Psychological Assessment
2001, Vol. 13, No. 4, 486-502

1040-3590/01/$5.00 DOI: 10.1037//1040-3590.13.4.486

The Hard Science of Rorschach Research: What Do We Know and Where Do We Go?

Gregory J. Meyer
University of Alaska Anchorage

Robert P. Archer
Eastern Virginia Medical School

As the final article in the Special Series on "The Utility of the Rorschach for Clinical Assessment," the authors provide an overview of this instrument's current status. They begin with a thorough review of global and focused meta-analyses, including an expanded analysis of K. C. H. Parker, R. K. Hanson, and J. Hunsley's (1988) data set, and conclude that Rorschach, Minnesota Multiphasic Personality Inventory, and IQ scales each produce roughly similar effect size magnitudes, although all tests have greater validity for some purposes than for others. Because this evidentiary foundation justifies addressing other issues, the authors build on contributions to the Special Series to identify 11 salient theoretical and empirical gaps in the Rorschach knowledge base and make recommendations for addressing these challenges to further the evolution of the Rorschach and document its strengths and inherent limitations.

As the final article in the Special Series on "The Utility of the Rorschach for Clinical Assessment," the double meaning of *hard science* in our title aptly conveys what we see as our dual agenda. The first is to help organize the basic, reproducible, quantitative, rigorous evidence addressing the Rorschach's validity. The second is to help identify the difficult challenges that face those conducting research on this instrument.

Without a doubt, progress in research is not easy in any discipline. However, two notable scientists (Diamond, 1987, 1999; Wilson, 1998) argued that the social sciences, not the physical sciences, were the true hard sciences because the variables are so complex and difficult to define and measure. As such, social science researchers must exert more concentrated and sustained effort to achieve success. Within the social sciences, one could argue that personality is one of its harder branches. Indeed, any textbook reveals a diverse array of noncohesive models (e.g., learning, evolutionary, psychodynamic, and trait) that endeavor to understand personality (e.g., Funder, 1997; Pervin & John, 1999). Although the serious study of personality began in the early 1900s, it is just in the last two decades that a reasonably broad descriptive model coalesced (e.g., McCrae & Costa, 1997). Even with this advance, however, salient psychometric and conceptual limitations remain evident (e.g., Block, 1995; Westen, 1995), not the least of which is the disconcerting propensity for a person's self-ratings to diverge substantially from ratings made by people who know him or her well (e.g., Meyer et al., 2001). Thus, many key pieces in the puzzle of personality have yet to be fit into place by the scientists striving to do so.

Within personality research, one could also argue that use of the Rorschach forms a harder domain of science. As a unique type of performance task, it is quite rare to see validity evidence proffered by correlating scales from two sets of inkblots. However, the inflated findings from these kinds of monomethod coefficients are pervasively reported for self-report inventories and cognitive ability tests. In addition, the Rorschach task itself and the meaning assigned to its scores may not always be obvious or intuitive (Hunsley & Bailey, 2001; Meyer, 1996a), and some interpretive links may seem odd (e.g., blot shading associated with anxiety; form vs. color associated with management of emotional experience). Because face validity influences attitudes and evaluations of a test's acceptability (Anastasi & Urbina, 1997), to some extent the Rorschach's limitations may be afforded less tolerance than the limitations inherent in other tests. Furthermore, it is now well documented that there is essentially no correlation between scales of similarly named constructs from the Rorschach and the Minnesota Multiphasic Personality Inventory (MMPI; e.g., Archer & Krishnamurthy, 1993a, 1993b), which are the two most frequently used and researched tests in applied practice. Although explanations have been put forth to account for these findings (e.g., Bornstein, 1998a; Meyer, 1996b), the results were not what most researchers would have predicted in advance.

The preceding highlights a broader issue that has not yet been sufficiently addressed. This concerns the Rorschach's locus of effectiveness (Hunsley & Bailey, 2001; Meyer, 1996a). Experts who write on Rorschach interpretation (e.g., Exner, 1993; Weiner, 1998), like those who write on interpreting the MMPI or other personality tests (e.g., Archer, 1997; Greene, 2000), often write as if scores from a given scale provide sufficient information to make

Gregory J. Meyer, Department of Psychology, University of Alaska Anchorage; Robert P. Archer, Department of Psychology, Eastern Virginia Medical School.

We are thankful for the extensive time and effort Kevin Parker devoted to helping us with the reanalysis of his meta-analytic data set. We also thank Steve Finn and Roger Greene for their helpful feedback on an earlier version of this article. At the request of the editor and reviewers, we have omitted some of the technical details related to the evidence presented in our first four tables. However, a document containing this information is available from Gregory J. Meyer, as is the data set used to prepare Table 1.

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

confident statements about personal characteristics and overt behavior. But just as no one mistakes the notes from a single instrument with the sound of a full band, accumulating evidence reveals that all assessment methods have inherent limitations when it comes to measuring the full scope of an individual's personality and functioning (Funder, 1997; Meyer, 1996b, in press-a; Ozer, 1999).

For instance, Meyer et al. (2001) provided 55 examples illustrating how distinct methods of assessment provide largely unique information about a person's personality and functioning. In general, cross-source correlations are in the range of .15 to .30, which clearly indicate that no single assessment measure is sufficiently comprehensive on its own to provide a complete picture of personality, unambiguous information about overt behavior, or reasonable, individualized predictions. Researchers and clinicians would be foolhardy to believe otherwise. Simultaneously, however, these data provide a clarion call to refine and advance our theories about what certain test methods can do well and what they simply cannot do.

