a large number of characteristics fit together uniquely within one person or in the context of a single life.

tally manipulate information. In a patient undergoing cognitive remediation for attentional problems secondary to a head injury, the same score may mean there has been a substantial recovery of cognitive functioning. In a third, otherwise very intelligent patient, a score of 10 may mean pronounced symptoms of anxiety and depression are impairing skills in active concentration. Thus, and consistent with Shea's (1985) observation that no clinical question can be answered solely by a test score, many different conditions can lead to an identical score on a particular test. The assessment task is to use test-derived sources of information in combination with historical data, presenting complaints, observations, interview results, and information from third parties to disentangle the competing possibilities (Eyde et al., 1993). The process is far from simple and requires a high degree of skill and sophistication to be implemented properly.

Distinctions Between Formal Assessment and Other Sources of Clinical Information

All mental health professionals assess patient problems. Almost universally, such evaluations rely on unstructured interviews and informal observations as the key sources of information about the patient. Although these methods can be efficient and effective ways to obtain data, they are also limited. When interviews are unstructured, clinicians may overlook certain areas of functioning and focus more exclusively on presenting complaints. When interviews are highly structured, clinicians can lose the forest for the trees and make precise but errant judgments (Hammond, 1996; Tucker, 1998). Such mistakes may occur when the clinician focuses on responses to specific interview questions (e.g., diagnostic criteria) without fully considering the salience of these responses in the patient's broader life context or without adequately recognizing how the individual responses fit together into a symptomatically coherent pattern (Arkes, 1981; Klein, Ouimette, Kelly, Ferro, & Riso, 1994; Perry, 1992).

Additional confounds derive from patients, who are often poor historians and/or biased presenters of information (see, e.g., John & Robins, 1994; Moffitt et al., 1997; Rogler, Malgady, & Tryon, 1992; Widom & Morris, 1997). For instance, neurologically impaired patients frequently lack awareness of their deficits or personality changes (Lezak, 1995), and response styles such as defensiveness or exaggeration affect the way patients are viewed by clinical interviewers or observers (see, e.g., Alterman et al., 1996; Pogge, Stokes, Frank, Wong, & Harvey, 1997). Defensive patients are seen as more healthy, whereas patients who exaggerate their distress are seen as more impaired. In contrast to less formal clinical methods, psychological testing can identify such biased self-presentation styles (see Entries 38, 99, 114, & 139 in Table 2), leading to a more accurate understanding of the patient's genuine difficulties.

There are several other ways that formal psychological assessment can circumvent problems associated with typical clinical interviews. First, psychological assessments generally measure a large number of personality, cognitive, or neuropsychological characteristics simultaneously. As a result, they are inclusive and often cover a range of functional domains, many of which might be overlooked during less formal evaluation procedures.

Second, psychological tests provide empirically quantified information, allowing for more precise measurement of patient characteristics than is usually obtained from interviews.

Third, psychological tests have standardized administration and scoring procedures. Because each patient is presented with a uniform stimulus that serves as a common yardstick to measure his or her characteristics, an experienced clinician has enhanced ability to detect subtle behavioral cues that may indicate psychological or neuropsychological complications (see, e.g., Lezak, 1995). Standardization also can reduce legal and ethical problems because it minimizes the prospect that unintended bias may adversely affect the patient. In less formal assessments, standardization is lacking, and the interaction between clinician and patient can vary considerably as a function of many factors.

Fourth, psychological tests are normed, permitting each patient to be compared with a relevant group of peers, which in turn allows the clinician to formulate refined inferences about strengths and limitations. Although clinicians using informal evaluation procedures generate their own internal standards over time, these are less systematic and are more likely to be skewed by the type of patients seen in a particular setting. Moreover, normed information accurately conveys how typical or unusual the patient is on a given characteristic, which helps clinicians to more adequately consider base rates—the frequency with which certain conditions occur in a setting (see, e.g., Finn & Kamphuis, 1995).

Fifth, research on the reliability and validity of individual test scales sets formal assessment apart from other sources of clinical information. These data allow the astute clinician to understand the strengths or limitations of various scores. Without this, practitioners have little ability to gauge the accuracy of the data they process when making judgments.

The use of test batteries is a final distinguishing feature of formal psychological assessment. In a battery, psychologists generally employ a range of methods to obtain information and cross-check hypotheses. These methods include self-reports, performance tasks, observations, and information derived from behavioral or functional assessment strategies (see Haynes et al., 1997). By incorporating multiple methods, the assessment psychologist is able to efficiently gather a wide range of information to facilitate understanding the patient.

Cross-Method Agreement

Our last point raises a critical issue about the extent to which distinct assessment methods provide unique versus redundant information. To evaluate this issue, Table 3 presents a broad survey of examples. As before, we attempted to draw on meta-analytic reviews or large-scale studies for this table, though this information was not often available. Consequently, many of the entries represent a

new synthesis of relevant literature.[10] To highlight independent methods, we excluded studies that used aggregation strategies to maximize associations (e.g., self-reports correlated with a composite of spouse and peer reports; see Cheek, 1982; Epstein, 1983; Tsujimoto, Hamilton, & Berger, 1990) and ignored moderators of agreement that may have been identified in the literature. We also excluded studies in which cross-method comparisons were not reasonably independent. For instance, we omitted studies in which patients completed a written self-report instrument that was then correlated with the results from a structured interview that asked comparable questions in an oral format (see, e.g., Richter, Werner, Heerlein, Kraus, & Sauer, 1998). However, to provide a wide array of contrasts across different sources, we at times report results that are inflated by criterion contamination.

A review of Table 3 indicates that distinct assessment methods provide unique information. This is evident from the relatively low to moderate associations between independent methods of assessing similar constructs. The findings hold for children and adults and when various types of knowledgeable informants (e.g., self, clinician, parent, peer) are compared with each other or with observed behaviors and task performance. For instance, child and adolescent self-ratings have only moderate correspondence with the ratings of parents (Table 3, Entries 1–4), teachers (Table 3, Entries 8–10), clinicians (Table 3, Entries 5 & 6), or observers (Table 3, Entry 7), and the ratings from each of these sources have only moderate associations with each other (Table 3, Entries 12–18, 20–21). For adults, self-reports of personality and mood have small to moderate associations with the same characteristics measured by those who are close to the target person (Table 3, Entries 23–25, 29–30), peers (Table 3, Entries 26–28), clinicians (Table 3, Entries 31–34), performance tasks (Table 3, Entries 38–44), or observed behavior (Table 3, Entries 45–47).

The substantial independence between methods clearly extends into the clinical arena. Not only do patients, clinicians, parents, and observers have different views about psychotherapy progress or functioning in treatment (see Table 3, Entries 3, 7, & 31) but diagnoses have only moderate associations when they are derived from self-reports or the reports of parents, significant others and clinicians (see Table 3, Entries 4, 6, 15, 17, 30, 33, 34, 48, & 49).[11]

The data in Table 3 have numerous implications, both for the science of psychology and for applied clinical practice. We emphasize just two points. First, at best, any single assessment method provides a partial or incomplete representation of the characteristics it intends to measure. Second, in the world of applied clinical practice, it is not easy to obtain accurate or consensually agreed on information about patients. Both issues are considered in more detail below.

Distinct Methods and the Assessment Battery

A number of authors have described several key features that distinguish assessment methods (see, e.g., Achenbach, 1995; Achenbach, McConaughy, & Howell, 1987; Finn, 1996; McClelland, Koestner, & Weinberger, 1989; Meyer, 1996b, 1997; S. B. Miller, 1987; Moskowitz, 1986; Winter, John, Stewart, Klohnen, & Duncan, 1998). Under optimal conditions, (a) unstructured interviews elicit information relevant to thematic life narratives, though they are constrained by the range of topics considered and ambiguities inherent when interpreting this information; (b) structured interviews and self-report instruments elicit details concerning patients' conscious understanding of themselves and overtly experienced symptomatology, though they are limited by the patients' motivation to communicate frankly and their ability to make accurate judgments; (c) performance-based personality tests (e.g., Rorschach, TAT) elicit data about behavior in unstructured settings or implicit dynamics and underlying templates of perception and motivation, though they are constrained by task engagement and the nature of the stimulus materials; (d) performance-based cognitive tasks elicit findings about problem solving and functional capacities, though they are limited by motivation, task engagement, and setting; and (e) observer rating scales elicit an informant's perception of the patient, though they are constrained by the parameters of a particular type of relationship (e.g., spouse, coworker, therapist) and the setting in which the observations transpire. These distinctions provide each method with particular strengths for measuring certain qualities, as well as inherent restrictions for measuring the full scope of human functioning.

More than 40 years ago, Campbell and Fiske (1959) noted how relative independence among psychological methods can point to unappreciated complexity in the phenomena under investigation. Thus, though low cross-method correspondence can potentially indicate problems with one or both methods under consideration, correlations can document only what is shared between two variables. As such, cross-method correlations cannot reveal what makes a test distinctive or unique, and they also cannot reveal how good a test is in any specific sense. Given the intricacy of human functioning and the method distinctions outlined above, psychologists should anticipate disagreements when similarly named scales are compared across diverse assessment methods. Furthermore, given the validity data provided in Table 2, psychologists should view the results in Table 3 as indicating that each assessment method identifies useful data not available from other sources. As is done in other scientific disciplines (Meyer,

(text continues on page 150)

[10] For Table 3, we searched PsycINFO using a variety of strategies. We also relied on bibliographic citations from contemporary articles and reviews. Although we undoubtedly overlooked pertinent studies, our search was extensive. The 55 entries in Table 3 integrate data from more than 800 samples and 190,000 participants, and we included all studies that fit within our search parameters. Thus, we are confident the findings are robust and generalizable.

[11] Methodologically, agreement between diagnoses derived from self-reports and clinicians is inflated by criterion contamination because clinicians must ground their diagnostic conclusions in the information reported by patients. Similar confounds also likely affect the associations between self-ratings and significant-other ratings.

Table 3
A Sample of Cross-Method Convergent Associations Across Single, Independent Sources of Information

Sources of data and constructs (study and notes)	*r*	κ	*N*
Children and adolescents			
1. Self vs. parent: behavioral and emotional problems (data combined from Achenbach, 1991a, 1997; Achenbach, McConaughy, & Howell, 1987; Cole, Peeke, Martin, Truglio, & Seroczynski, 1998 [average correlation estimated from ranges reported in Cole et al., 1998, p. 452, with *N* determined by the number of participants (288) multiplied by the number of data collection waves (6)]; Cole, Truglio, & Peeke, 1997; Epkins & Meyers, 1994; Forehand, Frame, Wierson, Armistead, & Kempton, 1991; Handwerk, Larzelere, Soper, & Friman, 1999; Henry, Moffitt, Caspi, Langley, & Silva, 1994; Lee, Elliott, & Barbour, 1994; McConaughy, Stanger, & Achenbach, 1992 [concurrent results only]; Meyer, 1996b [average associations between MMPI–A scales and conceptually matched parent ratings derived from the MMPI–A restandardization sample]; Pastorelli, Barbaranelli, Cermak, Rozsa, & Caprara; (1997); Phares & Compas, 1990; Phares, Compas, & Howell, 1989; Reynolds & Kamphaus, 1998 [using only scales with the same name]; Treiber & Mabe, 1987; Verhulst & van der Ende, 1991, 1992).	.29		14,102
2. Self vs. parent: behavioral and emotional problems—*Q* correlations of profile similarity (Achenbach, 1991a; the *Q* correlations were averaged across boys and girls and across 89 common items and eight syndrome scales).	.29		1,829
3. Self vs. parent: symptom change in treatment (Lambert, Salzer, & Bickman, 1998).	.19		199
4. Self vs. parent: DSM Axis I disorder (data combined from Frick, Silverthorn, & Evans, 1994; Puura et al., 1998; Rapee, Barrett, Dadds, & Evans, 1994; Reich, Herjanic, Welner, & Gandhy, 1982; Rubio-Stipec et al., 1994; and Vitiello, Malone, Buschle, Delaney, & Behar, 1990).		.24	1,136
5. Self vs. clinician: behavioral and emotional problems (data combined from Achenbach, McConaughy, & Howell, 1987; and Meyer, 1996b [average associations between MMPI–A scales and conceptually matched clinician ratings derived from the MMPI–A restandardization sample]).	.14		1,079
6. Self vs. clinician: DSM Axis I disorder (data summarize associations between diagnoses from fully structured interviews [i.e., self-report] and clinician-assigned diagnoses; data combined from Aronen, Noam, & Weinstein, 1993; Ezpeleta, de la Osa, Doménech, Navarro, & Losilla, 1997; Piacentini et al., 1993; Rubio-Stipec et al., 1994; Schwab-Stone et al., 1996 [excluding predictor and criterion data generated by the same clinician during the same interview]; Vitiello, Malone, Buschle, Delaney, & Behar, 1990; and Weinstein, Stone, Noam, Grives, & Schwab-Stone, 1989).		.23[a]	998
7. Self vs. clinical observer: change in treatment (Lambert, Salzer, & Bickman, 1998).	.28		199
8. Self vs. teacher: Behavioral and emotional problems (data combined with Achenbach, 1991a; Achenbach, McConaughy, & Howell, 1987; Cole, Truglio, & Peeke, 1997; Crowley, Worchel, & Ash, 1992; Epkins & Meyers, 1994; Forehand, Frame, Wierson, Armistead, & Kempton, 1991; Henry, Moffitt, Caspi, Langley, & Silva, 1994; Lee, Elliott, & Barbour, 1994; Malloy, Yarlas, Montvilo, & Sugarman, 1996; Phares, Compas, & Howell, 1989; Reynolds & Kamphaus, 1998 [using only scales with the same name]; Verhulst & van der Ende, 1991; and Wolfe et al., 1987).	.21		9,814
9. Self vs. teacher: behavioral and emotional problems—*Q* correlations of profile similarity (Achenbach, 1991a; the *Q* correlations were averaged across boys and girls and across 89 common items and eight syndrome scales).	.17		1,222
10. Self vs. teacher: test anxiety (Hembree, 1988; reported effect is the average for the lower and intermediate grade levels given in Table 4 of the article).	.23		3,099
11. Self vs. aggregated peer ratings: behavioral and emotional problems (data combined from Achenbach, McConaughy, & Howell, 1987; Cole, Truglio, & Peeke, 1997; Crowley, Worchel, & Ash, 1992; Epkins & Meyers, 1994; Malloy, Yarlas, Montvilo, & Sugarman, 1996; and Pastorelli, Barbaranelli, Cermak, Rozsa, & Caprara, 1997).[b]	.26		8,821
12. Parent vs. teacher: summed behavioral and emotional problems (data combined from Achenbach, 1991a; Achenbach, McConaughy, & Howell, 1987; Carter, Grigorenko, & Pauls, 1995; M. Cohen, Becker, & Campbell, 1990; Cole, Truglio, & Peeke, 1997; Epkins & Meyers, 1994; Forehand, Frame, Wierson, Armistead, & Kempton, 1991; Garrison & Earls, 1985; Henry, Moffitt, Caspi, Laugley, & Silva, 1994; P. S. Jensen, Traylor, Xanakis, & Davis, 1987; Kline & Lachar, 1992 [results limited to obvious correspondence in their Table 2]; Kumpulainen et al., 1999 [matched factor constructs only]; Lee, Elliott, & Barbour, 1994; McConaughy, Stanger, & Achenbach, 1992 [concurrent results only]; Phares, Compas, & Howell, 1989; Reynolds & Kamphaus, 1998 [using only scales with the same name]; Spiker, Kraemer, Constantine, & Bryant, 1992; Verhulst & Akkerhuis, 1989; and Verhulst & van der Ende, 1991).	.29		29,163
13. Parent vs. teacher: specific behavioral and emotional problems (Verhulst & Akkerhuis, 1989).	.16		1,161

Table 3 (*continued*)