Returning to the hard science theme, it has been obvious for thousands of years that a ruler (even a crude one, like a human foot) "works" to measure the length of an object yet provides no inherently accurate information about mass. For almost the same period of time, it has been obvious that a scale works to measure the mass of an object while providing no information about length. Perhaps because of the complexity of the task, applied personality assessment is still sorting out its rulers and scales. At present, there is not enough systematic evidence about the virtues and inherent limitations associated with self-ratings, spouse ratings, clinician ratings, Rorschach responses, Thematic Apperception Test (TAT) responses, and so on (Funder, 1997; Meyer, 1996b; Ozer, 1999). In essence, at this stage of development in our ability to understand and measure personality, the research literature provides an incomplete, multidimensional jigsaw puzzle.

Our goal in this article is to provide an overview of the emerging puzzle. Like inkblots themselves, the literature provides a complex array of evidentiary bits. Not surprisingly, contributors to this Rorschach Special Series have attended to different puzzle pieces and have emerged with substantially different impressions. To contend with these diverse conclusions, we begin at a broader level than some of the other contributors, with a thorough review of meta-analytic results addressing test validity. With these data in hand, we aim to identify the hard science that should ground all considerations of the Rorschach and indicate which pieces of the puzzle appear most properly situated and securely placed. Next, using the arguments and issues that have been raised in this Series and elsewhere, we identify gaps in the Rorschach knowledge base that appear to be in need of the most attention. Finally, we close with several recommendations for future research.

Each of us is committed to sound science, yet we have different views on the Rorschach, with public disagreements centered on how to interpret the lack of significant intercorrelations between Rorschach and MMPI scales that share similar names (e.g., Archer & Krishnamurthy, 1999; Meyer, Riethmiller, Brooks, Benoit, & Handler, 2000). Although we do not directly address that issue, by working together on this overview, we hope to provide a useful consensus blueprint for further evolution in Rorschach research and practice.

Meta-Analytic Findings

Global Meta-Analyses Examining Univariate Validity

To date, three global meta-analyses have examined the Rorschach's univariate validity (i.e., the validity of individual scales in relation to a criterion). Each examined the Rorschach and the MMPI, and one also examined the Wechsler Adult Intelligence Scale (WAIS). Atkinson (1986) and his colleagues (Atkinson, Quarrington, Alp, & Cyr, 1986) systematically sampled the Rorschach and MMPI literature. Their central findings emerged from 276 Rorschach and 237 MMPI effect sizes in which a rationale linked the predictor scale to the criterion variable as an index of test validity. The most recent meta-analysis was commissioned from Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib (1999) as a nonpartisan, expert review of the Rorschach and MMPI literature to inform this Special Series. Hiller et al. randomly sampled the 1977 to 1997 literature and used the consensus of expert judges to determine whether predictor–criterion relationships were indicative of test validity. The most extensively cited meta-analysis is Parker, Hanson, and Hunsley's (1988). They examined the reliability, stability, and validity of the Rorschach, MMPI, and WAIS using all data published in two journals over 12 years. Parker et al. coded 10,962 effects from 411 studies but then limited their published results to those from "core scales," defined as 14 MMPI scales, 14 WAIS scales, and 9 Rorschach scores in the Comprehensive System.

Garb, Florio, and Grove (1998) subsequently reanalyzed Parker et al.'s (1988) Rorschach and MMPI data. They used the same subset of studies as Parker et al. but conducted several new analyses. For instance, they eliminated studies confounded by method variance or studies in which evidence for test validity would have paradoxically corresponded to a small effect size.

An Expanded Examination of Parker et al.'s (1988) Data Set

A number of authors in this Rorschach Special Series (e.g., Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Hiller et al., 1999; Hunsley & Bailey, 1999, 2001; Weiner, 2001) expressed concerns or misgivings about Parker et al.'s (1988) meta-analysis and Garb et al.'s (1998) reanalysis of this data set. Because the results from this data set "have been widely cited to defend the validity of the Rorschach" (Garb et al., 1998, p. 402), these criticisms are not immaterial. Potential problems with the results would compromise any efforts to document the hard science behind the test.

Consequently, we undertook a detailed reexamination of Parker et al.'s (1988) data set to rectify potential confounds. Kevin Parker provided us with his complete data set, though we focused exclusively on the 199 studies Parker et al. coded as containing a hypothesized validity coefficient. Gregory J. Meyer obtained these studies and identified each hypothesized effect. Subsequently, five revisions were implemented.

First, previously published results from this data set were limited in scope, being based on a small number of core scales and the studies using them (e.g., Hiller et al., 1999; Hunsley & Bailey, 1999). To rectify this, our results included studies using any test scale. Second, Hiller et al. noted how some of Parker et al.'s

(1988) and Garb et al.'s (1998) results may have been artificially inflated by translating omnibus analysis of variance and chi-square statistics, or slightly underestimated at times by using ω^2 to translate F tests. To rectify these and other psychometric problems (e.g., use of multiple R coefficients), focused effect sizes were generated for each relevant association using standard formulas (Rosenthal, 1991; Rosnow & Rosenthal, 1996).

Third, Parker et al.'s (1988) decision to code hypothesized effects can produce paradoxical results when hypotheses do not equate with indications of test validity (Garb et al., 1998; Hiller et al., 1999). For instance, an author may reasonably hypothesize that examiner or administration factors will skew test results. However, if true, the results would not indicate the test was a valid measure of its intended construct. Conversely, if the effects were absent (indicating the test was impervious to these forms of bias), a near-zero coefficient should not be used to suggest the test was invalid. Thus, following Garb et al., we limited our analyses to effects that addressed the validity of a scale for measuring its target construct.

Fourth, because method variance is known to inflate associations (e.g., Campbell & Fiske, 1959), we used Garb et al.'s (1998) criteria to distinguish heteromethod and monomethod validity coefficients. We coded all Rorschach and MMPI associations, regardless of criterion type, but in the interest of time we only coded heteromethod WAIS effects (i.e., excluding instances when another standardized test of cognitive ability was the criterion). Finally, we corrected any coding mistakes we encountered (e.g., in the sign of an effect, *N*, or hypothesized effect vs. not).

To ensure accuracy, Kevin C. H. Parker examined changes proposed by Gregory J. Meyer (i.e., in *N*, *r*, or sign; ambiguity about computing a focused contrast; coding an effect as hypothesized; and coding an effect as indicating convergent validity). He consulted the original study to provide an answer, which in turn was then reviewed and approved by Robert P. Archer.

It is interesting to note that for all the effects coded by Parker et al. (1988) and initially by us, the original and revised data were highly correlated (i.e., $r > .92$ for N and effect size) and produced an almost identical level of overall validity (i.e., M difference = .005, Mdn difference = −.001, N = 1,260). Thus, whatever overestimates or underestimates may have been present in the initial data from omnibus statistics, problematic transformation equations, rounding, or miscoded signs, the impact was negligible and entirely random.

After applying all five of our refinements, 1,117 effects from 164 independent samples were examined. Table 1 provides summary data on the samples, effect size variability, and central tendency. Following Rosenthal, Hiller, Bornstein, Berry, and Brunell-Neuleib's (2001) and Wilcox's (1998) recommendation for creating robust central tendency estimates, we also present findings after trimming the highest and lowest 10% from each tail of the effect size distributions. The left side of Table 1 gives data for all the Rorschach and MMPI results, regardless of criteria, whereas the right side provides data for all three tests after controlling for shared method variance. Although all results are presented, we limit discussion to the 647 heteromethod effects on the right side of Table 1 (e.g., excluding data derived from the correlation of two self-report scales). For central tendency, we focus on coefficients that were not transformed to Fisher's Z prior to aggregation (see Hunter & Schmidt, 1990; Law, 1995) but include the Z-transformed mean for comparative purposes.

To parallel Atkinson (1986), we present in the upper portion of Table 1 results at the level of hypothesized effects (i.e., disregarding samples). These data indicate the validity of specific hypotheses considered in this literature.[1] As can be seen, in both the original and 10% trimmed distributions, the unweighted mean, median, and weighted mean produce similar results for each test. Validity coefficients are statistically significant and of a medium magnitude for the Rorschach (range = .27–.30), the MMPI (range = .23–.28), and the WAIS (range = .32–.36).

The lower portion of Table 1 presents findings after computing an average effect for each independent sample. For all three tests, the summary effects are statistically significant and of medium magnitude. In addition, for each test, the Fail-Safe N is greater than 2,000, indicating one would have to find more than 2,000 unpublished or unobtained studies with null results (i.e., r = .00) to add to the current findings to bring the summary coefficients down to a statistically nonsignificant level. In the unaltered and 10% trimmed distributions, the unweighted mean, median, and weighted mean produce fairly similar validity estimates for the Rorschach (range = .28–.30) and the WAIS (range = .27–.36). Similar consistency is present for the MMPI in the 10% trimmed distribution (range = .25–.34) but not in the full distribution (range = .25–.42), in which the weighted mean is substantially higher than the unweighted mean and median.

In this Special Series, authors have debated the merits of weighted and unweighted results (Garb et al., 2001; Hunsley & Bailey, 2001; Rosenthal et al., 2001), so we consider the issue in some detail. In general, weighted coefficients are superior when a meta-analysis quantifies a single predictor–criterion relationship (e.g., the Rorschach Schizophrenia Index [SCZI] and psychotic diagnoses) and methodological confounds are not correlated with sample size. All else being equal, larger studies have less sampling error and thus provide a more accurate estimate of the underlying population parameter (i.e., the true predictor–criterion relationship).

However, the issues are not straightforward for meta-analyses examining a wide array of predictor–criterion relationships (see Hunter & Schmidt, 1990, pp. 100–101 and pp. 146–150). In this context, the mathematical process of weighting by sample size gives appropriate credit to more trustworthy large-sample results but simultaneously confers greater importance to all the other features associated with the larger studies, including the topic investigated, the participants examined, and the experimental design used. Giving these factors excess weight may have negative consequences when estimating the general validity of a test be-

[1] Although many meta-analysts note complications that may result from using multiple effects from the same sample, this procedure does not systematically affect measures of central tendency (Hunter & Schmidt, 1990; Rosenthal, 1991; Tracz, Elmore, & Pohlmann, 1992). Hunter and Schmidt (1990, pp. 451–454) believed that nonindependence would produce an overestimate of true effect size variance, but simulation studies have either indicated no real impact on the standard deviation of effects (Tracz et al., 1992) or a slight underestimate when moderators are present (Martinussen & Bjørnstad, 1999). Given our focus on the central tendency of effects rather than variability, nonindependence should not pose a salient problem.

Table 1

An Expanded Meta-Analysis of the Parker, Hanson, and Hunsley (1988) Data Set Covering the Construct Validity of the Rorschach, MMPI, and WAIS From 1970 to 1981