Sources of data and constructs (study and notes)	*r*	κ	*N*
14. Parent vs. teacher: behavioral and emotional problems—*Q* correlations of profile similarity (Achenbach, 1991a; the *Q* correlations were averaged across boys and girls and across 89 common items and eight syndrome scales).	.22		2,274
15. Parent vs. teacher: DSM Axis I disorder (data combined from Frick, Silverthorn, & Evans, 1994; and Offord et al., 1996).		.13	1,229
16. Parent vs. clinician: behavioral and emotional problems (data combined from Achenbach, McConaughy, & Howell, 1987; and Kline & Lachar, 1992 [results limited to obvious correspondence in their Table 2]).	.34		1,725
17. Parent vs. clinician: DSM Axis I disoder (data summarize associations between diagnoses from fully structured interviews [i.e., parent report] or diagnostic questionnaires and clinician-assigned diagnoses; data combined from Ezpeleta, de la Osa, Doménech, Navarro, & Losilla, 1997; Morita, Suzuki, & Kamoshita, 1990; Piacentini et al., 1993; Rubio-Stipec et al., 1994; Schwab-Stone et al., 1996 [excluding predictor and criterion data generated by the same clinician during the same interview]; and Vitiello, Malone, Buschle, Delaney, & Behar, 1990).		.39[a]	786
18. Parent vs. direct observer of child behavior: behavioral and emotional problems (Achenbach, McConaughy & Howell, 1987).	.27		279
19. Parent vs. cognitive test: attentional problems (effect summarizes the association between parent ratings of inattention and the WISC–R/III Freedom From Distractibility Index; data combined from M. Cohen, Becker, & Campbell, 1990; Reinecke, Beebe, & Stein, 1999; and Riccio, Cohen, Hall, & Ross, 1997).	.03		451
20. Teacher vs. clinician: behavioral and emotional problems (Achenbach, McConaughy, & Howell, 1987).	.34		1,325
21. Teacher vs. direct observer of child behavior: behavioral and emotional problems (Achenbach, McConaughy, & Howell, 1987).	.42		732
22. Teacher vs. cognitive test: attentional problems (effect summarizes the association between teacher ratings of inattention and the WISC–R/III Freedom From Distractibility Index; data combined from Anastopoulos, Spisto, & Maher, 1994; M. Cohen, Becker, & Campbell, 1990; Lowman, Schwanz, & Kamphaus, 1996; Reinecke, Beebe, & Stein, 1999; and Riccio, Cohen, Hall, & Ross, 1997).	.10		483
Adults			
23. Self vs. spouse/partner: personality and mood (data combined from A. L. Edwards & Klockars, 1981; and Meyer, 1996b [average association between MMPI–2 scales and conceptually matched spouse ratings derived from the MMPI–2 restandardization sample]).	.29		2,011
24. Self vs. spouse/partner: Big Five personality traits—domains and facets (data combined from Bagby et al., 1998 [included friend and spouse ratings]; Borkenau & Liebler, 1993; Conley, 1985 [concurrent ratings only]; Costa & McCrae, 1988 [only concurrent correlations were used], 1992; Foltz, Morse, Calvo, & Barber, 1997; McCrae, 1982; McCrae, Stone, Fagan, & Costa, 1998; Mutén, 1991; and Yang et al., 1999).	.44		1,774
25. Self vs. parent: personality characteristics (including the Big Five; data combined from Caldwell-Andrews, Baer, & Berry, 2000; Funder, Kolar, & Blackman, 1995; Harkness, Tellegen, & Waller, 1995; and Harlan & Clark, 1999; if results for both mothers and fathers were reported for the same participants, they were treated as independent findings. The median correlation for self–father ratings was used from Harlan & Clark because this was all that was reported).	.33		828
26. Self vs. peer: personality and mood (data combined from Funder & Colvin, 1988; Funder, Kolar, & Blackman, 1995; Harkness, Tellegen, & Waller, 1995; A. F. Hayes & Dunning, 1997; Hill, Zrull, & McIntire, 1998; Kurokawa & Weed, 1998; Oltmanns, Turkheimer, & Strauss, 1998; Paunonen, 1989 [estimates derived from unpartialed correlations reported in Paunonen's Figures 2 and 3 using only degree of acquaintanceship rated 6–9]; Watson & Clark, 1991; and Zuckerman et al., 1988. Funder and Colvin reported correlations between self-ratings and the composite of two informants. Because the average interinformant correlation was also reported, an estimate of the correlation between self-ratings and the ratings of a single informant was generated using the formula provided by Tsujimoto, Hamilton, & Berger, 1990. The same formula was used with data in Oltmanns et al. to estimate the correlation between self-ratings and the ratings of a single peer).	.27		2,119

(*table continues*)

Table 3 (*continued*)

Sources of data and constructs (study and notes)	*r*	κ	*N*
27. Self vs. peer: Big Five personality traits—domains and facets (data combined from Cheek, 1982; Costa & McCrae, 1992; Funder, Kolar, & Blackman, 1995 [the two sets of self–peer associations in their Table 1 were treated as independent samples]; John & Robins, 1993; Koestner, Bernieri, & Zuckerman, 1994; McCraee & Costa, 1987; Paulhus & Reynolds, 1995; Piedmont, 1994; Zuckerman, Bernieri, Koestner, & Rosenthal, 1989; and Zuckerman, Miyake, Koestner, Baldwin, & Osborne, 1991. For Paulhus & Reynolds, the Wave 2 validity coefficients from their Table 4 were adjusted to reflect the validity of a single rater. This was done by assuming the initial findings were generated from four-rater composites and using the formula presented in Tsujimoto, Hamilton, & Berger, 1990. The same formula was used to estimate validity for a single rater from Piedmont's data, though it could not be used with Koestner et al.).	.31		1,967
28. Self vs. peer: job performance (Conway & Huffcutt, 1997).	.19		6,359
29. Self vs. significant other: attentional problems and impulsivity (Ryan, 1998).	.22		202
30. Self vs. significant other: DSM Axis II personality disorder diagnosis (data combined from Bernstein et al., 1997; Dowson, 1992 [kappa estimated to be 0.0 when values were not reported but said to be nonsignificant]; Dreessen, Hildebrand, & Arntz, 1998; Ferro & Klein, 1997; Riso, Klein, Anderson, Ouimette, & Lizardi, 1994; and Zimmerman, Pfohl, Coryell, Stangl, & Corenthal, 1988).		.12	768
31. Self vs. clinician: treatment-related functioning, symptomatology, and outcome (data combined from Cribbs & Niva, 2000, and Nebeker, Lambert, & Huefner, 1995).	.29		7,903
32. Self vs. clinician: DSM Axis II personality disorder characteristics (findings examine the correspondence between self-report scales of personality disorders and clinician ratings on the same dimensions; data were combined from Barber & Morse, 1994 [using only the dimensional scores reported in their Table 5]; Burgess, 1991; de Ruiter & Greeven, 2000; Ekselius, Lindström, von Knorring, Bodlund, & Kullgren, 1994 [coefficients were Spearman correlations]; Fossati et al., 1988; Hart, Forth, & Hare, 1991; Hunt & Andrews, 1992 [intraclass correlations were used in this study]; Kennedy et al., 1995; Marlowe, Husband, Bonieskie, Kirby, & Platt, 1997; Millon, 1994; Overholser, 1994 [Studies 1, 5, 8, 9, 12, and 13 from Overholser's Table III were used]; Rogers, Salekin, & Sewell, 1999 [Studies 12, 19, 20, and 22 from their Table 3 were used]; Soldz, Budman, Demby, & Merry, 1993; and Trull & Larson, 1994).	.33[a]		2,778
33. Self vs. clinician: DSM Axis II personality disorder diagnosis (findings examine the correspondence between diagnostic cutoff criteria from self-report scales and clinician-assigned diagnoses; data were combined from de Ruiter & Greeven, 2000; Ekselius, Lindström, von Knorring, Bodlund, & Kullgren, 1994 [kappa was calculated from their Tables 1 and 2]; Fossati et al., 1998; Jacobsberg, Perry, & Frances, 1995 [kappa was calculated from their Table 1]; Kennedy et al., 1995; Marlowe, Husband, Bonieskie, Kirby, & Platt, 1997 [kappa was calculated from their Table 3 using BR > 84 data; BR > 74 data led to a smaller average kappa]; Nussbaum & Rogers, 1992; Perry, 1992; Renneberg, Chambless, Dowdall, Fauerbach, & Gracely, 1992 [kappa coefficients were available for all disorders using BR > 74 as the cutoff, so they were used here]; Rogers, Salekin, & Sewell, 1999 [Studies 2 and 11 were used]; Soldz, Budman, Demby, & Merry, 1993; and Trull & Larson, 1994).		.18[a]	2,859
34. Self vs. clinician: DSM Axis I disorders (Meyer, in press; coefficient summarizes the association between diagnoses from a fully structured interview [i.e., self-report] and clinician-assigned diagnoses, excluding designs in which both diagnoses were derived from the same interview).		.34[a]	5,990
35. Self vs. clinician: Big Five personality traits (domains only; Piedmont & Ciarrocchi, 1999).	.32		132
36. Self vs. supervisor: job performance (Conway & Huffcutt, 1997).	.22		10,359
37. Self vs. subordinate: job performance (Conway & Huffcutt, 1997).	.14		5,925
38. Self vs. cognitive test or grades: general intelligence (data combined from Borkenau & Liebler, 1993; Mabe & West, 1982 [using the *ns* reported in their Table 1]; and Paulhus, Lysy, & Yik, 1998).	.24		904
39. Self vs. cognitive test or grades: scholastic ability (Mabe & West, 1982; the reported *N* was derived from their Table 1 using studies that reported on the strength of association).[c]	.38		8,745

Table 3 (*continued*)

Sources of data and constructs (study and notes)	*r*	κ	*N*
40. Self vs. cognitive test: memory problems (data combined from Branca, Giordani, Lutz, & Saper, 1995; Brown, Dodrill, Clark, & Zych, 1991; Gagnon et al., 1994; Gass, Russell, & Hamilton, 1990 [using only the memory-specific self-report scale]; Herzog & Rodgers, 1989; Johansson, Allen-Burge, & Zarit, 1997; Olsson & Juslin, 1999; Seidenberg, Haltiner, Taylor, Hermann, & Wyler, 1994; G. E. Smith, Petersen, Ivnik, Malec, & Tangalos, 1996; J. L. Taylor, Miller, & Tinklenberg, 1992; and Zelinski, Gilewski, & Anthony-Bergstone, 1990).	.13		5,717
41. Self vs. cognitive test: attentional problems (data combined from Meyer, 1996b; Paulhus, Aks, & Coren, 1990; Ryan, 1998; Seidenberg, Haltiner, Taylor, Hermann, & Wyler, 1994; and Turner & Gilliland, 1997 [unreported but nonsignificant correlations were considered to be zero]).	.06		522
42. Self vs. Thematic Apperception Test: achievement motivation (Spangler, 1992).[c]	.09		2,785
43. Self vs. Thematic Apperception Test: problem solving (Ronan, Colavito, & Hammontree, 1993).	.13		199
44. Self vs. Rorschach: emotional distress, psychosis, and interpersonal wariness (data combined from Meyer, 1997; and Meyer, Riethmiller, Brooks, Benoit, & Handler, 2000).	.04		689
45. Self vs. observed behavior: personality characteristics (data combined from Gosling, John, Craik, & Robins, 1998; Kolar, Funder, & Colvin, 1996; and Moskowitz, 1990. Kolar et al. used the aggregated ratings of six observers on average, whereas Moskowitz relied on the aggregated ratings of four observers; thus, the overall coefficient reported here is larger than it would be if each study had relied on behavior ratings from a single observer).	.16		274
46. Self vs. observed behavior: attitudes (Kraus, 1995; the reported *N* was derived from the total number of studies times the average *n* per study. Kim & Hunter, 1993, also conducted a meta-analysis of attitude–behavior relations. However, in their criterion measures, they did not distinguish between self-reported behavior and observed behavior).	.32		15,624
47. Peers vs. observed behavior: personality characteristics (Kolar, Funder, & Colvin, 1996. Coefficient reflects the average of two sets of single-peer ratings correlated with observed behavior. Ratings of observed behavior were aggregated from six observers on average, so the reported correlation is larger than would be found if behavior was rated by a single observer).	.15		264
48. Clinician vs. consensus best estimate: DSM Axis II personality disorder diagnosis (data combined from Perry, 1992 [using only the Skodol et al. data]; Pilkonis et al., 1995 [all diagnostic data in their Table 1 were averaged]); and Pilkonis, Heape, Ruddy, & Serrao, 1991 [excluding PAF data but including baseline and follow-up kappa for "any personality disorder"].		.28	218
49. Significant other vs. significant other: target patient's DSM personality disorder diagnosis (Ferro & Klein, 1997).		.32	386
50. Significant other vs. clinician: target patient's depressive signs and symptoms (G. Parker et al., 1992; average agreement computed from their Tables 1 and 2).		.13	141
51. Judgments from one source of test data vs. another: personality, needs, and IQ (data combined from L. R. Goldberg & Werts, 1966; Howard, 1962 [total *N* was determined by multiplying the 10 patients by the seven raters]; and Little & Shneidman, 1959. For Little and Shneidman, congruence across judgments from the Rorschach, Thematic Apperception Test, MMPI, and Make a Picture Story Test was estimated by subtracting the average coefficient in their Table 10 from the average test coefficient reported in their Table 9).[d]	.12		158
52. Supervisor vs. peers: Job performance (Conway & Huffcutt, 1997).	.34		7,101
53. Supervisor vs. subordinate: Job performance (Conway & Huffcutt, 1997).	.22		4,815
54. Peers vs. subordinate: Job performance (Conway & Huffcutt, 1997).	.22		3,938
55. Objective criteria vs. managerial ratings: Job success (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995).	.32		8,341

Note. *r* = Pearson correlation; κ = kappa coefficient; BR = base rate; *N* = number of participants; DSM = Diagnostic and Statistical Manual of Mental Disorders; IQ = intelligence quotient; MMPI = Minnesota Multiphasic Personality Inventory; MMPI–A = adolescent version of MMPI; PAF = Personality Assessment Form; WISC–R/III = Wechsler Intelligence Scale for Children—Revised & Third Edition.

[a] These coefficients are inflated by criterion contamination. For instance, in an effort to maximize cross-observer correspondence, one study (Ekselius, Lindstrom, von Knorring, Bodlund, & Kullgren, 1994) went so far as to exclude the inferences that clinicians developed from their direct observations of the patient as a way to increase diagnostic agreement between patients and clinicians. [b] Because much of this data reflects the correlation between aggregated peer ratings and self-ratings, the coefficient is larger than would be obtained between self-ratings and the ratings of a single peer. [c] Result combines some data from children and adolescents with adults. [d] These studies were from the late 1950s and early 1960s. It is unclear whether the data may be different using more contemporary scoring and interpretive practices.

1996b), clinicians and researchers should recognize the unique strengths and limitations of various assessment methods and harness these qualities to select methods that help them more fully understand the complexity of the individual being evaluated.[12]

Test batteries, particularly in the area of personality assessment, have been criticized at times because evidence for the incremental validity of each test within the battery has not been consistently demonstrated (see, e.g., Garb, 1984). However, several logical and empirical considerations support the multimethod battery as a means to maximize assessment validity.

In particular, we believe that there is a direct parallel between empirical research and applied clinical practice on this issue. In research, *monomethod* bias and *monooperation* bias are critical threats to the validity of any investigation (Cook & Campbell, 1979). Thus, research validity is compromised when information is derived from a single method of measurement (e.g., self-report) and when a construct has been operationally defined in a single way (e.g., depression delineated by emotional rather than physiological, interpersonal, or cognitive symptoms).

The optimal methodology to enhance the construct validity of nomothetic research consists of combining data from multiple methods and multiple operational definitions (see, e.g., Cole, Martin, Powers, & Truglio, 1996; Cook & Campbell, 1979; Epstein, 1980, 1983). To our knowledge, the same standards have not been directly linked to principles for guiding the idiographic clinical assessments that are designed to understand the full complexity of a single individual. We believe the parallels should be explicit.

Just as optimal research recognizes that any method of measurement and any single operational definition of a construct are incomplete, optimal clinical assessment should recognize that the same constraints exist when measuring phenomena in the life of a single person. Furthermore, just as effective nomothetic research recognizes how validity is maximized when variables are measured by multiple methods, particularly when the methods produce meaningful discrepancies (Cheek, 1982; Cole et al., 1996; Tsujimoto et al., 1990), the quality of idiographic assessment can be enhanced by clinicians who integrate the data from multiple methods of assessment (Achenbach, 1995; Colvin, Block, & Funder, 1995; Ganellen, 1994; McClelland et al., 1989; Meyer, 1996b, 1997; S. B. Miller, 1987; Shedler, Mayman, & Manis, 1993; Winter et al., 1998).