	All effects		Unconfounded by shared method variance		
Result and measure	Rorschach	MMPI	Rorschach	MMPI	WAIS
Hypothesis level					
Sample information					
No. of hypotheses	286	727	247	296	104
Mean *N*	87.2	99.7	91.5	125.2	97.3
Median *N*	40.0	85.0	39.0	88.0	80.0
Kurtosis of *N*	14.0	135.3	11.6	64.7	38.9
Effect size dispersion					
SD	.281	.266	.261	.211	.233
25th percentile	.127	.000	.100	.108	.210
75th percentile	.500	.400	.446	.386	.479
Effect size central tendency					
Unweighted mean *r*	.303	.222	.268	.253	.355
Unweighted mean *Zr* to *r*	.332	.246	.289	.270	.383
Unweighted median *r*	.320	.210	.291	.230	.354
Weighted mean *r*	.310	.236	.288	.275	.316
10% trimmed distribution					
Unweighted mean *r*	.317	.216	.284	.243	.354
Unweighted median *r*	.320	.210	.291	.230	.354
Weighted mean *r*	.318	.222	.303	.246	.328
Sample level					
Sample information					
No. of samples	44	103	43	58	25
Total *N*	4,855	15,105	4,807	11,531	3,593
Mean *N*	110.3	146.7	111.8	198.8	143.7
Median *N*	42.5	69.0	42.0	80.5	80.0
Kurtosis of *N*	8.9	27.2	8.6	14.4	8.3
Effect size dispersion					
SD	.212	.223	.185	.245	.182
25th percentile	.135	.187	.136	.118	.238
75th percentile	.415	.522	.410	.444	.447
Effect size central tendency					
Unweighted mean *r*	.273	.357	.283	.297	.326
Unweighted mean *Zr* to *r*	.292	.391	.300	.330	.342
Unweighted median *r*	.302	.340	.298	.252	.327
Weighted mean *r*	.288	.428	.291	.422	.271
10% trimmed distribution					
Unweighted mean *r*	.284	.351	.285	.283	.341
Unweighted median *r*	.302	.340	.298	.252	.327
Weighted mean *r*	.306	.366	.304	.338	.356
Fail-Safe *N*	4,036	57,181	4,097	18,949	2,240

Note. The Fail-Safe *N* was computed from the unweighted mean *r* in the full distributions. MMPI = Minnesota Multiphasic Personality Inventory; WAIS = Wechsler Adult Intelligence Scale.

cause larger studies do not necessarily investigate topics that are more prototypic for the test (Hiller et al., 1999).

In our sample, the initial distribution of weighted and unweighted MMPI results differ noticeably because three studies used markedly larger samples to test their hypothesis. Bloom (1977) examined Masculinity–femininity (Mf) raw scores and biological sex in 1,957 Air Force recruits ($r = .78$). Lebovits and Ostfeld (1970) examined the K scale and education level in 1,805 men ($r = .42$). Newmark, Gentry, Simpson, and Jones (1978) compared a highly selected group of schizophrenic patients with a control group diagnosed by other procedures ($r = .46$, $N = 1{,}769$). These samples had an average *N* of 1,844. The three next largest samples had *N*s between 321 and 373 ($M = 347$). They examined the ability of the MMPI to (a) differentiate college students seeking vocational counseling from those seeking treatment for an emotional problem ($r = .26$; Johnson, 1970), (b) identify Veterans Affairs patients with a service-connected disability ($r = .25$; Schneider, 1979), and (c) determine suicide severity in female outpatients ($r = .13$; Pallis & Birtchnell, 1976).

Obviously, these six samples investigated very different issues (as did all those not mentioned). One could legitimately question whether the *topics* in the three largest studies should be considered 5 times more important than the next three largest studies or 23 times more important than the typical MMPI study (median $N = 80.5$). This question is particularly relevant because doing so pushes the MMPI's weighted mean (.42) very close to the 75th percentile across all MMPI effects (.44). This statistical problem disappears using the 10% trimmed distribution and is not present with the unweighted statistics. These factors, among others, suggest the MMPI's weighted mean is a less stable estimate of validity for this data set.

Summary of Global Meta-Analyses Examining Univariate Validity

With the expanded and psychometrically refined results from Parker et al.'s (1988) data set in hand, we consider the findings in light of those reported by other authors using this data set and in light of the two other global meta-analyses. Table 2 lists relevant summary information for the Rorschach, MMPI, and WAIS. Individually, all three of the meta-analytic data sets are limited and possesses methodological shortcomings. Collectively, however, their findings are more trustworthy because they used different procedures to sample the literature, select test scales, classify what constituted a meaningful finding, and aggregate information. If one looks across all of the results, Table 2 indicates that the estimates of Rorschach validity are quite consistent, hovering around r = .30. The findings are less consistent for the MMPI. Garb et al. (1998) found noticeably higher estimates than the other analyses, particularly when self-report method confounds were controlled (i.e., .55 vs. .25, .30, and .29). Dramatic differences are also seen with the WAIS. Parker et al.'s original meta-analysis included a large number of studies that evaluated WAIS validity by correlating it with other standardized cognitive tests. When these monomethod relationships are excluded, typical validity drops from .57 to .33.

Overall, when all three tests are placed on a comparable methodological footing that excludes concurrent validity yielded by an alternative test of the same type, the Rorschach, MMPI, and WAIS obtain generally similar estimates of global validity, falling in the range between .25 and .35. Although effects of this magnitude are not dramatic, they are not unimportant either. For instance, these effects are about the same size as those found for the effectiveness

Table 2
Summary of Meta-Analytic Results Examining the Global Validity of the Rorschach, MMPI, and WAIS

Study and level of aggregation	Description	Effects (k)/ Samples (K)	N	Summary mean r		
				Rorschach	MMPI	WAIS
Atkinson (1986)	1930–1980, any journal, any Rorschach scale, no method confound	k = 276	?	.36		
Hypothesis level	1960–1980, any journal, any MMPI scale, MMPI not criterion	k = 237	?		.40	
Parker et al. (1988)[a]	1970–1981 in JPA/JCP, 9 Rorschach scales, any criterion, no χ^2	K = 13	872	.37		
Citation level	1970–1981 in JPA/JCP, 14 MMPI scales, any criterion	K = 66	10,776		.43	
	1970–1981 in JPA/JCP, 14 WAIS scales, any criterion	K = 39	5,795			.57
Garb et al. (1998)[b]	Rorschach, same studies as Parker et al. but including χ^2	K = 18	1,302	.29		
Citation level	MMPI, same studies as Parker et al.	K = 66	10,776		.48	
	Rorschach, as above but no method confound, no χ^2	K = 10	656	.36		
	MMPI, as above but no method confound	K = 36	5,640		.55	
Current analysis	1970–1981 in JPA/JCP, any Rorschach scale, any criterion	k = 286	(24,952)	.33		
Hypothesis level	1970–1981 in JPA/JCP, any MMPI scale, any criterion	k = 727	(72,509)		.22	
Hypothesis level	Rorschach, as above but no method confound	k = 247	(22,597)	.27		
	MMPI, as above but no method confound	k = 296	(37,048)		.25	
	1970–1981 in JPA/JCP, any WAIS scale, no method confound	k = 104	(10,122)			.36
Sample level	1970–1981 in JPA/JCP, any Rorschach scale, any criterion	K = 44	4,855	.27		
	1970–1981 in JPA/JCP, any MMPI scale, any criterion	K = 103	15,105		.36	
Sample level	Rorschach, as above but no method confound	K = 43	4,807	.28		
	MMPI, as above but no method confound	K = 58	11,531		.30	
	1970–1981 in JPA/JCP, any WAIS scale, no method confound	K = 25	3,593			.33
Hiller et al. (1999)	1977–1997 in any journal, any Rorschach scale, any criterion	K = 30	1,713	.29		
Citation level	1977–1997 in any journal, any MMPI scale, any criterion	K = 30	4,920		.30	
	Rorschach, as above but no method confound[c]	K = 30	1,713	.29		
	MMPI, as above but no method confound[d]	K = 27	4,454		.29	

Note. *N*s in parentheses are nonindependent totals. MMPI = Minnesota Multiphasic Personality Inventory; WAIS = Wechsler Adult Intelligence Scale; JPA = *Journal of Personality Assessment;* JCP = *Journal of Clinical Psychology.*

[a] N obtained from Parker et al.'s (1988) data set. Average effect sizes were computed from their Table 2, using the corrected mean reported in Parker, Hunsley, and Hanson (1999).

[b] N was obtained from Parker et al.'s (1988) data set based on study inclusion information provided by Howard N. Garb.

[c] No studies used Rorschach scales as criterion variables, which was the definition of monomethod results for Atkinson (1986), Garb et al. (1998), and our reanalysis of Parker et al. (1988). However, Hiller et al. (1999) conducted an analysis that excluded other "projective" tests as criteria. They found Rorschach validity to be slightly higher than what we report here (r = .30, K = 27, N = 1,509).

[d] Results are from Table 9 in Hiller et al. (1999). We believe two studies should have been excluded from this analysis. If so, the unconfounded validity of the MMPI would be lower (r = .26, K = 25, N = 4,357).

of psychological, educational, and behavioral treatments, or the extent to which therapists and clients agree on treatment-related variables (see Meyer et al., 2001).

What Global Meta-Analyses Do and Do Not Indicate

Almost all of the authors contributing to this Special Series have pointed out limitations associated with the global meta-analyses, so we briefly review what they can and cannot reveal. Global meta-analyses are inherently limited because they provide diffuse information. They do not cumulatively organize evidence for specific test scales and thus fail to provide fine-grained and clinically useful information about the value of a scale in relation to specific criteria. This is a genuine limitation of global meta-analyses, and it is impossible to circumvent this shortcoming.

In addition, analyses that compare the global validity of one test with another are potentially misleading. The WAIS, MMPI, and Rorschach do not contain equivalent predictor constructs, and they have not been evaluated in relation to a common set of criterion variables. Furthermore, the studies contributing to the meta-analysis of each test have not been equated for methodological rigor or corrected for design features known to have an impact on results (e.g., reliability and range restriction; see Hunter & Schmidt, 1990). These features make relative comparisons difficult.

Nonetheless, global meta-analyses document characteristics that could not be demonstrated from a more narrowly focused review. In particular, the meta-analyses in Table 2 show that the Rorschach, MMPI, and WAIS all have positive and meaningful evidence of construct validity regardless of predictor scales, criterion variables, target populations, literature sampling strategies, or determinations about what constitutes a meaningful hypothesized effect.

So, although global meta-analyses cannot reveal everything one needs to know about test validity, the important information provided by these reviews should not be underestimated. Across journals, decades, aggregation procedures, predictor scales, criteria, and participants, reasonable hypotheses for the vast array of Rorschach, MMPI, and WAIS scales that have been empirically tested produce convincing evidence for their construct validity. In turn, this knowledge provides a global foundation that documents the hard science behind the Rorschach.

Focused Meta-Analyses on the Validity of the Rorschach, MMPI, and IQ Tests

However, the general knowledge provided by global meta-analyses is not sufficient. In conjunction with the global results, one should ask whether some scales are better for certain criteria than others. This question can be addressed in two ways. First, one can examine how the Rorschach fares in relation to alternative tests when predicting the same criterion. Second, one can examine how specific scales fare in relation to a fixed criterion, regardless of alternative tests.

To address the first issue, Table 3 presents an overview of focused meta-analyses that have compared the Rorschach with alternative predictors of the same criterion. The table includes all the published studies we know of on this topic. The results appear clear. For these particular meta-analyses, the Rorschach can validly predict a range of criterion variables and can do so about as well as alternative tests. (Note, however, that the association between Rorschach dependency and physical illness [r = .72] should be treated with considerable caution because it is based on just two samples and 56 participants.)

To address the second issue, we considered a broader pool of Rorschach meta-analyses that examined specific test scales. To place the Rorschach results in a meaningful context, we also considered results from focused meta-analyses on the MMPI, WAIS, or similar IQ tests. Findings included all relevant effects from Meyer et al.'s (2001) review of meta-analyses on psychological and medical tests and several results that were not available for that review, including Zalewski's (1989) large MMPI meta-analysis. To use the latter, we systematically sampled the largest studies in her literature to obtain standard deviation estimates.