It is well known that lapses in reasoning often may accompany clinical judgment (see, e.g., Arkes, 1981; Borum, Otto, & Golding, 1993; Garb, 1994; Hammond, 1996; Holt, 1986). Although these pitfalls also can affect assessments, the evaluation process incorporates some inherent checks on clinical reasoning. An assessment battery is likely to generate findings that, at least superficially, appear conflicting or contradictory. When assessors systematically integrate this information, they are forced to consider questions, symptoms, dynamics, and behaviors from multiple perspectives—simply because everything does not fit together in a neat and uncomplicated package. Clinicians must consider the nature of the information provided by each testing method, the peculiarities associated with the specific way different scales define a construct, the reliability and validity of different scales, and the motivational and environmental circumstances that were present during the testing. Assuming no data can be deemed invalid and ignored, then the assessment clinician must conceptualize the patient in a way that synthesizes all of the test scores. Next, these test-based conceptualizations must be reconciled with what is known from history, referral information, and observation. Finally, all of this information must be integrated with the clinician's understanding of the complex condition(s) being assessed (e.g., narcissistic personality disorder, learning disability, transference reactions, contingencies that maintain obsessive behaviors) and the many other complex conditions that need to be considered and then ruled out as unimportant or irrelevant. Although there are many places in this process for errors to develop, the careful consideration of multimethod assessment data can provide a powerful antidote to the normal judgment biases that are inherent in clinical work (also see Borum et al., 1993; Spengler, Strohmer, Dixon, & Shivy, 1995). This line of reasoning also suggests that by relying on a multimethod assessment battery, practitioners have historically used the most efficient means at their disposal to maximize the validity of their judgments about individual clients.

Method Disparities and Errors in Practice

Current knowledge about the substantial disagreements between methods of information gathering has important implications for health care. The data indicate that even though it may be less expensive at the outset, a single clinician using a single method (e.g., interview) to obtain information from a patient will develop an incomplete or biased understanding of that patient. To the extent that such impressions guide diagnostic and treatment decisions, patients will be misunderstood, mischaracterized, misdiagnosed, and less than optimally treated. Over the long term, this should increase health care costs.

These issues are not trivial. The evidence indicates that clinicians who use a single method to obtain patient information regularly draw faulty conclusions. For instance, Fennig, Craig, Tanenberg-Karant, and Bromet (1994) reviewed the diagnoses assigned to 223 patients as part of usual hospital practice. Clinical diagnoses were then compared with diagnoses derived from a comprehensive multimethod assessment that consisted of a semistructured patient interview, a review of the patient's medical record, a semistructured interview with the treating clinician, and an interview with the patient's significant other, all of which were then reviewed and synthesized by two clini-

[12] Unlike other scientific disciplines, a factor that contributes to divergence across psychological methods undoubtedly emerges from a discipline-wide propensity to ignore the fundamental measurement question, which is whether the objects or attributes psychologists aspire to measure actually have quantitative properties (Michell, 1997). In part, this question is ignored because test results can have practical utility even without this knowledge. Utility does not demand cross-method convergence. However, precise convergence would be required for any two methods that purported to measure the same quantitative attribute.

cians to derive final diagnoses from the multimethod assessment.

Even though Fennig, Craig, Tanenberg-Karant, et al. (1994) used very liberal criteria to define diagnostic agreement (e.g., major depression with psychotic features was treated as equivalent to dysthymia), the diagnoses assigned during the course of typical clinical practice had poor agreement with the diagnostic formulations derived from the more extensive synthesis of multiple assessment methods. Overall, after discounting chance agreement, the clinical diagnoses agreed with the multimethod conclusions only about 45–50% of the time.[13] This was true for a range of disorders on the schizophrenic, bipolar, and depressive spectrums. Because these conditions are treated in decidedly different ways, such frequent misdiagnoses in typical practice suggest that many patients erroneously receive antipsychotic, antimanic, and antidepressant medications.

Another example involves fully structured interviews like the Composite International Diagnostic Interview (CIDI), which have a format that makes them essentially equivalent to an oral self-report instrument. A salient question concerns the extent to which diagnoses from CIDI-type scales agree with those derived from clinicians who also rely on their impression of the patient (e.g., from semistructured interviews, from clinical consensus after following the patient over time). Although diagnoses from the CIDI and diagnoses derived from semistructured interviews suffer from criterion contamination because both the predictor and criterion rely on the patient's report as a primary source of information (see, e.g., Malgady, Rogler, & Tryon, 1992), Table 3 indicates that across 33 samples and 5,990 patients, the correspondence between CIDI-type diagnoses and clinician diagnoses was quite modest (κ = .34; Table 3, Entry 34; see Meyer, in press). Similar findings have been observed when Axis I diagnoses from the Structured Clinical Interview for the *Diagnostic and Statistical Manual of Mental Disorders* were compared with clinician diagnoses (mean κ = .26, N = 100; Steiner, Tebes, Sledge, & Walker, 1995), suggesting again that the source of information for diagnostic inferences exerts a prominent influence over final classifications (see, e.g., Offord et al., 1996).

Although the above disagreements are pronounced, even more drastic errors have been found for personality disorders. Perry (1992) and Pilkonis et al. (1995) compared diagnoses derived from a semistructured clinical interview with diagnoses based on more extensive and complex assessments using multiple methods of gathering patient information. Across studies, there was a meager correspondence between the diagnoses derived from a single clinician using the single method of assessment and the diagnoses derived from the multimethod evaluations (κ = .28; N = 218; see Entry 48 in Table 3). In fact, after correcting for agreements due to chance, about 70% of the interview-based diagnoses were in error.

The evidence also indicates that personality disorder diagnoses diverge substantially across other sources of information. For instance, Table 3 shows that diagnoses derived from self-report bear little resemblance to those derived from clinicians (κ = .18, N = 2,859; Table 3, Entry 33) and that diagnoses from semistructured patient interviews bear little resemblance to those based on semistructured interviews with significant others in the patient's life (κ = .12, N = 768; Table 3, Entry 30).

Though the latter results are sobering, they are open to interpretation about which perspective is more correct. The most relevant evidence is that which compared interviews with the multimethod synthesis of information. These data clearly demonstrate how conclusions derived from a typical evaluation using a single method of assessment had little correspondence with those derived from a more comprehensive evaluation. By necessity then, the research findings indicate that many patients may be misunderstood or improperly treated when they do not receive thorough assessments. Errors of misappraisal and mistreatment are most likely when administrative efforts to save money restrict clinicians to very brief and circumscribed evaluations.

Issues at the Interface of Assessment Research and Practice

Virtually all research with purported relevance to assessment has examined the nomothetic association between isolated test scores and equally isolated criterion measures (e.g., MMPI Depression scores in patients with depression vs. patients without that diagnosis). In such an approach, the scores from one scale are evaluated out of context from other test scores and sources of information. This strategy is ideal for scale validation because it allows for an understanding of the strengths and limitations of a single scale, divorced from the array of other factors that impinge on any assessment (Cronbach & Meehl, 1955). However, this research strategy does very little for the assessment clinician, who is almost never concerned with a single scale but rather with one scale in the context of other scales and other sources of information.

Because the nomothetic association between different methods is generally small to moderate, if the results from most testing research are considered in isolation, the observed validity coefficients suggest that psychologists have a limited capacity to make sound, individualized judgments from test scales alone. This is true even for the substantial coefficients presented in Table 2. In fact, if the value of clinical assessment could be supported only by the testing evidence that documents the validity of test scales divorced from contextual factors (i.e., Tables 2 and 3), then, as a profession, psychologists might be forced to abandon assessment as a justifiable activity. When one considers the errors associated with measurement and the infrequent occurrence of most clinical conditions, validity coefficients are too small to justify testing-based decisions for individuals (Hummel, 1999). Thus, someone with a high score on the Depression scale of the MMPI cannot be assigned a

[13] In a separate study with the same population, Fennig, Craig, Lavelle, Kovasznay, and Bromet (1994) demonstrated how clinicians who derived psychiatric diagnoses after synthesizing information from multiple sources had much higher correspondence with the gold standard criterion diagnoses.

depressive diagnosis with conviction, just as someone with a low score on the Wechsler Memory Scale (WMS) cannot be assigned a diagnosis of Alzheimer's disease with confidence. This is true even when scores deviate substantially from normal.[14]

The fact that one cannot derive unequivocal clinical conclusions from test scores considered in isolation should not be a surprise, as sophisticated clinicians would never expect to make a diagnosis from just a single test or scale. However, failure to appreciate the testing-versus-assessment distinction has led some to seriously question the utility of psychological tests in clinical contexts (see, e.g., Hummel, 1999; Rogers, Salekin, & Sewell, 1999). When this important difference is not recognized or fully appreciated, the testing literature may lead to pessimism about psychological assessment, even though they are quite different activities.

Because most research studies do not use the same type of data that clinicians do when performing an individualized assessment, the validity coefficients from testing research may underestimate the validity of test findings when they are integrated into a systematic and individualized psychological assessment. To illustrate, when conducting an idiographic assessment using an MMPI, the clinician begins by examining the validity scales to understand the patient's test-taking approach. This analysis is completed first because all other scale elevations need to be interpreted in this light. The same elevation on the MMPI Depression scale means something very different when the validity scales indicate the patient was open and straightforward during the evaluation, rather than guarded and defensive. Other contextual factors must also be considered. A *T* score of 100 on the F Scale (Infrequency) may have very different implications if the patient is tested on an acute inpatient ward rather than in an outpatient clinic. In the latter setting, this elevation is more likely to indicate that the MMPI–2 data are invalid because the patient responded to items in an inconsistent manner or magnified the extent of his or her disturbance. However, in an inpatient setting, the very same score is more likely to be an accurate reflection of the patient's acute distress and genuine disturbance. Competently trained clinicians recognize these contextual factors and interpret scale scores accordingly.

The same type of reasoning is used when evaluating data from other assessment methods. For example, neuropsychological test scores are considered in light of the patient's level of fatigue, attention, cooperation, estimated premorbid level of functioning, and so forth because all of these factors can influence performance and the proper interpretation of obtained scores.

The important point here is that contextual factors play a very large role in determining the final scores obtained on psychological tests. In methodological terms, when test scores are studied across large groups of people, the contextual factors associated with each individual contribute to what is known as *method variance* (see, e.g., Campbell & Fiske, 1959; L. K. Edwards & Edwards, 1991; Glutting, Oakland, & Konold, 1994; Jackson, Fraboni, & Helmes, 1997; Meyer, 1997; Oakland & Glutting, 1990). Tests employed in other scientific disciplines are less affected by these factors, as results from an x-ray, blood chemistry panel, seismograph, or carbon-14 dating test never depend on the motivation, rapport, or drowsiness of the object under study. However, these are all critical factor that influence the scores obtained on any psychological test.

Although skilled clinicians appear to recognize the contextual factors described above, it is much more difficult to make such individualized adjustments when conducting research. This is because scale scores are not given differential trustworthiness weights to reflect the fact that some are obtained from patients who are exaggerating, some from patients who are unmotivated, some from patients who are open and frank, some from patients who are highly guarded and defended, and so on. Rather, every test score is identically weighted and regarded as if it were equally valid. (Of course, every criterion score is treated in the same fashion.)

The salience of these individualized contextual factors may be easier to recognize with two specific examples. First, consider a clinician who is asked to determine if a man is depressed given (a) an MMPI–2 Depression score that is unusually low, (b) a mild elevation on MMPI–2 Scale 3 (Hysteria), (c) an elevated Rorschach Depression Index, (d) clinical observations on the Brief Psychiatric Rating Scale (BPRS) that yield somewhat elevated scores for emotional withdrawal and guilt feelings but a suppressed score for depressive mood, (e) the patient's report that he recently lost a loved one and now has sleeping difficulties, and (f) a report from the patient's sister that, since childhood, he has successfully coped with problems by "looking on the bright side of things." With these data, the clinician could conclude the man is struggling with an underlying depressive condition (as evident on portions of the BPRS, Rorschach, and history) brought about by his recent loss (from the history), even though his generally upbeat coping strategy (from his sister's description and MMPI–2 Scale 3) prevents him from acknowledging his troubles (as evident from the MMPI–2 Depression scale and part of the BPRS). One might also infer that his defenses serve an important function and that treatment that abruptly confronted his underlying emotions could leave him in a psychologically unbalanced state.

Note how in this individualized context, the MMPI–2 Depression score supports the valid conclusion that the patient is struggling with depression despite the fact that it

[14] Psychologists can of course still use testing data (i.e., scores derived from a single scale or a single prediction equation) if the data are applied in a selection context, such as with employment screening tests, the Graduate Record Examination, the Scholastic Aptitude Test, and so on. This is because one can choose a small number of applicants from a large pool as a way to maximize validity (H. C. Taylor & Russell, 1939). However, this strategy reflects an application of nomothetically derived validity coefficients in an appropriate nomothetic context. Such procedures are not helpful when applying nomothetic validity coefficients to the idiographic practice of psychological assessment.

indicates less depression than would be found in an average person without psychiatric difficulties. The MMPI–2 score is low for this man because it accurately reflects his efforts to cope by keeping depressive experiences at bay (cf. Finn, 1996; Meyer, 1997). Unfortunately, the clinical accuracy of a score like this is lost in a typical statistical analysis because correlations, *t* tests, *F* tests, and so on do not take into account the complex array of unique contextual variables associated with individual patients. In fact, in a typical study, the clinical accuracy of this man's MMPI–2 score would be treated as error, and including his score in research would serve only to reduce the size of a correlation, *t* value, or *F* value that quantified the validity of the MMPI–2. Thus, even though this man's MMPI–2 would provide valid information for an idiographic assessment, it would actually make the MMPI–2 scale appear less valid in nomothetic research.

As another example, early stage dementia is more likely when an elderly person's memory is poor yet other cognitive abilities are intact. Thus, the diagnosis is more probable if assessment data reveal low memory test performance (e.g., on the WMS) in combination with high scores on a test like the National Adult Reading Test (NART), which estimates premorbid intelligence on the basis of the pronunciation of irregularly spelled words. This idiographic contrast quantifies a key feature of the disorder. Dementia is also more likely if the patient minimizes memory problems even though his or her spouse reports instances of poor memory, if the family history is positive for Alzheimer's disease, if there is no evidence of localized dysfunction on other neuropsychological tests, and if recent MRI or CT scans do not show localized signs of stroke.

In a large meta-analysis, D. Christensen, Hadzi-Pavlovic, and Jacomb (1991) found scores from the WMS and similar tests had a strong ability to differentiate patients with dementia from normal controls (see Entry 137 in Table 2). However, NART scores had a minimal ability to make this kind of discrimination ($r = .14$). Thus, the testing results indicated NART scores were not very useful for diagnosis. In clinical practice, however, an assessment clinician would be most inclined to diagnose dementia when test scores indicated high premorbid cognitive functioning (i.e., high NART scores) in the presence of currently compromised memory (e.g., low WMS scores). Thus, because the NART is not only a valid measure of preexisting cognitive abilities (Spreen & Strauss, 1998) but also relatively insensitive to dementia symptoms, it can be a critical asset for diagnosing dementia on an individual-by-individual basis. If one had relied on just the nomothetic effect size, one would have concluded that the NART was of little value to the diagnosis of dementia, even though its applied clinical value is actually much higher because it allows the clinician to estimate an individual's memory decline relative to his or her premorbid cognitive abilities.

More generally, to the extent that clinicians view all test data in a contextually differentiated fashion, the practical value of tests used in clinical assessment is likely greater than what is suggested by the research on their nomothetic associations.[15] However, trying to document the validity of individualized, contextually embedded inferences is incredibly complex—and virtually impossible if one hopes to find a relatively large sample of people with the same pattern of test and extratest information (i.e., history, observed behavior, motivational context, etc.). Research cannot realistically hope to approximate such an ideal. Nevertheless, using just test scores, a growing body of findings support the value of combining data from more than one type of assessment method, even when these methods disagree within or across individuals (see, e.g., Colvin et al., 1995; Davidson, 1996; Ganellen, 1994; Klein et al., 1994; McClelland et al., 1989; Meyer, 1997; Meyer, Riethmiller, Brooks, Benoit, & Handler, 2000; Power et al., 1998; Robertson & Kinder, 1993; Shedler et al., 1993; Winter et al., 1998).

Future Research

Assessment is a complicated activity that requires (a) sophisticated understanding of personality, psychopathology, or the many ways in which neurological disorders are manifested in cognition and behavior; (b) knowledge of psychological measurement, statistics, and research methods; (c) recognition that different assessment methods produce qualitatively distinct kinds of information; (d) understanding of the particular strengths and limitations of each method and of different scales within each method; (e) a capacity to conceptualize the diverse real-world conditions that could give rise to a particular pattern of test data; (f) the ability to challenge one's judgment by systematically linking the presence and absence of test indicators to the psychological characteristics under consideration; and (g) the interpersonal skill and emotional sensitivity to effectively communicate findings to patients, significant others, and referral sources.