A Summary of Focused and Global Meta-Analyses on the Rorschach, MMPI, and IQ Tests

To have all the relevant meta-analytic findings summarized in one place, we present in Table 4 results for the focused and global meta-analyses. For the latter, we generated hypothesis-level results by combining our reanalyses of Parker et al.'s (1988) data set with Atkinson's (1986) findings. For sample-level results, we combined our findings with Hiller et al.'s (1999). We did not use Parker et al.'s or Garb et al.'s (1998) results because they were less comprehensive in scope and subject to other methodological problems (Hiller et al., 1999). We also excluded the focused Rorschach-physical illness finding from Bornstein (1998b; see Table 3) because this is likely an overestimate of validity.

Table 4 reveals that Rorschach, MMPI, and IQ validity varies as a function of the predictor–criterion relationships under consideration. The Rorschach is more valid for some purposes (e.g., the SCZI to detect psychotic diagnoses) than for others (e.g., the Depression Index [DEPI] to detect depressive diagnoses). The same is true for all tests. For instance, the MMPI is better at differentiating psychiatric patients from controls than predicting prison misconduct, and IQ tests are better at differentiating patients with dementia from controls than predicting pilot success. However, no test produces consistently better or worse validity coefficients. This is as true for global estimates of validity as it is for a wide array of focused predictor–criterion relationships.[2]

Thus, the best available evidence points to two conclusions for grounding the hard science of Rorschach research. First, there is no

[2] In Table 4, the second to last effect quantifying MMPI validity scales for detecting malingered pathology is huge (r = .74). Two factors suggest it is artificially large. First, it is virtually as large as the MMPI's alternate forms reliability (i.e., written vs. computerized; r = .78). It is implausible one could detect malingering with almost the same certainty one could predict scores on one form of the test from parallel scores on an alternate form. Second, there is a substantial disparity between this validity coefficient (.74) and the next largest MMPI coefficient (.45), even though both indicate the ability of the same MMPI scales to detect malingering. The disparity appears to be due to methodological factors. The coefficient of .74 was derived largely from analog studies, in which volunteers were instructed to fake pathology, whereas the coefficient of .45 was derived from actual patients who were known to be or strongly suspected of malingering.

Table 3
Results From Focused Meta-Analyses Comparing the Rorschach to Alternative Predictors of the Same Criterion

Study and criterion/predictor scale	No. of samples	*N*	Mean *r*	
			Rorschach	Other
Bornstein (1999): Observed dependent behavior				
Rorschach Oral Dependence Scale	21	1,320	.37	
TAT Dependency Scale	4	125		.34
Blacky Picture Test Oral Dependence Scale	6	323		.50
MMPI Dependency Scale	5	320		.20
MCMI Dependency Scale	9	720		.17
EPPS Succorance Scale	9	485		.35
IDI Dependency Scale	9	424		.33
Meyer and Handler (1997, 2000) and Meyer (2000): Psychotherapy outcome				
Baseline Rorschach PRS	17	624	.45	
Baseline MMPI Ego Strength Scale	5	280		.02
Baseline IQ	6	246		.15
Incremental validity of Rorschach PRS over IQ	8	290	.36	
Romney (1990): Relatives of schizophrenic patients vs. relatives of controls				
Rorschach Communication Deviance	3	230	.22	
Lovibond Object Sorting Test Thought Processes[a]	5	464		.23
All Non-Rorschach Tests[b] of Thought Processes	11	872		.23
Bornstein (1998b): Physical illness (retrospective designs)				
Rorschach Oral Dependence Scale	2	56	.72	
Dependency by Thematic Story	4	269		.29
Dependency by DSM Interview	2	200		.09
Dependency by Self-Report Questionnaire	6	539		.18

Note. TAT = Thematic Apperception Test; MMPI = Minnesota Multiphasic Personality Inventory; MCMI = Millon Clinical Multiaxial Inventory; EPPS = Edwards Personal Preference Schedule; IDI = Interpersonal Dependency Inventory; PRS = Prognostic Rating Scale; DSM = *Diagnostic and Statistical Manual of Mental Disorders.*

[a] Romney (1990) reported results for a study he conducted using two predictor variables. We obtained an effect size for just the Lovibond Object Sorting Test based on data reported in Catts, McConaghy, Ward, Fox, and Hadzi-Pavlovic (1993). With slightly different inclusion and exclusion criteria, the meta-analysis by Catts et al. reported nearly identical validity for the Lovibond Test in relation to the same criterion (r = .24, k = 7, N = 534).

[b] Tasks included proverbs, object sorting, verbal associates, repertory grid, the TAT, and observation of structured interactions.

reason for the Rorschach to be singled out for particular criticism or specific praise. It produces reasonable validity, roughly on par with other commonly used tests. Second, validity is conditional. It varies as a function of the predictor and the criterion. While not earth-shattering, these conservative inferences can be seen as constituting the basics of what we know about the Rorschach.

Rorschach Limitations and Challenges That Must Be Addressed for Further Evolution

Given the positive foundational evidence just reviewed, it is reasonable to now look forward to the hard (i.e., difficult) science that remains. Indeed, a fair degree of consistency appeared in this Special Series about limitations and challenges facing the Rorschach. We agree with many of the challenges that have been identified and describe below those that appear most central. Although we discuss these issues in reference to the Rorschach, to varying degrees, the shortcomings, limitations, and challenges that have been identified for the Rorschach confront all instruments used in contemporary clinical practice.