Although psychological tests can assist clinicians with case formulation and treatment recommendations, they are only tools. Tests do not think for themselves, nor do they directly communicate with patients. Like a stethoscope, a blood pressure gauge, or an MRI scan, a psychological test is a dumb tool, and the worth of the tool cannot be separated from the sophistication of the clinician who draws inferences from it and then communicates with patients and other professionals. Because assessment competence re-

[15] Our argument is not that clinical judgment will consistently surpass statistical decision rules in a head-to-head comparison (Meyer et al., 1998; see Grove, Zald, Lebow, Snitz, & Nelson, 2000, for a meta-analytic review). Rather, it is that the practical validity of psychological assessment (i.e., the sophisticated integration of data from multiple tests and sources of contextual information) is probably greater than what is suggested by the validity coefficients found in the testing literature (i.e., scale data in which the many contextual factors affecting all observed scores are treated as error variance). Also, if this line of reasoning is extended, one should expect nomothetic validity coefficients for testing data to increase when researchers begin to differentially weigh scores to reflect individualized contextual influences. As a simple example that builds on the text discussion, if researchers attend to premorbid intelligence as an important contextual variable, dementia studies should produce larger effect sizes when the NART–WMS discrepancy is the dependent variable than when WMS and NART scores are considered in isolation.

quires a considerable investment of time and effort, further documenting the worth of this investment is our final consideration.

More than 20 years ago, psychologists with an interest in treatment took the lead in demonstrating how clinicians have practical utility for enhancing patient outcome (M. L. Smith & Glass, 1977). Today, the beneficial impact of treatment continues to be documented (see, e.g., Lipsey & Wilson, 1993; Seligman, 1995; Shadish et al., 1997). Assessment research—in both psychology and medicine—has generally followed a path that differs from treatment research. Although notable exceptions exist (see Entries 77 & 91 in Table 2), researchers have historically focused at a micro level to evaluate the psychometric reliability and validity of test scales that are divorced from an individualized context. This focus is certainly important. However, researchers should also focus at a macro level to evaluate the practical value of clinicians who use tests as tools that help them provide professional consultation and benefit to patients and allied health care providers.

We are not the first to recognize this imbalance in the literature. It has been noted regularly over the years (see, e.g., Finn & Tonsager, 1997; S. C. Hayes, Nelson, & Jarrett, 1987; Korchin & Schuldberg, 1981; McReynolds, 1985; Meehl, 1959; Moreland, Fowler, & Honaker, 1994; Persons, 1991). Unfortunately, recognizing the imbalance has not yet been sufficient to correct it.

Research designs for evaluating assessment utility have been proposed by S. C. Hayes et al. (1987) and recently discussed again by Finn and Tonsager (1997). Even a relatively simple design addressing the utility of psychological assessment for affecting referral sources, patient care, and patient well-being would be of considerable value. For example, a group of patients deemed to be in need of psychological assessment could be provided with (a) a flexible, multimethod assessment battery using tests typically employed in practice and selected on the basis of idiographic referral questions by a clinician competent in the relevant domain, (b) personal feedback from the assessment, and (c) feedback to their treating and referring clinicians. These patients could then be contrasted with an appropriate control group, such as patients who also were deemed to be in need of a psychological assessment but received a comparable amount of therapy rather than any of the above.[16] Given that the main purpose of assessment is to provide useful information to patients and referral sources, key outcomes would directly address these issues (e.g., resolution of patient and therapist referral questions, congruence over treatment goals, confidence that treatment is moving in a helpful direction).[17] Conducting this type of research would complement the very strong findings in Table 2 by documenting the extent to which the test-informed assessment clinician is useful and effective in everyday clinical practice.

A second important issue concerns the accuracy of judgments made by assessment clinicians. This could be addressed by building on the basic design mentioned above to have clinicians describe the patients in the experimental and control groups using standard measures of symptomatology and functioning. The accuracy of the ratings given to patients who received a flexible, multimethod assessment battery would then be compared with those generated for patients who did not receive an assessment but were deemed to be in need of one. This comparison would quantify the value of assessment for the accurate understanding of patients.

The key to the latter type of study—and what would set it apart from prior research in this area—is ensuring that the criterion judgments that determine accuracy are as systematic, comprehensive, and true as possible. Particularly for personality assessment, there is no ready gold standard that allows psychologists to know a patient with certainty. Table 3 reveals unequivocally that psychologists cannot use self-, clinician, teacher, spouse, or peer ratings as a criterion because judgments from these different perspectives agree only modestly. Thus, every single source of information diverges substantially from every other potential source, and it is impossible to say that one (e.g., clinician) is more true than any other (e.g., spouse). Yet if one wants to evaluate the accuracy of judgments derived from a psychological assessment, one must have excellent criteria available first. Thus, following Meehl (1959), criterion ratings should be obtained by the consensus of experts after patients have been followed over time, after interviews have been conducted with significant others, after interviews have been conducted with mental health and medical personnel who have encountered the patients, and after systematic consideration has been given to all the available data for each person (see Klein et al., 1994, and Pilkonis et al., 1995, for examples applied to diagnostic criteria; see Faraone & Tsuang, 1994, Meyer, 1996a, and Tsujimoto et al., 1990, for alternative ways to maximize criterion validity). Ensuring that the criterion measures are sufficient gold standards will require a considerable investment of time and resources. However, if psychologists wish to clearly document whether judgments and inferences are more accurate when they are derived from a multimethod psychological assessment, it is necessary to spend the time

[16] The experimental and control groups should consist of patients deemed to be in need of an assessment according to some reasonable clinical criteria. Just as every patient does not need a CT scan, every patient does not need a psychological assessment. Randomly assigning all patients to experimental and control conditions would serve only to drastically reduce the statistical power of the design and the size of any observed effect. Also, in the current health care climate, it should be possible to find providers who refuse to authorize psychological assessments regardless of need (Eisman et al., 1998, 2000). Thus, the design could provide a new assessment service, rather than withhold appropriate care from patients otherwise eligible for it.

[17] Previously, we said it may be valuable to measure the impact of assessment on outcomes like length, cost, or speed of improvement in treatment (Meyer et al., 1998). However, these are distal outcomes that do not have direct relationships to the reasons that prompt an assessment referral. Thus, although it may be interesting to learn about these derivative effects, the sample sizes required to detect differences of this sort are likely to be huge (Sturm, Unützer, & Katon, 1999) and tangential to the core purpose of assessment. (In many respects, the mismatch in this design would be analogous to a situation where researchers tried to evaluate the effectiveness of treatment by determining how much an intervention aided differential diagnosis.)

and resources on a design that can actually answer the question.

Conclusions

Formal assessment is a vital element in psychology's professional heritage and a central part of professional practice today. This review has documented the very strong and positive evidence that already exists on the value of psychological testing and assessment for clinical practice. We have demonstrated that the validity of psychological tests is comparable to the validity of medical tests and indicated that differential limits on reimbursement for psychological and medical tests cannot be justified on the basis of the empirical evidence. We have also demonstrated that distinct assessment methods provide unique sources of data and have documented how sole reliance on a clinical interview often leads to an incomplete understanding of patients. On the basis of a large array of evidence, we have argued that optimal knowledge in clinical practice (as in research) is obtained from the sophisticated integration of information derived from a multimethod assessment battery. Finally, to advance research, we have identified critical implications that flow from the distinction between testing and assessment and have called for future investigations to focus on the practical value of assessment clinicians who provide test-informed services to patients and referral sources. We hope this review simultaneously clarifies the strong evidence that supports testing while helping to initiate new research that can further demonstrate the unique value of well-trained psychologists providing formal assessments in applied health care settings. We invite all psychologists to join us in advancing the utility of this core and distinctive aspect of our profession.

REFERENCES

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133.

Achenbach, T. M. (1991a). *Integrative guide for the 1991 CBCL/4–18, YSR, and TRF profiles.* Burlington: University of Vermont Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Child Behavior Checklist and 1991 Profile.* Burlington: University of Vermont Department of Psychiatry.

Achenbach, T. M. (1995). Empirically based assessment and taxonomy: Applications to clinical research. *Psychological Assessment, 7,* 261–274.

Achenbach, T. M. (1997). *Manual for the Young Adult Self-Report and Young Adult Behavior Checklist.* Burlington: University of Vermont Department of Psychiatry.

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

Ahmad, N., Grad, H. A., Haas, D. A., Aronson, K. J., Jokovic, A., & Locker, D. (1997). The efficacy of nonopioid analgesics for postoperative dental pain: A meta-analysis. *Anesthesia Progress, 44,* 119–126.

Aleman, A., Hijman, R., de Haan, E. H. F., & Kahn, R. S. (1999). Memory impairment in schizophrenia: A meta-analysis. *American Journal of Psychiatry, 156,* 1358–1366.

Alfirevic, Z., & Neilson, J. P. (1995). Doppler ultrasonography in high-risk pregnancies: Systematic review with meta-analysis. *American Journal of Obstetrics and Gynecology, 172,* 1379–1387.

Alterman, A. I., Snider, E. C., Cacciola, J. S., Brown, L. S., Jr., Zaballero, A., & Siddiqui, N. (1996). Evidence for response set effects in structured research interviews. *Journal of Nervous and Mental Disease, 184,* 403–410.

Amato, P. R., & Keith, B. (1991). Parental divorce and the well-being of children: A meta-analysis. *Psychological Bulletin, 110,* 26–46.

American Psychological Association. (1992). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author.

Anastopoulos, A. D., Spisto, M. A., & Maher, M. C. (1994). The WISC–III Freedom From Distractibility factor: Its utility in identifying children with attention deficit hyperactivity disorder. *Psychological Assessment, 6,* 368–371.

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

Aronen, E. T., Noam, G. G., & Weinstein, S. R. (1993). Structured diagnostic interviews and clinicians' discharge diagnoses in hospitalized adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 32,* 674–681.

Arthur, W., Barrett, G. V., & Alexander, R. A. (1991). Prediction of vehicular accident involvement: A meta-analysis. *Human Performance, 4,* 89–105.

Atkinson, L. (1986). The comparative validities of the Rorschach and MMPI: A meta-analysis. *Canadian Psychology, 27,* 238–247.

Aziz, D. C., & Barathur, R. B. (1993). Prostate-specific antigen and prostate volume: A meta-analysis of prostate cancer screening criteria. *Journal of Clinical Laboratory Analysis, 7,* 283–292.

Baer, R. A., Wetter, M. W., & Berry, D. T. R. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review, 12,* 509–525.

Bagby, R. M., Rector, N. A., Bindseil, K., Dickens, S. E., Levitan, R. D., & Kennedy, S. H. (1998). Self-report ratings and informants' ratings of personalities of depressed outpatients. *American Journal of Psychiatry, 155,* 437–438.

Barber, J. P., & Morse, J. Q. (1994). Validation of the Wisconsin Personality Disorders Inventory with the SCID–II and PDE. *Journal of Personality Disorders, 8,* 307–319.

Baron, J., & Norman, M. F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. *Educational and Psychological Measurement, 52,* 1047–1055.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Bastian, L. A., Nanda, K., Hasselblad, V., & Simel, D. L. (1998). Diagnostic efficiency of home pregnancy test kits: A meta-analysis. *Archives of Family Medicine, 7,* 465–469.

Beck, A. T., Brown, G., Berchick, R. J., Stewart, B. L., & Steer, R. A. (1990). Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients. *American Journal of Psychiatry, 147,* 190–195.

Beck, A. T., Steer, R. A., Kovacs, M., & Garrison, B. (1985). Hopelessness and eventual suicide: A 10-year prospective study of patients hospitalized with suicidal ideation. *American Journal of Psychiatry, 142,* 559–563.

Becker, D. M., Philbrick, J. T., Bachhuber, T. L., & Humphries, J. E. (1996). D-dimer testing and acute venous thromboembolism: A shortcut to accurate diagnosis? *Archives of Internal Medicine, 156,* 939–946.

Bender, D. J., Contreras, T. A., & Fahrig, L. (1998). Habitat loss and population decline: A meta-analysis of the patch size effect. *Ecology, 79,* 517–533.

Bernstein, D. P., Kasapis, C., Bergman, A., Weld, E., Mitropoulou, V., Horvath, T., Klar, H., Silverman, J., & Siever, L. J. (1997). Assessing Axis II disorders by informant interview. *Journal of Personality Disorders, 11,* 158–167.

Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11,* 585–598.

Berthelot, J.-M., Garnier, P., Glémarec, J., & Flipo, R.-M. (1998). Diagnostic value for rheumatoid arthritis of antiperinuclear factor at the 1:100 threshold: Study of 600 patients and meta-analysis of the literature. *Revue du Rhumatisme, 65,* 9–14.

Binder, L. M., Rohling, M. L., & Larrabee, G. J. (1997). A review of mild head trauma: Part I. Meta-analytic review of neuropsychological stud-

ies. *Journal of Clinical and Experimental Neuropsychology, 19,* 421–431.
Blakley, B. R., Quiñones, M. A., & Crawford, M. S. (1994). The validity of isometric strength tests. *Personnel Psychology, 47,* 247–274.
Bommer, W. H., Johnson, J. L., Rich, G., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48,* 587–605.
Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119,* 111–137.
Bonis, P. A., Ioannidis, J. P., Cappelleri, J. C., Kaplan, M. M., & Lau, J. (1997). Correlation of biochemical response to interferon alfa with histological improvement in hepatitis C: A meta-analysis of diagnostic test characteristics. *Hepatology, 26,* 1035–1044.
Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123,* 123–142.
Booth-Kewley, S., & Friedman, H. S. (1987). Psychological predictors of heart disease: A quantitative review. *Psychological Bulletin, 101,* 343–362.
Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65,* 546–553.
Bornstein, R. F. (1998). Interpersonal dependency and physical illness: A meta-analytic review of retrospective and prospective studies. *Journal of Research in Personality, 32,* 480–497.
Bornstein, R. F. (1999). Criterion validity of objective and projective dependency tests: A meta-analytic assessment of behavioral prediction. *Psychological Assessment, 11,* 48–57.
Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry and Law, 21,* 35–76.
Boult, C., Boult, L., Murphy, C., Ebbitt, B., Luptak, M., & Kane, R. L. (1994). A controlled trial of outpatient geriatric evaluation and management. *Journal of the American Geriatrics Society, 42,* 465–470.
Branca, B., Giordani, B., Lutz, T., & Saper, J. R. (1995). Self-report of cognition and objective test performance in posttraumatic headache. *Headache, 36,* 300–306.
Brown, F. H., Jr., Dodrill, C. B., Clark, T., & Zych, K. (1991). An investigation of the relationship between self-report of memory functioning and memory test performance. *Journal of Clinical Psychology, 47,* 772–777.
Bucher, H. C., & Schmidt, J. G. (1993). Does routine ultrasound scanning improve outcome in pregnancy? Meta-analysis of various outcome measures. *British Medical Journal, 307,* 13–17.
Büla, C. J., Bérod, A. C., Stuck, A. E., Alessi, C. A., Aronow, H. U., Santos-Eggimann, B., Rubenstein, L. Z., & Beck, J. C. (1999). Effectiveness of preventive in-home geriatric assessment in well functioning, community-dwelling older people: Secondary analysis of a randomized trial. *Journal of the American Geriatrics Society, 47,* 389–395.
Burgess, J. W. (1991). The Personality Inventory Scales: A self-rating clinical instrument for diagnosis of personality disorder. *Psychological Reports, 69,* 1235–1246.
Burns, R., Nichols, L. O., Graney, M. J., & Cloar, F. T. (1995). Impact of continued geriatric outpatient management on health outcomes of older veterans. *Archives of Internal Medicine, 155,* 1313–1318.
Butzlaff, R. L., & Hooley, J. M. (1998). Expressed emotion and psychiatric relapse: A meta-analysis. *Archives of General Psychiatry, 55,* 547–552.
Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin, 125,* 367–383.
Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response set on NEO-PI–R scores and their relations to external criteria. *Journal of Personality Assessment, 74,* 472–488.
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.
Campens, D., & Buntinx, F. (1997). Selecting the best renal function tests: A meta-analysis of diagnostic studies. *International Journal of Technology Assessment in Health Care, 13,* 343–356.
Carson, K. P., & Gilliard, D. J. (1993). Construct validity of the Miner Sentence Completion Scale. *Journal of Occupational and Organizational Psychology, 66,* 171–175.
Carter, A. S., Grigorenko, E. L., & Pauls, D. L. (1995). A Russian adaption of the Child Behavior Checklist: Psychometric properties and associations with child and maternal affective symptomatology and family functioning. *Journal of Abnormal Child Psychology, 23,* 661–684.
Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: I. Psychosocial characteristics. *JAMA, 259,* 2701–2707.
Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology, 43,* 1254–1269.
Christensen, D., Hadzi-Pavlovic, D., & Jacomb, P. (1991). The psychometric differentiation of dementia from normal aging: A meta-analysis. *Psychological Assessment, 3,* 147–155.
Christensen, H., Griffiths, K., MacKinnon, A., & Jacomb, P. (1997). A quantitative review of cognitive deficits in depression and Alzheimer-type dementia. *Journal of the International Neuropsychological Society, 3,* 631–651.
Christensen, H., & Mackinnon, A. (1992). Wechsler Intelligence Scale profiles in Alzheimer type dementia and healthy ageing. *International Journal of Geriatric Psychiatry, 7,* 241–246.
Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Cohen, M., Becker, M. G., & Campbell, R. (1990). Relationships among four methods of assessment of children with attention deficit–hyperactivity disorder. *Journal of School Psychology, 28,* 189–202.
Cole, D. A., Martin, J. M., Powers, B., & Truglio, R. (1996). Modeling causal relations between academic and social competence and depression: A multitrait–multimethod longitudinal study of children. *Journal of Abnormal Psychology, 105,* 258–270.
Cole, D. A., Peeke, L. G., Martin, J. M., Truglio, R., & Seroczynski, A. D. (1998). A longitudinal look at the relation between depression and anxiety in children and adolescents. *Journal of Consulting and Clinical Psychology, 66,* 451–460.
Cole, D. A., Truglio, R., & Peeke, L. (1997). Relation between symptoms of anxiety and depression in children: A multitrait–multimethod–multigroup assessment. *Journal of Consulting and Clinical Psychology, 65,* 110–119.
Collins, N. L., & Miller, L. C. (1994). Self-disclosure and liking: A meta-analytic review. *Psychological Bulletin, 116,* 457–475.
Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68,* 1152–1162.
Conde-Agudelo, A., & Kafury-Goeta, A. C. (1998). Triple-marker test as screening for Down syndrome: A meta-analysis. *Obstetrical and Gynecological Survey, 53,* 369–376.
Conley, J. J. (1985). Longitudinal stability of personality traits: A multitrait–multimethod–multioccasion analysis. *Journal of Personality and Social Psychology, 49,* 1266–1282.
Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331–360.
Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.
Costa, P. T., Jr., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology, 54,* 853–863.
Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.
Cribbs, J. B., & Niva, E. J. (2000, April). *The extent of client and therapist agreement on therapeutic constructs: A meta-analysis.* Poster session presented at the annual meeting of the Western Psychological Association, Portland, OR.
Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Crowley, S. L., Worchel, F. F., & Ash, M. J. (1992). Self-report, peer-report, and teacher-report measures of childhood depression: An analysis by item. *Journal of Personality Assessment, 59,* 189–203.

D'Agostino, R. B., Sr., Weintraub, M., Russell, H. K., Stepanians, M., D'Agostino, R. B., Jr., Cantilena, L. R., Graumlich, J. F., Maldonado, S., Honig, P., & Anello, C. (1998). The effectiveness of antihistamines in reducing the severity of runny nose and sneezing: A meta-analysis. *Clinical Pharmacology and Therapeutics, 64,* 579–596.

Damos, D. L. (1993). Using meta-analysis to compare the predictive validity of single- and multiple-task measures to flight performance. *Human Factors, 35,* 615–628.

D'Andrade, R., & Dart, J. (1990). The interpretation of r versus r^2 or why percent of variance accounted for is a poor measure of size of effect. *Journal of Quantitative Anthropology, 2,* 47–59.

Davidson, K. W. (1996). Self- and expert-reported emotion inhibition: On the utility of both data sources. *Journal of Research in Personality, 30,* 535–549.

de Ruiter, C., & Greeven, P. G. J. (2000). Personality disorders in a Dutch forensic psychiatric sample: Convergence of interview and self-report measures. *Journal of Personality Disorders, 14,* 162–170.

de Vries, S. O., Hunink, M. G. M., & Polak, J. F. (1996). Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Academic Radiology, 3,* 361–369.

Del Mar, C., Glasziou, P., & Hayem, M. (1997). Are antibiotics indicated as initial treatment for children with acute otitis media? A meta-analysis. *British Medical Journal, 314,* 1526–1529.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin, 124,* 197–229.

Di Fabio, R. P. (1996). Meta-analysis of the sensitivity and specificity of platform posturography. *Archives of Otolaryngology—Head and Neck Surgery, 122,* 150–156.

Dowson, J. H. (1992). Assessment of DSM–III–R personality disorders by self-report questionnaire: The role of informants and a screening test for co-morbid personality disorders (STCPD). *British Journal of Psychiatry, 161,* 344–352.

Dreessen, L., Hildebrand, M., & Arntz, A. (1998). Patient–informant concordance on the Structured Clinical Interview for DSM–III–R Personality Disorders (SCID–II). *Journal of Personality Disorders, 12,* 149–161.

Early Breast Cancer Trialists' Collaborative Group. (1988). Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. *New England Journal of Medicine, 319,* 1681–1692.

Edwards, A. L., & Klockars, A. J. (1981). Significant others and self-evaluation: Relationships between perceived and actual evaluations. *Personality and Social Psychology Bulletin, 7,* 244–251.

Edwards, L. K., & Edwards, A. L. (1991). A principal-components analysis of the Minnesota Multiphasic Personality Inventory factor scales. *Journal of Personality and Social Psychology, 60,* 766–772.

Eisenberg, E., Berkey, C. S., Carr, D. B., Mosteller, F., & Chalmers, T. C. (1994). Efficacy and safety of nonsteroidal antiinflammatory drugs for cancer pain: A meta-analysis. *Journal of Clinical Oncology, 12,* 2756–2765.

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

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Meyer, G. J., & Moreland, K. (2000). Problems and limitations in the use of psychological assessment in the contemporary health care delivery system. *Professional Psychology: Research and Practice, 31,* 131–140.

Ekselius, L., Lindström, E., von Knorring, L., Bodlund, O., & Kullgren, G. (1994). SCID II interviews and the SCID Screen questionnaire as diagnostic tools for personality disorders in DSM–III–R. *Acta Psychiatrica Scandinavica, 90,* 120–123.

Engelhardt, J. B., Toseland, R. W., O'Donnell, J. C., Richie, J. T., Jue, D., & Banks, S. (1996). The effectiveness and efficiency of outpatient geriatric evaluation and management. *Journal of the American Geriatrics Society, 44,* 847–856.

Epkins, C. C., & Meyers, A. W. (1994). Assessment of childhood depression, anxiety, and aggression: Convergent and discriminant validity of self-, parent-, teacher-, and peer-report measures. *Journal of Personality Assessment, 62,* 364–381.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790–806.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360–392.

Erel, O., & Burman, B. (1995). Interrelatedness of marital relations and parent–child relations: A meta-analytic review. *Psychological Bulletin, 118,* 108–132.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., Harrison, P. L., Hammer, A. L., & Primoff, E. S. (1993). *Responsible test use: Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Ezpeleta, L., de la Osa, N., Doménech, J. M., Navarro, J. B., & Losilla, J. M. (1997). Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents (DICA–R) in an outpatient sample. *Journal of Child Psychology and Psychiatry, 38,* 431–440.

Fabacher, D., Josephson, K., Pietruszka, F., Linderborn, K., Morley, J. E., & Rubenstein, L. Z. (1994). An in-home preventive assessment program for independent older adults: A randomized controlled trial. *Journal of the American Geriatrics Society, 42,* 630–638.

Fahey, M. T., Irwig, L., & Macaskill, P. (1995). Meta-analysis of Pap test accuracy. *American Journal of Epidemiology, 141,* 680–689.

Faraone, S. V., & Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a "gold standard." *American Journal of Psychiatry, 151,* 650–657.

Faron, G., Boulvain, M., Irion, O., Bernard, P.-M., & Fraser, W. D. (1998). Prediction of preterm delivery by fetal fibronectin: A meta-analysis. *Obstetrics and Gynecology, 92,* 153–158.

Feingold, A. (1988). Matching for attractiveness in romantic partners and same-sex friends: A meta-analysis and theoretical critique. *Psychological Bulletin, 104,* 226–235.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456.

Fennig, S., Craig, T. J., Lavelle, J., Kovasznay, B., & Bromet, E. J. (1994). Best-estimate versus structured interview-based diagnosis in first-admission psychosis. *Comprehensive Psychiatry, 35,* 341–348.

Fennig, S., Craig, T. J., Tanenberg-Karant, M., & Bromet, E. J. (1994). Comparison of facility and research diagnoses in first-admission psychotic patients. *American Journal of Psychiatry, 151,* 1423–1429.

Ferro, T., & Klein, D. N. (1997). Family history assessment of personality disorders: I. Concordance with direct interview and between pairs of informants. *Journal of Personality Disorders, 11,* 123–136.

Finger, M. S., & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment, 11,* 58–66.

Finn, S. E. (1982). Base rates, utilities, and DSM–III: Shortcomings of fixed-rule systems of psychodiagnosis. *Journal of Abnormal Psychology, 91,* 294–302.

Finn, S. E. (1996). Assessment feedback integrating MMPI–2 and Rorschach findings. *Journal of Personality Assessment, 67,* 543–557.

Finn, S. E., & Kamphuis, J. H. (1995). What a clinician needs to know about base rates. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 224–235). New York: Oxford University Press.

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

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9,* 374–385.

Fiore, M. C., Smith, S. S., Jorenby, D. E., & Baker, T. B. (1994). The effectiveness of the nicotine patch for smoking cessation: A meta-analysis. *JAMA, 271,* 1940–1947.

Fleischmann, K. E., Hunink, M. G. M., Kuntz, K. M., & Douglas, P. S. (1998). Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. *JAMA, 280,* 913–920.

Foltz, C., Morse, J. Q., Calvo, N., & Barber, J. P. (1997). Self- and observer ratings on the NEO–FFI in couples: Initial evidence of the psychometric properties of an observer form. *Assessment, 4,* 287–295.

Forehand, R., Frame, C. L., Wierson, M., Armistead, L., & Kempton, T. (1991). Assessment of incarcerated juvenile delinquents: Agreement across raters and approaches to psychopathology. *Journal of Psychopathology and Behavioral Assessment, 13,* 17–25.

Forster, A. A., & Leckliter, I. N. (1994). The Halstead–Reitan Neuropsychological Test Battery for older children: The effects of age versus clinical status on test performance. *Developmental Neuropsychology, 10,* 299–312.

Fossati, A., Maffei, C., Bagnato, M., Donati, D., Donini, M., Fiorilli, M., Novella, L., & Ansoldi, M. (1998). Brief communication: Criterion validity of the Personality Diagnostic Questionnaire—4+ (PDQ–4+) in a mixed psychiatric sample. *Journal of Personality Disorders, 12,* 172–178.

Fretwell, M. D., Raymond, P. M., McGarvey, S. T., Owens, N., Traines, M., Silliman, R. A., & Mor, V. (1990). The Senior Care Study: A controlled trial of a consultative/unit-based geriatric assessment program in acute care. *Journal of the American Geriatrics Society, 38,* 1073–1081.

Frick, P. J., Silverthorn, P., & Evans, C. (1994). Assessment of childhood anxiety using structured interviews: Patterns of agreement among informants and association with maternal anxiety. *Psychological Assessment, 6,* 372–379.

Fuller, H. D. (1992). The validity of cardiac output measurement by thoracic impedance: A meta-analysis. *Clinical and Investigative Medicine, 15,* 103–112.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55,* 149–158.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69,* 656–672.

Gagnon, M., Dartigues, J. F., Mazaux, J. M., Dequae, L., Letenneur, L., Giroire, J. M., & Barberger-Gateau, P. (1994). Self-reported memory complaints and memory performance in elderly French community residents: Results of the PAQUID research program. *Neuroepidemiology, 13,* 145–154.

Ganellen, R. J. (1994). Attempting to conceal psychological disturbance: MMPI defensive response sets and the Rorschach. *Journal of Personality Assessment, 63,* 423–437.

Ganellen, R. J. (1996). Comparing the diagnostic efficiency of the MMPI, MCMI–II, and Rorschach: A review. *Journal of Personality Assessment, 67,* 219–243.

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

Garb, H. N. (1994). Cognitive heuristics and biases in personality assessment. In L. Heath, R. S. Tindale, J. Edwards, E. Posavac, F. Bryant, E. Henderson, Y. Suarez-Balcazar, & J. Myers (Eds.), *Applications of heuristics and biases to social issues* (pp. 73–90). New York: Plenum.

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

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

Garrison, W. T., & Earls, F. (1985). The Child Behavior Checklist as a screening instrument for young children. *Journal of the American Academy of Child Psychiatry, 24,* 76–80.

Gass, C. S., Russell, E. W., & Hamilton, R. A. (1990). Accuracy of MMPI-based inferences regarding memory and concentration in closed-head-trauma patients. *Psychological Assessment, 2,* 175–178.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Gendreau, P., Goggin, C. E., & Law, M. A. (1997). Predicting prison misconduct. *Criminal Justice and Behavior, 24,* 414–431.

Germain, M., Knoeffel, F., Wieland, D., & Rubenstein, L. Z. (1995). A geriatric assessment and intervention team for hospital inpatients awaiting transfer to a geriatric unit: A randomized trial. *Aging: Clinical and Experimental Research, 7,* 55–60.

Gianrossi, R., Detrano, R., Colombo, A., & Froelicher, V. (1990). Cardiac fluoroscopy for the diagnosis of coronary artery disease: A meta analytic review. *American Heart Journal, 120,* 1179–1188.

Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the Guide to the Assessment of Test-Session Behavior for the WISC–III and WIAT: Possible race/ethnicity, gender, and SES effects. *Journal of School Psychology, 32,* 355–369.

Goffinet, F., Paris-Llado, J., Nisand, I., & Bréart, G. (1997). Umbilical artery Doppler velocimetry in unselected and low risk pregnancies: A review of randomized controlled trials. *British Journal of Obstetrics and Gynaecology, 104,* 425–430.

Goldberg, E. L., & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement, 52,* 1019–1027.

Goldberg, L. R., & Werts, C. E. (1966). The reliability of clinicians' judgments: A multitrait–multimethod approach. *Journal of Consulting Psychology, 30,* 199–206.

Goldstein, I., Lue, T. F., Padma-Nathan, H., Rosen, R. C., Steers, W. D., & Wicker, P. A. (1998). Oral Sildenafil in the treatment of erectile dysfunction: Sildenafil Study Group. *New England Journal of Medicine, 338,* 1397–1404.

Goodman, C. M., Cohen, V., Thornby, J., & Netscher, D. (1998). The life span of silicone gel breast implants and a comparison of mammography, ultrasonography, and magnetic resonance imaging in detecting implant rupture: A meta-analysis. *Annals of Plastic Surgery, 41,* 577–586.

Goodman, M., Lamm, S. H., Engel, A., Shepherd, C. W., Houser, O. W., & Gomez, M. R. (1997). Cortical tuber count: a biomarker indicating neurologic severity of tuberous sclerosis complex. *Journal of Child Neurology, 12,* 85–90.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line coding of observers. *Journal of Personality and Social Psychology, 74,* 1337–1349.

Graves, P. L., Phil, M., Mead, L. A., & Pearson, T. A. (1986). The Rorschach Interaction Scale as a potential predictor of cancer. *Psychosomatic Medicine, 48,* 549–563.

Greenberg, S., Smith, I. L., & Muenzen, P. M. (1995). *Executive summary: Study of the practice of licensed psychologists in the United States and Canada.* New York: Professional Examination Service.

Griffith, L. F. (1997). Surviving no-frills mental healthcare: The future of psychological assessment. *Journal of Practical Psychiatry and Behavioral Health, 3,* 255–258.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Hallan, S., & Åsberg, A. (1997). The accuracy of C-reactive protein in diagnosing acute appendicitis: A meta-analysis. *Scandinavian Journal of Clinical and Laboratory Investigation, 57,* 373–380.