Clarifying the Rorschach's Locus of Effectiveness

Perhaps the most important issue concerns the type of information the Rorschach provides. Building on the pioneering work of McClelland (1980; McClelland, Koestner, & Weinberger, 1989) and others (Kagan, 1988; Shedler, Mayman, & Manis, 1993), researchers have proposed that the Rorschach assesses implicit or underlying personality characteristics (e.g., Bornstein, 1998a; Meyer, 1996a, 1996b). This conceptualization is an initial step in the right direction. It recognizes the most consistent finding in the literature, which is the general lack of correlation between Rorschach scores and similarly named self-report scales. It also more adequately parallels the nature of much of the information derived from the task, which consists of articulated images and in vivo observations of problem-solving behavior. It is not necessary for this type of material to be represented in consciousness, so viewing the task as a measure of implicit or underlying characteristics removes these unwarranted connotations (Meyer, 1996a).

Undoubtedly, however, this view of the Rorschach will prove to be overly vague and simplistic. People certainly censor their responses (Exner, 1993) and differ in the extent to which they spontaneously engage with the task (Meyer, 1997), both of which affect the quality of obtained information. Censored or constricted protocols lack richness and thus provide much less material about implicit representations or underlying propensities.

Furthermore, to use Tellegen's (1991) notion, Rorschach variables differ in the extent to which they should penetrate into the domain of overt behavior. Consider the form quality and cognitive

Table 4
Summary Effect Sizes (r) From Focused and General Meta-Analyses Examining the Validity of the Rorschach, MMPI, and IQ Tests

Predictor and criterion	Rorschach	MMPI	IQ	*N*
1. MMPI Ego Strength scores and subsequent psychotherapy outcome		.02		280
2. Unique contribution of an MMPI high point code (vs. other codes) to relevant criteria[a]		.07		8,614
3. MMPI scores and subsequent prison misconduct		.07		17,636
4. MMPI elevations on Scales F, 6, or 8 and criminal defendant incompetency		.08		1,461
5. MMPI Scale 8 and differentiation of schizophrenic vs. depressed disorders		.12		2,435
6. Lower general cognitive ability and involvement in automobile accidents			.12	1,020
7. General intelligence and success in military pilot training			.13	15,403
8. Rorschach DEPI and detection of depressive diagnosis	.14			994
9. MMPI Scale 2 and differentiation of neurotic vs. psychotic disorders		.14		6,156
10. MMPI Scale 8 and differentiation of neurotic vs. psychotic disorders		.14		6,156
11. Baseline IQ and subsequent psychotherapy outcome			.15	246
12. MMPI Cook–Medley Hostility Scale elevations and subsequent death from all causes		.16		4,747
13. MMPI validity scales and detection of known or suspected underreported psychopathology		.18		328
14. MMPI Dependency Scale and dependent behavior		.20		320
15. Rorschach to detect thought disturbance in relatives of schizophrenic patients	.22			230
16. WISC Distractibility subscales and learning disability diagnoses			.24	(K = 54)
17. General intelligence test scores and functional effectiveness across jobs			.25	40,230
18. General validity of Rorschach studies without method confounds	.29			6,520
19. General validity of MMPI studies without method confounds		.29		15,985
20. MMPI Scale 2 and differentiation of schizophrenic vs. depressed disorders		.31		2,435
21. General validity of Rorschach hypotheses without method confounds	.32			(k = 523)
22. General validity of MMPI hypotheses (includes some method confounds)		.32		(k = 533)
23. General validity of WAIS studies without method confounds			.33	3,593
24. MMPI Scale 2 or Depression Scale and detection of depressive diagnosis		.35		2,905
25. Incremental contribution of Rorschach PRS scores over IQ to predict treatment outcome	.36			290
26. General validity of WAIS hypotheses without method confounds			.36	(k = 104)
27. Rorschach Oral Dependence Scale and dependent behavior	.37			1,320
28. MMPI validity scales to detect underreported psychopathology (primarily analog studies)		.39		2,297
29. MMPI Scale 8 and differentiation of psychiatric patients vs. controls		.42		23,747
30. Rorschach SCZI and detection of psychotic diagnosis	.44			717
31. MMPI Scale 2 and differentiation of psychiatric patients vs. controls		.44		23,747
32. WAIS IQ and obtained level of education			.44	(k = 9)
33. Rorschach PRS scores and subsequent psychotherapy outcome	.45			624
34. MMPI validity scales and detection of known or suspected malingered psychopathology		.45		771
35. Rorschach X+% and differentiation of clinical/target group from controls	.46			1,517
36. WAIS IQ subtests and differentiation of dementia from normal controls			.52	516
37. MMPI validity scales and detection of malingered psychopathology (primarily analog studies)		.74		11,204
38. MMPI basic scales: Booklet vs. computerized form		.78		732

Note. Table entries are from Meyer et al. (2001), except as follows: 5, 9, 10, 20, 29, and 31 are from Zalewski (1989); 8 and 30 are from Jørgensen et al. (2000); 11, 14, 15, and 27 are from Table 3; 18, 19, 21–23, and 26 are from Table 2; 35 is from Meyer (2001); and 24 is from Gross, Keyes, and Greene (2000). MMPI = Minnesota Multiphasic Personality Inventory; DEPI = Depression Index; WISC = Wechsler Intelligence Scale for Children; WAIS = Wechsler Adult Intelligence Scale; SCZI = Schizophrenia Index; PRS = Prognostic Rating Scale. K = number of samples; k = number of effects.
[a] The design in this research should produce results more akin to incremental validity than univariate validity.

special scores that compose the Rorschach SCZI. These variables quantify task behaviors that should have fairly clear parallels in everyday behaviors of idiosyncratic perception and disorganized thought processes. In contrast, the DEPI variables are largely based on determinants (e.g., vista, form dimension, achromatic color, blends), which intrinsically have less of a direct link to observable behavior. This is not to say they lack behavioral correlates (e.g., Netter & Viglione, 1994). Rather, there are just more inferential steps between a color-shading blend on the Rorschach, for example, and overt behaviors indicative of mixed or confused emotional reactions. Alternatively, perceiving a three-dimensional image on the inkblots does not equate with overt behaviors indicating a person views the environment in three dimensions. As models about the Rorschach's locus of effectiveness develop, it will be important to account for these kinds of score distinctions.