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

Handwerk, M. L., Larzelere, R. E., Soper, S. H., & Friman, P. C. (1999). Parent and child discrepancies in reporting severity of problem behaviors in three out-of-home settings. *Psychological Assessment, 11,* 14–23.

Hansen, F. R., Poulsen, H., & Sørensen, K. H. (1995). A model of regular geriatric follow-up by home visits to selected patients discharged from a geriatric ward: A randomized control trial. *Aging: Clinical and Experimental Research, 7,* 202–206.

Hanson, R. K., Hunsley, J., & Parker, K. C. H. (1988). The relationship between WAIS subtest reliability, "g" loadings, and meta-analytically derived validity estimates. *Journal of Clinical Psychology, 44,* 557–563.

Harkness, A. R., Tellegen, A., & Waller, N. (1995). Differential convergence of self-report and informant data for Multidimensional Personality Questionnaire traits: Implications for the construct of negative emotionality. *Journal of Personality Assessment, 64,* 185–204.

Harlan, E., & Clark, L. A. (1999). Short forms of the Schedule for

Nonadaptive and Adaptive Personality (SNAP) for self- and collateral ratings: Development, reliability, and validity. *Assessment, 6,* 131–145.
Harris, R. D., Chalmers, J. P., Henschke, P. J., Tonkin, A., Popplewell, P. Y., Stewart, A. M., Radford, A. J., O'Brien, K. P., Bond, M. J., Harris, M. G., Turnbull, R. J., Champion, G., Hobbin, E. R., & Andrews, G. R. (1991). A randomized study of outcomes in a defined group of acutely ill elderly patients managed in a geriatric assessment unit or a general medical unit. *Australian and New Zealand Journal of Medicine, 21,* 230–234.
Hart, S. D., Forth, A. E., & Hare, R. D. (1991). The MCMI–II and psychopathy. *Journal of Personality Disorders, 5,* 318–327.
Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117,* 167–178.
Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self–peer agreement in personality judgment. *Journal of Personality and Social Psychology, 72,* 664–677.
Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment. *American Psychologist, 42,* 963–974.
Haynes, S. N., Leisen, M. B., & Blaine, D. (1997). The design of individualized behavioral treatment programs using functional analytic clinical case models. *Psychological Assessment, 9,* 334–348.
Heatley, M. K. (1999). Systematic review and meta-analysis in anatomic pathology: The value of nuclear DNA content in predicting progression in low grade CIN, the significance of the histological subtype on prognosis in cervical carcinoma. *Histology and Histopathology, 14,* 203–215.
Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58,* 47–77.
Henry, B., Moffitt, T. E., Caspi, A., Langley, J., & Silva, P. A. (1994). On the "remembrance of things past": A longitudinal evaluation of the retrospective method. *Psychological Assessment, 6,* 92–101.
Herbert, T. B., & Cohen, S. (1993). Depression and immunity: A meta-analytic review. *Psychological Bulletin, 113,* 472–486.
Herzog, A. R., & Rodgers, W. L. (1989). Age differences in memory performance and memory ratings as measured in a sample survey. *Psychology and Aging, 4,* 173–182.
Hill, R. W., Zrull, M. C., & McIntire, K. (1998). Differences between self- and peer ratings of interpersonal problems. *Assessment, 5,* 67–83.
Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment, 11,* 278–296.
Holden, G. (1991). The relationship of self-efficacy appraisals to subsequent health related outcomes: A meta-analysis. *Social Work in Health Care, 16,* 53–93.
Holt, R. R. (1986). Clinical and statistical prediction: A retrospective and would-be integrative perspective. *Journal of Personality Assessment, 50,* 376–386.
Howard, K. I. (1962). The convergent and discriminant validation of ipsative ratings from three projective instruments. *Journal of Clinical Psychology, 18,* 183–188.
Howell, W. H., McNamara, D. J., Tosca, M. A., Smith, B. T., & Gaines, J. A. (1997). Plasma lipid and lipoprotein responses to dietary fat and cholesterol: A meta-analysis. *American Journal of Clinical Nutrition, 65,* 1747–1764.
Huicho, L., Campos, M., Rivera, J., & Guerrant, R. L. (1996). Fecal screening tests in the approach to acute infectious diarrhea: A scientific overview. *Pediatric Infectious Disease Journal, 15,* 486–494.
Hummel, T. J. (1999). The usefulness of tests in clinical decisions. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation.* Boston: Allyn & Bacon.
Hunt, C., & Andrews, G. (1992). Measuring personality disorders: The use of self-report questionnaires. *Journal of Personality Disorders, 6,* 125–133.
Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.
Ie, Y. L., & Verdonschot, E. H. (1994). Performance of diagnostic systems in occlusal caries detection compared. *Community Dentistry and Oral Epidemiology, 22,* 187–191.
Irle, E. (1990). An analysis of the correlation of lesion size, localization and behavioral effects in 283 published studies of cortical and subcortical lesions in old-world monkeys. *Brain Research Reviews, 15,* 181–213.
Irvin, J. E., Bowers, C. A., Dunn, M. E., & Wang, M. C. (1999). Efficacy of relapse prevention: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 67,* 563–570.
Islam, S. S., & Schottenfeld, D. (1994). Declining FEV_1 and chronic productive cough in cigarette smokers: A 25-year prospective study of lung cancer incidence in Tecumseh, Michigan. *Cancer Epidemiology, Biomarkers and Prevention, 3,* 289–298.
Ito, T. A., Miller, N., & Pollock, V. E. (1996). Alcohol and aggression: A meta-analysis on the moderating effects of inhibitory cues, triggering events, and self-focused attention. *Psychological Bulletin, 120,* 60–82.
Jackson, D. N., Fraboni, M., & Helmes, E. (1997). MMPI–2 content scales: How much content do they measure? *Assessment, 4,* 111–117.
Jacobsberg, L., Perry, S., & Frances, A. (1995). Diagnostic agreement between the SCID–II Screening Questionnaire and the Personality Disorder Examination. *Journal of Personality Assessment, 65,* 428–433.
Jacobson, J. L., Jacobson, S. W., Sokal, R. J., Martier, S. S., Ager, J. W., & Shankaran, S. (1994). Effects of alcohol use, smoking, and illicit drug use on fetal growth in Black infants. *Journal of Pediatrics, 124,* 757–764.
Janick, P. G., Davis, J. M., Gibbons, R. D., Ericksen, S., Chang, S., & Gallagher, P. (1985). Efficacy of ECT: A meta-analysis. *American Journal of Psychiatry, 142,* 297–302.
Jensen, L. A., Onyskiw, J. E., & Prasad, N. G. N. (1998). Meta-analysis of arterial oxygen saturation monitoring by pulse oximetry in adults. *Heart and Lung, 27,* 387–408.
Jensen, P. S., Traylor, J., Xanakis, S. N., & Davis, H. (1987). Child psychopathology rating scales and interrater agreement: I. Parents' gender and psychiatric symptoms. *Journal of the American Academy of Child and Adolescent Psychiatry, 27,* 442–450.
Johansson, B., Allen-Burge, R., & Zarit, S. H. (1997). Self-reports on memory functioning in a longitudinal study of the oldest old: Relation to current, prospective, and retrospective performance. *Journal of Gerontology: Psychological Sciences, 52B,* P139–P146.
John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521–551.
John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology, 66,* 206–219.
Jorgensen, R. S., Johnson, B. T., Kolodziej, M. E., & Schreer, G. E. (1996). Elevated blood pressure and personality: A meta-analytic review. *Psychological Bulletin, 120,* 293–320.
Karpf, D. B., Shapiro, D. R., Seeman, E., Ensrud, K. E., Johnston, C. C., Adami, S., Harris, S. T., Santora, A. C., Hirsch, L. J., Oppenheimer, L., Thompson, D., & the Alendronate Osteoporosis Treatment Study Groups. (1997). Prevention of nonvertebral fractures by alendronate: A meta-analysis. *JAMA, 277,* 1159–1164.
Karppi, P., & Tilvis, R. (1995). Effectiveness of a Finnish geriatric inpatient assessment: Two-year follow up of a randomized clinical trial on community-dwelling patients. *Scandinavian Journal of Primary Health, 13,* 93–98.
Kavale, K. A., & Forness, S. R. (1984). A meta-analysis of the validity of Wechsler scale profiles and recategorizations: Patterns or parodies? *Learning Disability Quarterly, 7,* 136–156.
Kavale, K. A., & Nye, C. (1985). Parameters of learning disabilities in achievement, linguistic, neuropsychological, and social/behavioral domains. *Journal of Special Education, 19,* 443–458.
Kennedy, S. H., Katz, R., Rockert, W., Mendlowitz, S., Ralevski, E., & Clewes, J. (1995). Assessment of personality disorders in anorexia nervosa and bulimia nervosa: A comparison of self-report and structured interview methods. *Journal of Nervous and Mental Disease, 183,* 358–364.
Kim, M.-S., & Hunter, J. E. (1993). Attitude–behavior relations: A meta-analysis of attitudinal relevance and topic. *Journal of Communication, 43,* 101–142.
Klein, D. N., Ouimette, P. C., Kelly, H. S., Ferro, T., & Riso, L. P. (1994). Test–retest reliability of team consensus best-estimate diagnoses of Axis I and II disorders in a family study. *American Journal of Psychiatry, 151,* 1043–1047.
Kliegman, R. M., Madura, D., Kiwi, R., Eisenberg, I., & Yamashita, T.

(1994). Relation of maternal cocaine use to the risks of prematurity and low birth weight. *Journal of Pediatrics, 124,* 751–756.

Kline, R. B., & Lachar, D. (1992). Evaluation of age, sex, and race bias in the Personality Inventory for Children (PIC). *Psychological Assessment, 4,* 333–339.

Koestner, R., Bernieri, F., & Zuckerman, M. (1994). Self–peer agreement as a function of two kinds of trait relevance: Personal and social. *Social Behavior and Personality, 22,* 17–30.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64,* 311–337.

Korchin, S. J., & Schuldberg, D. (1981). The future of clinical assessment. *American Psychologist, 36,* 1147–1158.

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin, 21,* 58–75.

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

Kumpulainen, K., Räsänen, E., Heuttonen, L., Moilanen, I., Piha, J., Puura, K., Tamminen, T., & Almqvist, F. (1999). Children's behavioral/emotional problems: A comparison of parents' and teachers' reports for elementary school-aged children. *European Child and Adolescent Psychiatry, 8*(Suppl. 4), IV/41–IV/47.

Kurokawa, N. K. S., & Weed, N. C. (1998). Interrater agreement on the Coping Inventory for Stressful Situations (CISS). *Assessment, 5,* 93–100.

Kwok, Y., Kim, C., Grady, D., Segal, M., & Redberg, R. (1999). Meta-analysis of exercise testing to detect coronary artery disease in women. *American Journal of Cardiology, 83,* 660–666.

Lambert, W., Salzer, M. S., & Bickman, L. (1998). Clinical outcome, consumer satisfaction, and ad hoc ratings of improvement in children's mental health. *Journal of Consulting and Clinical Psychology, 66,* 270–279.

Lee, S. W., Elliott, J., & Barbour, J. D. (1994). A comparison of cross-informant behavior ratings in school-based diagnosis. *Behavioral Disorders, 19,* 87–97.

Lester, D. (1992). The dexamethasone suppression test as an indicator of suicide: A meta-analysis. *Pharmacopsychiatry, 25,* 265–270.

Lester, D. (1995). The concentration of neurotransmitter metabolites in the cerebrospinal fluid of suicidal individuals: A meta-analysis. *Pharmacopsychiatry, 28,* 45–50.

Lewin, A. (1999, April). Critics' choice: The nation's top critics rate the 100 most noteworthy films of 1998. *Premiere, 12,* 86–87.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lijmer, J. C., Mol, B. W., Heisterkamp, S., Bonsel, G. J., Prins, M. H., van der Meulen, J. H. P., & Bossuyt, P. M. M. (1999). Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA, 282,* 1061–1066.

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

Littenberg, B., Mushlin, A. I., & the Diagnostic Technology Assessment Consortium. (1992). Technetium bone scanning in the diagnosis of osteomyelitis: A meta-analysis of test performance. *Journal of General Internal Medicine, 7,* 158–163.

Little, K. B., & Shneidman, E. S. (1959). Congruencies among interpretations of psychological test and anamnestic data. *Psychological Monographs: General and Applied, 73,* 1–42.

Losier, B. J., McGrath, P. J., & Klein, R. M. (1996). Error patterns of the Continuous Performance Test in non-medicated and medicated samples of children with and without ADHD: A meta-analytic review. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37,* 971–987.

Lowman, M. G., Schwanz, K. A., & Kamphaus, R. W. (1996). WISC–III third factor: Critical measurement issues. *Canadian Journal of School Psychology, 12,* 15–22.

Lyness, S. A. (1993). Predictors of differences between Type A and B individuals in heart rate and blood pressure reactivity. *Psychological Bulletin, 114,* 266–295.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 434–452.

Malgady, R. G., Rogler, L. H., & Tryon, W. W. (1992). Issues of validity in the Diagnostic Interview Schedule. *Journal of Psychiatric Research, 26,* 59–67.

Malloy, T. E., Yarlas, A., Montvilo, R. K., & Sugarman, D. B. (1996). Agreement and accuracy in children's interpersonal perceptions: A social relations analysis. *Journal of Personality and Social Psychology, 71,* 692–702.

Mantha, S., Roizen, M. F., Barnard, J., Thisted, R. A., Ellis, J. E., & Foss, J. (1994). Relative effectiveness of four preoperative tests for predicting adverse cardiac outcomes after vascular surgery: A meta-analysis. *Anesthesia and Analgesia, 79,* 422–433.

Marlowe, D. B., Husband, S. D., Bonieskie, L. M., Kirby, K. C., & Platt, J. J. (1997). Structured interview versus self-report test vantages for the assessment of personality pathology in cocaine dependence. *Journal of Personality Disorders, 11,* 177–190.

Marshall, D., Johnell, O., & Wedel, H. (1996). Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *British Medical Journal, 312,* 1254–1259.

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. *International Journal of Aviation Psychology, 6,* 1–20.

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

McConaughy, S. H., Stanger, C., & Achenbach, T. M. (1992). Three-year course of behavioral/emotional problems in a national sample of 4- to 16-year-olds: I. Agreement among informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 31,* 932–940.

McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self-reports and ratings. *Journal of Personality and Social Psychology, 43,* 293–303.

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T., Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66,* 285–313.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

McGrath, R. E., & Ingersoll, J. (1999a). Writing a good cookbook: I. A review of MMPI high-point code system studies. *Journal of Personality Assessment, 73,* 149–178.

McGrath, R. E., & Ingersoll, J. (1999b). Writing a good cookbook: II. A synthesis of MMPI high-point code system study effect sizes. *Journal of Personality Assessment, 73,* 179–198.

McKenna, M. C., Zevon, M. A., Corn, B., & Rounds, J. (1999). Psychosocial factors and the development of breast cancer: A meta-analysis. *Health Psychology, 18,* 520–531.

McReynolds, P. (1985). Psychological assessment and clinical practice: Problems and prospects. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 4, pp. 1–30). Hillsdale, NJ: Erlbaum.

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

Meier, S. T. (1994). *The chronic crisis in psychological measurement and assessment.* San Diego, CA: Academic Press.

Meiran, N., & Jelicic, M. (1995). Implicit memory in Alzheimer's disease: A meta-analysis. *Neuropsychology, 9,* 291–303.

Merrit, R. M., Williams, M. F., James, T. H., & Porubsky, E. S. (1997). Detection of cervical metastasis: A meta-analysis comparing computed tomography with physical examination. *Archives of Otolaryngology and Head and Neck Surgery, 123,* 149–152.

Meyer, G. J. (1996a). Construct validation of scales derived from the Rorschach method: A review of issues and introduction to the Rorschach Rating Scale. *Journal of Personality Assessment, 67,* 598–628.

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

Meyer, G. J. (1997). On the integration of personality assessment methods: The Rorschach and MMPI–2. *Journal of Personality Assessment, 68,* 297–330.

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

Meyer, G. J. (in press). Distinctions among information gathering methods and implications for a refined taxonomy of psychopathology. In L. E. Beutler & M. Malik (Eds.), *Alternatives to the DSM–IV.* Washington, DC: American Psychological Association.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., Eisman, E. J., & Dies, R. R. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part 1.* Washington, DC: American Psychological Association.

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

Meyer, G. J., & Handler, L. (2000). "The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach Prognostic Rating Scale": Correction. *Journal of Personality Assessment, 74,* 504–506.

Meyer, G. J., Riethmiller, R. J., Brooks, G. D., Benoit, W. A., & Handler, L. (2000). A replication of Rorschach and MMPI–2 convergent validity. *Journal of Personality Assessment, 74,* 175–215.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology, 88,* 355–383.

Miller, S. B. (1987). A comparison of methods of inquiry. *Bulletin of the Menninger Clinic, 51,* 505–518.

Miller, T. Q., Smith, T. W., Turner, C. W., Guijarro, M. L., & Hallet, A. J. (1996). A meta-analytic review of research on hostility and physical health. *Psychological Bulletin, 119,* 322–348.

Millon, T. (1994). *Millon Clinical Multiaxial Inventory—III manual.* Minneapolis, MN: National Computer Systems.

Mirvis, S. E., Shanmuganathan, K., Miller, B. H., White, C. S., & Turney, S. Z. (1996). Traumatic aortic injury: Diagnosis with contrast-enhanced thoracic CT: Five-year experience at a major trauma center. *Radiology, 200,* 413–422.

Mitchell, M. F., Schottenfeld, D., Tortolero-Luna, G., Cantor, S. B., & Richards-Kortum, R. (1998). Colposcopy for the diagnosis of squamous intraepithelial lesions: A meta-analysis. *Obstetrics and Gynecology, 91,* 626–631.

Moffitt, T. E., Caspi, A., Krueger, R. F., Magdol, L., Margolin, G., Silva, P. A., & Sydney, R. (1997). Do partners agree about abuse in their relationship? A psychometric evaluation of interpartner agreement. *Psychological Assessment, 9,* 47–56.

Mol, B. W. J., Bayram, N., Lijmer, J. G., Wiegerinck, M. A. H. M., Bongers, M. Y., van der Veen, F., & Bossuyt, P. M. M. (1998). The performance of CA-125 measurement in the detection of endometriosis: A meta-analysis. *Fertility and Sterility, 70,* 1101–1108.

Mol, B. W. J., Lijmer, J. G., Ankum, W. M., van der Veen, F., & Bossuyt, P. M. M. (1998). The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: A meta-analysis. *Human Reproduction, 13,* 3220–3227.

Mol, B. W. J., Lijmer, J., Dijkman, B., van der Veen, F., Wertheim, P., & Bossuyt, P. M. M. (1997). The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: A meta-analysis. *Fertility and Sterility, 67,* 1031–1037.

Mol, B. W. J., Meijer, S., Yuppa, S., Tan, E., de Vries, J., Bossuyt, P. M. M., & van der Veen, F. (1998). Sperm penetration assay in predicting successful in vitro fertilization. *Journal of Reproductive Medicine, 43,* 503–508.

Moreland, K. L., Fowler, R. D., & Honaker, L. M. (1994). Future directions in the use of psychological assessment for treatment planning and outcome assessment: Predictions and recommendations. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 581–602). Hillsdale, NJ: Erlbaum.

Morita, H., Suzuki, M., & Kamoshita, S. (1990). Screening measures for detecting psychiatric disorders in Japanese secondary school children. *Journal of Child Psychology and Psychiatry, 31,* 603–617.

Moroney, J. T., Bagiella, E., Desmond, D. W., Hachinski, V. C., Mölsä, P. K., Gustafson, L., Brun, A., Fischer, P., Erkinjuntti, T., Rosen, W., Paik, M. C., & Tatemichi, T. K. (1997). Meta-analysis of the Hachinski Ischemic Score in pathologically verified dementias. *Neurology, 49,* 1096–1105.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of success. *Educational and Psychological Measurement, 55,* 309–316.

Moskowitz, D. S. (1986). Comparison of self-reports, reports by knowledgeable informants, and behavioral observation data. *Journal of Personality, 54,* 294–317.

Moskowitz, D. S. (1990). Convergence of self-reports and independent observers: Dominance and friendliness. *Journal of Personality and Social Psychology, 58,* 1096–1106.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11,* 145–165.

The Movie Times. (1999). *Movies of 1998 box office totals.* No place of publication given: Author. Retrieved December 27, 2000, from the World Wide Web: http://www.the-movie-times.com/thrsdir/moviesof98.html

Mushlin, A. I., Kouides, R. W., & Shapiro, D. E. (1998). Estimating the accuracy of screening mammography: A meta-analysis. *American Journal of Preventive Medicine, 14,* 143–153.

Mutén, E. (1991). Self-reports, spouse ratings, and psychophysiological assessment in a behavioral medicine program: An application of the five-factor model. *Journal of Personality Assessment, 57,* 449–464.

National Oceanic and Atmospheric Administration. (1999). *Climate research data: The daily historical climatology network.* Raw data retrieved December 9, 1999, from the World Wide Web: http://www.ncdc.noaa.gov/ol/climate/research/ushen/daily.html.

Naughton, B. J., Moran, M. B., Feinglass, J., Falconer, J., & Williams, M. E. (1994). Reducing hospital costs for the geriatric patient admitted from the emergency department: A randomized trial. *Journal of the American Geriatrics Society, 42,* 1045–1049.

Nebeker, R. S., Lambert, M. J., & Huefner, J. C. (1995). Ethnic differences on the Outcome Questionnaire. *Psychological Reports, 77,* 875–879.

Needleman, H. L., & Gatsonis, C. A. (1990). Low-level lead exposure and the IQ of children: A meta-analysis of modern studies. *JAMA, 263,* 673–678.

Nelson, J. C., & Davis, J. M. (1997). DST studies in psychotic depression: A meta-analysis. *American Journal of Psychiatry, 154,* 1497–1503.

Newman, M. L., & Greenway, P. (1997). Therapeutic effects of providing MMPI–2 test feedback to clients at a university counseling service: A collaborative approach. *Psychological Assessment, 9,* 122–131.

Ng, P. C., & Dear, P. R. F. (1990). The predictive value of a normal ultrasound scan in the preterm baby: A meta-analysis. *Acta Paediatrica Scandinavica, 79,* 286–291.

Nicholson, R. A., & Kugler, K. E. (1991). Competent and incompetent criminal defendants: A quantitative review of the comparative research. *Psychological Bulletin, 109,* 355–370.

Norcross, J. C., Karg, R. S., & Prochaska, J. O. (1997). Clinical psychologists in the 1990s: Part II. *Clinical Psychologist, 50,* 4–11.

Nowell, P. D., Mazumdar, S., Buysse, D. J., Dew, M. A., Reynolds, C. F., III, & Kupfer, D. F. (1997). Benzodiazepines and zolpidem for chronic insomnia: A meta-analysis of treatment efficacy. *JAMA, 278,* 2170–2177.

Nussbaum, D., & Rogers, R. (1992). Screening psychiatric patients for Axis II disorders. *Canadian Journal of Psychiatry, 37,* 658–660.

Oakland, T., & Glutting, J. J. (1990). Examiner observations of children's WISC–R test-related behaviors: Possible socioeconomic status, race, and gender effects. *Psychological Assessment, 2,* 86–90.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 1078–1085.

Oldridge, N. B., Guyatt, G. H., Fischer, M. E., & Rimm, A. A. (1988). Cardiac rehabilitation after myocardial infarction: Combined experience of randomized clinical trials. *JAMA, 260,* 945–950.

Oler, A., Whooley, M. A., Oler, J., & Grady, D. (1996). Adding heparin

to aspirin reduces the incidence of myocardial infarction and death in patients with unstable angina. *JAMA, 276,* 811–815.

Olsson, N., & Juslin, P. (1999). Can self-reported encoding strategy and recognition skill be diagnostic of performance in eyewitness identifications? *Journal of Applied Psychology, 84,* 42–49.

Oltmanns, T. F., Turkheimer, E., & Strauss, M. E. (1998). Peer assessment of personality traits and pathology in female college students. *Assessment, 5,* 53–65.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124,* 54–74.

Overholser, J. C. (1994). The personality disorders: A review and critique of contemporary assessment strategies. *Journal of Contemporary Psychotherapy, 24,* 223–243.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

Parker, G., Boyce, P., Mitchell, P., Hadzi-Pavlovic, D., Wilhelm, K., Hickie, I., & Brodaty, H. (1992). Comparison of clinician rated and family corrobative witness data for depressed patients. *Journal of Affective Disorders, 24,* 25–34.

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

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

Pastorelli, C., Barbaranelli, C., Cermak, I., Rozsa, S., & Caprara, G. V. (1997). Measuring emotional instability, prosocial behavior and aggression in pre-adolescents: A cross-national study. *Personality and Individual Differences, 23,* 691–703.

Paulhus, D. L., Aks, D. J., & Coren, S. (1990). Independence of performance and self-report measures of distractibility. *Journal of Social Psychology, 130,* 781–787.

Paulhus, D. L., Lysy, D. C., & Yik, M. S. M. (1998). Self-report measures of intelligence: Are they useful as proxy IQ tests? *Journal of Personality, 64,* 525–554.

Paulhus, D. L., & Reynolds, S. (1995). Enhancing target variance in personality impressions: Highlighting the person in person perception. *Journal of Personality and Social Psychology, 69,* 1233–1242.

Paulson, W. D., Ram, S. J., Birk, C. G., & Work, J. (1999). Does blood flow accurately predict thrombosis or failure of hemodialysis synthetic grafts? A meta-analysis. *American Journal of Kidney Diseases, 34,* 478–485.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target–rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology, 56,* 823–833.

Perry, J. C. (1992). Problems and considerations in the valid assessment of personality disorders. *American Journal of Psychiatry, 149,* 1645–1653.

Persky, V. W., Kempthorne-Rawson, J., & Shekelle, R. B. (1987). Personality and risk of cancer: 20-year follow-up of the Western Electric study. *Psychosomatic Medicine, 49,* 435–449.

Persons, J. B. (1991). Psychotherapy outcome studies do not accurately represent current models of psychotherapy: A proposed remedy. *American Psychologist, 46,* 99–106.

Peter, J. P., & Churchill, G. A., Jr. (1986). Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research, 23,* 1–10.

Petersen, J. R., Smith, E., Okorodudu, A. O., & Bissell, M. G. (1996). Comparison of four methods (L/S ratio, TDx FLM, lamellar bodies, PG) for fetal lung maturity using meta-analysis. *Clinical Laboratory Management Review, 10,* 169–175.

Phares, V., & Compas, B. E. (1990). Adolescents' subjective distress over their emotional/behavioral problems. *Journal of Consulting and Clinical Psychology, 58,* 596–603.

Phares, V., Compas, B. E., & Howell, D. C. (1989). Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. *Psychological Assessment, 1,* 68–71.

Phelps, R., Eisman, E. J., & Kohout, J. (1998). Psychological practice and managed care: Results of the CAPP practitioner survey. *Professional Psychology: Research and Practice, 29,* 31–36.

Piacentini, J., Shaffer, D., Fisher, P., Schwab-Stone, M., Davies, M., & Gioia, P. (1993). The Diagnostic Interview Schedule for Children—Revised Version (DISC–R): III. Concurrent criterion validity. *Journal of the American Academy of Child and Adolescent Psychiatry, 32,* 658–665.

Piedmont, R. L. (1994). Validation of the NEO-PI–R observer form for college students: Toward a paradigm for studying personality development. *Assessment, 1,* 259–268.

Piedmont, R. L., & Ciarrocchi, J. W. (1999). The utility of the Revised NEO Personality Inventory in an outpatient, drug rehabilitation context. *Psychology of Addictive Behaviors, 13,* 213–226.

Pilkonis, P. A., Heape, C. L., Proietti, J. M., Clark, S. W., McDavid, J. D., & Pitts, T. E. (1995). The reliability and validity of two structured diagnostic interviews for personality disorders. *Archives of General Psychiatry, 52,* 1025–1033.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment, 3,* 46–54.

Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology, 55,* 787–796.

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

Po, A. L., & Zhang, W. Y. (1998). Analgesic efficacy of ibuprofen alone and in combination with codeine or caffeine in post-surgical pain: A meta-analysis. *European Journal of Clinical Pharmacology, 53,* 303–311.

Pogge, D. L., Stokes, J. M., Frank, J., Wong, H., & Harvey, P. D. (1997). Association of MMPI validity scales and therapist ratings of psychopathology in adolescent psychiatric inpatients. *Assessment, 4,* 17–27.

Power, T. J., Andrews, T. J., Eiraldi, R. B., Doherty, B. J., Ikeda, M. J., DuPaul, G. J., & Landau, S. (1998). Evaluating attention deficit hyperactivity disorder using multiple informants: The incremental utility of combining teacher with parent reports. *Psychological Assessment, 10,* 250–260.

Psaty, B. M., Smith, N. L., Siscovick, D. S., Koepsell, T. D., Weiss, N. S., Heckbert, S. R., Lemaitre, R. N., Wagner, E. H., & Furberg, C. D. (1997). Health outcomes associated with antihypertensive therapies used as first-line agents: A systematic review and meta-analysis. *JAMA, 277,* 739–745.

Puura, K., Almqvist, F., Tamminen, T., Piha, J., Räsänen, E., Kumpulainen, K., Moilanen, I., & Koivisto, A.-M. (1998). Psychiatric disturbances among prepubertal children in Southern Finland. *Social Psychiatry and Psychiatric Epidemiology, 33,* 310–318.

Rao, J. K., Weinberger, M., Oddone, E. Z., Allen, N. B., Landsman, P., & Feussner, J. R. (1995). The role of antineutrophil cytoplasmic antibody (c-ANCA) testing in the diagnosis of Wegener granulomatosis: A literature review and meta-analysis. *Annals of Internal Medicine, 123,* 925–932.

Rapee, R. M., Barrett, P. M., Dadds, M. R., & Evans, L. E. (1994). Reliability of the DSM–III–R childhood anxiety disorders using structured interview: Interrater and parent–child agreement. *Journal of the American Academy of Child and Adolescent Psychiatry, 33,* 984–992.

Raz, S., & Raz, N. (1990). Structural brain abnormalities in the major psychoses: A quantitative review of the evidence from computerized imaging. *Psychological Bulletin, 108,* 93–108.

Reich, W., Herjanic, B., Welner, Z., & Gandhy, P. R. (1982). Development of a structured psychiatric interview for children: Agreement on diagnoses comparing child and parent interviews. *Journal of Abnormal Child Psychology, 10,* 325–336.

Reinecke, M. A., Beebe, D. W., & Stein, M. A. (1999). The third factor of the WISC–III: It's (probably) not Freedom From Distractibility. *Journal of the American Academy of Child and Adolescent Psychiatry, 38,* 322–328.

Renneberg, B., Chambless, D. L., Dowdall, D. J., Fauerbach, J. A., & Gracely, E. J. (1992). The Structured Clinical Interview for DSM–III–R, Axis II and the Millon Clinical Multiaxial Inventory: A concur-

rent validity study of personality disorders among anxious outpatients. *Journal of Personality Disorders, 6,* 117–124.

Reuben, D. B., Borok, G. M., Wolde-Tsadik, G., Ershoff, D. H., Fishman, L. K., Ambrosini, V. L., Liu, Y., Rubenstein, L. Z., & Beck, J. C. (1995). A randomized trial of comprehensive geriatric assessment in the care of hospitalized patients. *New England Journal of Medicine, 332,* 1345–1350.

Reynolds, C. R., & Kamphaus, R. W. (1998). *BASC: Behavioral Assessment for Children manual.* Circle Pines, MN: American Guidance Service.

Ribeiro, S. C. M., Tandon, R., Grunhaus, L., & Greden, J. F. (1993). The DST as a predictor of outcome in depression: A meta-analysis. *American Journal of Psychiatry, 150,* 1618–1629.

Riccio, C. A., Cohen, M. J., Hall, J., & Ross, C. M. (1997). The third and fourth factors of the WISC–III: What they don't measure. *Journal of Psychoeducational Assessment, 15,* 27–39.

Richter, P., Werner, J., Heerlein, A., Kraus, A., & Sauer, H. (1998). On the validity of the Beck Depression Inventory: A review. *Psychopathology, 31,* 160–168.

Riso, L. P., Klein, D. N., Anderson, R. L., Ouimette, P. C., & Lizardi, H. (1994). Concordance between patients and informants on the Personality Disorder Examination. *American Journal of Psychiatry, 151,* 568–573.

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology, 66,* 225–244.

Rogers, R., Salekin, R. T., & Sewell, K. W. (1999). Validation of the Millon Clinical Multiaxial Inventory for Axis II disorders: Does it meet the *Daubert* standard? *Law and Human Behavior, 23,* 425–443.

Rogers, R., Sewell, K. W., & Salekin, R. T. (1994). A meta-analysis of malingering on the MMPI–2. *Assessment, 1,* 227–237.

Rogler, L. H., Malgady, R. G., & Tryon, W. W. (1992). Evaluation of mental health: Issues of memory in the Diagnostic Interview Schedule. *Journal of Nervous and Mental Disease, 180,* 215–222.

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

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777.

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

Rosenthal, R. (1995). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice, 2,* 133–150.

Roth, P. L., BeVier, C. A., Switzer, F. S., & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81,* 548–556.

Rothbaum, F., & Weisz, J. R. (1995). Parental caregiving and child externalizing behavior in nonclinical samples: A meta-analysis. *Psychological Bulletin, 116,* 55–74.

Rubenstein, L. Z., Josephson, K. R., Harker, J. O., Miller, D. K., & Wieland, D. (1995). The Sepulveda GEU Study revisited: Long-term outcomes, use of services, and costs. *Aging (Milano), 7,* 212–217.

Rubenstein, L. Z., Stuck, A. E., Siu, A. L., & Wieland, D. (1991). Impacts of geriatric evaluation and management programs on defined outcomes: Overview of the evidence. *Journal of the American Geriatrics Society, 39,* 8S–16S.

Rubin, C. D., Sizemore, M. T., Loftis, P. A., & de Mola, N. L. (1993). A randomized, controlled trial of outpatient geriatric evaluation and management in a large public hospital. *Journal of the American Geriatrics Society, 41,* 1023–1028.

Rubio-Stipec, M., Canino, G. J., Shrout, P., Dulcan, M., Freeman, D., & Bravo, M. (1994). Psychometric properties of parents and children as informants in child psychiatry epidemiology with the Spanish Diagnostic Interview Schedule for Children (DISC–2). *Journal of Abnormal Child Psychology, 22,* 703–720.

Russell, C. J., Settoon, R. P., McGrath, R. N., Blanton, A. E., Kidwell, R. E., Lohrke, F. T., Scifres, E. L., & Danforth, G. W. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. *Journal of Applied Psychology, 79,* 163–170.

Ryan, K. J. (1998). *Heteromethod validity of self-reports, observational scales, and performance measures in the assessment of attention and impulsivity.* Unpublished master's thesis, University of Alaska Anchorage.

Salekin, R. T., Rogers, R., & Sewell, K. W. (1996). A review and meta-analysis of the Psychopathy Checklist and Psychopathy Checklist—Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice, 3,* 203–215.

Salgado, J. S. (1997). The five-factor Model of personality and job performance in the European Community. *Journal of Applied Psychology, 82,* 30–43.

Salgado, J. S. (1998). Big Five personality dimensions and job performance in army and civil occupations: A European perspective. *Human Performance, 11,* 271–288.

Scheidler, J., Hricak, H., Yu, K. K., Subak, L., & Segal, M. R. (1997). Radiological evaluation of lymph node metastases in patients with cervical cancer: A meta-analysis. *JAMA, 278,* 1096–1101.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Schwab-Stone, M. E., Shaffer, D., Dulcan, M. K., Jensen, P. S., Fisher, P., Bird, H. R., Goodman, S. H., Lahey, B. B., Lichtman, J. H., Canino, G., Rubio-Stipec, M., & Rae, D. S. (1996). Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC–2.3). *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 878–888.

Seidenberg, M., Haltiner, A., Taylor, M. A., Hermann, B. B., & Wyler, A. (1994). Development and validation of a multiple ability self-report questionnaire. *Journal of Clinical and Experimental Neuropsychology, 16,* 93–104.

Seligman, M. E. (1995). The effectiveness of psychotherapy: The *Consumer Reports* study. *American Psychologist, 50,* 965–974.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D., Jorm, A. F., Lyons, L. C., Nietzel, M. T., Prout, H. T., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Shea, V. (1985). Overview of the assessment process. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (pp. 1–10). Boston: Allyn & Bacon.

Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist, 48,* 1117–1131.

Siegman-Igra, Y., Anglim, A. M., Shapiro, D. E., Adal, K. A., Strain, B. A., & Farr, B. M. (1997). Diagnosis of vascular catheter-related bloodstream infection: A meta-analysis. *Journal of Clinical Microbiology, 35,* 928–936.

Silverman, M., Musa, D., Martin, D. C., Lave, J. R., Adams, J., & Ricci, E. M. (1995). Evaluation of outpatient geriatric assessment: A randomized multi-site trial. *Journal of the American Geriatric Society, 43,* 733–740.

Silvestri, G. A., Littenberg, B., & Colice, G. L. (1995). The clinical evaluation for detecting metastic lung cancer: A meta-analysis. *American Journal of Respiratory and Critical Care Medicine, 152,* 225–230.

Siu, A. L., Kravitz, R. L., Keeler, E., Hemmerling, K., Kington, R., Davis, J. W., Michell, A., Burton, T. M., Morgenstern, H., Beers, M. H., & Reuben, D. B. (1996). Postdischarge geriatric assessment of hospitalized frail elderly patients. *Archives of Internal Medicine, 156,* 76–81.

Smith, G. E., Petersen, R. C., Ivnik, R. J., Malec, J. F., & Tangalos, E. G. (1996). Subjective memory complaints, psychological distress, and longitudinal change in objective memory performance. *Psychology and Aging, 11,* 272–279.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Smith-Bindman, R., Kerlikowske, K., Feldstein, V. A., Subak, L., Scheidler, J., Segal, M., Brand, R., & Grady, D. (1998). Endovaginal ultrasound to exclude endometrial cancer and other abnormalities. *JAMA, 280,* 1510–1517.

Soldz, S., Budman, S., Demby, A., & Merry, J. (1993). Diagnostic agreement between the Personality Disorder Examination and the MCMI–II. *Journal of Personality Assessment, 60,* 486–499.

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

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

Spiker, D., Kraemer, H. C., Constantine, N. A., & Bryant, D. (1992). Reliability and validity of behavior problem checklists as measures of stable traits in low birth weight, premature preschoolers. *Child Development, 63,* 1481–1496.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary* (2nd ed.). New York: Oxford University Press.

Stanton, M. D., & Shadish, W. R. (1997). Outcome, attrition, and family–couples treatment for drug abuse: A meta-analysis and review of the controlled, comparative studies. *Psychological Bulletin, 122,* 170–191.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318,* 262–264.

Steiner, J. L., Tebes, J. K., Sledge, W. H., & Walker, M. L. (1995). A comparison of the Structured Clinical Interview for DSM–III–R and clinical diagnoses. *Journal of Nervous and Mental Disease, 183,* 365–369.

Sturm, R., Unützer, J., & Katon, W. (1999). Effectiveness research and implications for study design: Sample size and statistical power. *General Hospital Psychiatry, 21,* 274–283.

Sweeney, P. D., Anderson, K., & Bailey, S. (1986). Attributional style in depression: A meta-analytic review. *Journal of Personality and Social Psychology, 50,* 974–991.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121,* 371–394.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Taylor, J. L., Miller, T. P., & Tinklenberg, J. R. (1992). Correlates of memory decline: A 4-year longitudinal study of older adults with memory complaints. *Psychology and Aging, 7,* 185–193.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.

Tett, R. P., Jackson, D. N., Rothstein, M., & Reddon, J. R. (1994). Meta-analysis of personality–job performance relations: A reply to Ones, Mount, Barrick, and Hunter (1994). *Personnel Psychology, 47,* 157–172.

Thomas, D. R., Brahan, R., & Haywood, B. P. (1993). Inpatient community-based geriatric assessment reduces subsequent mortality. *Journal of the American Geriatrics Society, 41,* 101–104.

Thornton, A. E., & Raz, N. (1997). Memory impairment in multiple sclerosis: A quantitative review. *Neuropsychology, 11,* 357–366.

Towler, B., Irwig, L., Glasziou, P., Kewenter, J., Weller, D., & Silagy, C. (1998). A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, Hemoccult. *British Medical Journal, 317,* 559–565.

Treiber, F. A., & Mabe, P. A., III (1987). Child and parent perceptions of children's psychopathology in psychiatric outpatient children. *Journal of Abnormal Child Psychology, 15,* 115–124.

Trentini, M., Semeraro, S., Rossi, E., Giannandrea, E., Vanelli, M., Pandiani, G., Bardelli, E., Tassini, D., Lacetera, A., Cortesi, P., Chioma, V., Capitelli, M., & Bianchini, G. (1995). A multicenter randomized trial of comprehensive geriatric assessment and management: Experimental design, baseline data, and six-month preliminary results. *Aging (Milano), 7,* 224–233.

Trull, T. J., & Larson, S. L. (1994). External validity of two personality disorder inventories. *Journal of Personality Disorders, 8,* 96–103.

Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment, 2,* 432–437.

Tucker, G. J. (1998). Putting DSM–IV in perspective. *American Journal of Psychiatry, 155,* 159–161.

Turner, R. G., & Gilliland, L. (1977). Comparison of self-report and performance measures of attention. *Perceptual and Motor Skills, 45,* 409–410.

U.S. Department of Health and Human Services National Center for Health Statistics. (1996). *Third National Health and Nutrition Examination Survey, 1988–1994.* Hyattsville, MD: Center for Disease Control and Prevention. Retrieved from NHANES III Laboratory data file (CD-ROM, No. 76200).

Uchino, B. N., Cacioppo, J. T., & Kiecolt-Glaser, J. K. (1996). The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin, 119,* 488–531.

Van IJzendoorn, M. H. (1995). Adult attachment representations, parental responsiveness, and infant attachment: A meta-analysis on the predictive validity of the adult attachment interview. *Psychological Bulletin, 117,* 387–403.

Van IJzendoorn, M. H., & Schuengel, C. (1996). The measurement of dissociation in normal and clinical populations: Meta-analytic validation of the Dissociative Experiences Scale (DES). *Clinical Psychology Review, 16,* 365–382.

Van Rijkom, H. M., & Verdonschot, E. H. (1995). Factors involved in validity measurements of diagnostic tests for approximal caries: A meta-analysis. *Caries Research, 29,* 364–370.

Veiel, H. O. F. (1997). A preliminary profile of neuropsychological deficits associated with major depression. *Journal of Clinical and Experimental Neuropsychology, 19,* 587–603.

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analysis of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin, 122,* 231–249.

Verhulst, F. C., & Akkerhuis, G. W. (1989). Agreement between parents' and teachers' ratings of behavioral/emotional problems of children aged 4–12. *Journal of Child Psychology and Psychiatry, 30,* 123–136.

Verhulst, F. C., & van der Ende, J. (1991). Assessment of child psychopathology: Relationships between different methods, different informants and clinical judgment of severity. *Acta Psychiatrica Scandinavica, 84,* 155–159.

Verhulst, F. C., & van der Ende, J. (1992). Agreement between parents' reports and adolescents' self-reports of problem behavior. *Journal of Child Psychology and Psychiatry, 33,* 1011–1023.

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23,* 15–32.

Videbech, P. (1997). MRI findings in patients with affective disorder: A meta-analysis. *Acta Psychiatrica Scandinavica, 96,* 157–168.

Vinchur, A. J., Schippman, J. S., Switzer, F. S., III, & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83,* 586–597.

Vitiello, B., Malone, R., Buschle, P. R., Delaney, M. A., & Behar, D. (1990). Reliability of DSM–III diagnoses of hospitalized children. *Hospital and Community Psychiatry, 41,* 63–67.

Wahlbeck, K., Cheine, M., Essali, A., & Adams, C. (1999). Evidence of clozapine's effectiveness in schizophrenia: A systematic review and meta-analysis of randomized trials. *American Journal of Psychiatry, 156,* 990–999.

Watson, D., & Clark, L. A. (1991). Self-versus peer ratings of specific emotional traits: Evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology, 60,* 927–940.

Wechsler, D. (1997). *WAIS–III: Wechsler Adult Intelligence Scale—Third Edition.* San Antonio, TX: Psychological Corporation.

Weinstein, S. R., Stone, K., Noam, G. G., Grives, K., & Schwab-Stone, M. (1989). Comparison of DISC with clinicians' DSM–III diagnoses in psychiatric patients. *Journal of the American Academy of Child and Adolescent Psychiatry, 28,* 53–60.

Wells, P. S., Lensing, A. W. A., Davidson, B. L., Prins, M. H., & Hirsh, J. (1995). Accuracy of ultrasound for the diagnosis of deep venous thrombosis in asymptomatic patients after orthopedic surgery: A meta-analysis. *Annals of Internal Medicine, 122,* 47–53.

Welten, D. C., Kemper, H. C. G., Post, G. B., & Van Staveren, W. A. (1995). A meta-analysis of the effect of calcium intake on bone mass in young and middle-aged females and males. *Journal of Nutrition, 125,* 2802–2813.

Widom, C. S., & Morris, S. (1997). Accuracy of adult recollections of childhood victimization: Part 2. Childhood sexual abuse. *Psychological Assessment, 9,* 34–46.

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C., & Duncan, L. E.

(1998). Traits and motives: Toward an integration of two traditions in personality research. *Psychological Review, 105,* 230–250.
Wishart, H., & Sharpe, D. (1997). Neuropsychological aspects of multiple sclerosis: A quantitative review. *Journal of Clinical and Experimental Neuropsychology, 19,* 810–824.
Wolfe, V. V., Finch, A. J., Saylor, C. F., Blount, R. L., Pallmeyer, T. P., & Carek, D. J. (1987). Negative affectivity in children: A multitrait–multimethod investigation. *Journal of Consulting and Clinical Psychology, 55,* 245–250.
Wolraich, M. L., Wilson, D. B., & White, J. W. (1995). The effect of sugar on behavior or cognition in children. *JAMA, 274,* 1617–1621.
Wood, W., Wong, F. Y., & Chachere, J. G. (1991). Effects of media violence on viewers' aggression in unconstrained social interaction. *Psychological Bulletin, 109,* 371–383.
Yang, J., McCrae, R. R., Costa, P. T., Jr., Dai, X., Yao, S., Cai, T., & Gao, B. (1999). Cross-cultural personality assessment in psychiatric populations: The NEO-PI–R in the People's Republic of China. *Psychological Assessment, 11,* 359–368.
Yankowitz, J., Fulton, A., Williamson, R., Grant, S. S., & Budelier, W. T. (1998). Prospective evaluation of prenatal maternal serum screening for Trisomy 18. *American Journal of Obstetrics and Gynecology, 178,* 446–450.
Yusuf, S., Zucker, D., Peduzzi, P., Fisher, L. D., Takaro, T., Kennedy, J. W., Davis, K., Killip, T., Passamani, E., Norris, R., Morris, C., Mathur, V., Varnauskas, E., & Chalmers, T. C. (1994). Effect of coronary artery bypass graft surgery on survival: Overview of 10-year results from the randomized trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet, 344,* 563–570.
Zakzanis, K. K. (1998). Quantitative evidence for neuroanatomic and neuropsychological markers in dementia of the Alzheimer type. *Journal of Clinical and Experimental Neuropsychology, 20,* 259–269.
Zelinski, E. M., Gilewski, M. J., & Anthony-Bergstone, C. R. (1990). Memory Functioning Questionnaire: Concurrent validity with memory performance and self-reported memory failures. *Psychology and Aging, 5,* 388–399.
Zimmerman, M., Pfohl, B., Coryell, W., Stangl, D., & Corenthal, C. (1988). Diagnosing personality disorder in depressed patients: A comparison of patient and informant interviews. *Archives of General Psychiatry, 45,* 733–737.
Zorrilla, E. P., McKay, J. R., Luborsky, L., & Schmidt, K. (1996). Relation of stressor and depressive symptoms to clinical progression of viral illness. *American Journal of Psychiatry, 153,* 626–635.
Zuckerman, M., Bernieri, F., Koestner, R., & Rosenthal, R. (1989). To predict some of the people some of the time: In search of moderators. *Journal of Personality and Social Psychology, 57,* 279–293.
Zuckerman, M., Koestner, R., DeBoy, T., Garcia, T., Maresca, B. C., & Sartoris, J. M. (1988). To predict some of the people some of the time: A reexamination of the moderator variable approach in personality theory. *Journal of Personality and Social Psychology, 54,* 1006–1019.
Zuckerman, M., Miyake, K., Koestner, R., Baldwin, C. H., & Osborne, J. W. (1991). Uniqueness as a moderator of self–peer agreement. *Personality and Social Psychology Bulletin, 17,* 385–391.