The number of inferential steps between a Rorschach test score and everyday observed behaviors also has implications for validation research. For instance, Table 4 demonstrates the SCZI has much stronger validity for detecting a psychotic disorder ($r = .44$) than the DEPI has for detecting a depressive disorder ($r = .14$). This can be understood as a function of differences in the predictor scales and criterion diagnoses. Psychotic diagnoses can be assigned on the basis of a patient's observed behavior even if this conflicts with his or her direct report of internal experiences. For instance, clinicians do not directly ask patients if they have delusions, disorganized speech, or grossly disorganized behavior. Patients with psychoses often do not recognize these symptoms, so clinicians trust their observations more than a patient's verbal report. In terms of test validation, this means the behaviors of perception and thought organization captured by the SCZI are synergistically aligned with the behaviors that are essential to the diagnosis.

In contrast, opposing methodological forces are generally at play when diagnosing depression. Variables on the DEPI can be viewed as more classically implicit, as they do not require the patient's conscious awareness to be evident (e.g., patients can produce Sum Shading > animal and inanimate movement even if they do not consciously experience emotional distress), and their links to overt behavior require more inferential steps. However, to diagnose depression, a patient's experience of sadness, worthlessness, guilt, and loss of interest or pleasure is essential. Observed behaviors of weight loss, insomnia, hypersomnia, tearfulness, agitation, or retardation may play a role in making the diagnosis, but it would be rare for these behaviors to prompt a diagnosis in a patient who denies the central cognitive and emotional experiences. In either case, the observed behaviors and conscious experiences that are required for a depressive diagnosis do not have direct parallels in formal Rorschach scores. No scores indicate whether a patient has increased tearfulness, early morning awakening, psychomotor agitation, or weight loss. In addition, no scores directly indicate whether a patient consciously experiences depressed mood that persists most of each day, diminished interest or pleasure in almost all activities, excessive or inappropriate guilt, or deliberate suicidal ideation. On the basis of certain Rorschach scores and responses, one could speculate about some of the latter. However, to really know whether these specific symptoms are present and represented in consciousness, one needs to ask the patient.

Thus, despite Exner's (1993) ambitions and the procedures he used to develop and revise the DEPI, the index has never worked well for children or adolescents (e.g., Archer & Krishnamurthy, 1997; Exner & Weiner, 1995), and its ability to detect depressive diagnoses in adults has not replicated well (Garb et al., 2001; Jørgensen, Andersen, & Dam, 2000; Viglione, 1999). From the beneficial perspective of hindsight, given the specific symptoms that are required for a depressive diagnosis, and given that those things cannot be determined in a direct or unequivocal way from inkblot responses, the field would have been better off if the Rorschach had not been asked to do things it does not do well. Even though it is logical to now reconceptualize the DEPI as a measure of implicit depressive propensities, this legitimately can be seen as a defensive retreat because it is based more on empirical failures than affirmations. Thus, it is essential to gather evidence that directly supports this revised view of the index (Hunsley & Bailey, 1999, 2001). Suggestive data are emerging (Jansak, 1996/1997; Renteria & Meyer, 2001), though much more empirical work is needed.

More generally, researchers and clinicians must develop a differentiated view of information-gathering methods. As Kagan (1988) argued, researchers and clinicians must stop thinking of test scales in terms of the theoretical constructs they aspire to measure (e.g., depression) without also recognizing the unique influence of the methods used to assess those constructs (e.g., self-rating vs. ward behavior vs. significant other rating vs. Rorschach behavior). We also must design creative studies that put these issues front and center (e.g., Bornstein, 1998a). Just as one should select an accountant over a bloodhound to complete tax forms but a bloodhound over an accountant to search for a child lost in the woods, we need to do a better job of documenting what distinct assessment methods do best. To the extent that we are interested in diagnoses, and to the extent that certain diagnoses emphasize the patient's self-reported conscious experience over observed behaviors, certain methods of information gathering (i.e., self-report) will be most relevant. This does not mean other data sources become irrelevant—just as bloodhounds do not become worthless when accountants are hired to complete taxes. Rather, the goal should be to empirically clarify what each procedure effectively accomplishes for applied clinical practice.

Normative Reference Groups

Consistent with recommendations and concerns found in this Special Series (Hunsley & Bailey, 2001; Viglione & Hilsenroth, 2001; Weiner, 2001), Exner has been collecting a new nonpatient reference sample for the Comprehensive System (CS). As with any test, original norms may become dated, so reference samples should be updated occasionally to ensure they are contemporary.

Exner (1993) historically has used relatively healthy and functional people as his reference standard (see Meyer, 2001). In contrast, many other tests, such as the MMPI–2/A, WAIS–III, and Personality Assessment Inventory (PAI), use a representative sample of the U.S. population as the reference standard. (Anchoring the far end of this continuum, Millon [1994] uses a clinical sample as the reference standard for his test.) A representative sample of the population includes a greater prevalence of psychiatric illness than a sample that excludes people with a history of mental health treatment. However, it is not yet clear what impact distinct reference standards have on Rorschach scores. If Exner's nonpatient samples are contrasted with those collected by others in the United States or around the world (e.g., Erdberg & Shaffer, 2001; Meyer, 2001; Shaffer, Erdberg, & Haroian, 1999; Viglione & Hilsenroth, 2001; Wood, Nezworski, Garb, & Lilienfeld, 2001), differences are evident for several variables. To more thoroughly understand these differences and the interpretive implications that flow from using a relatively healthy versus representative sample as the reference standard, a carefully defined sample that is representative of the U.S. population should also be collected for the CS. We hope this will be done in the current normative efforts.

Revised CS reference data should also explore and report potential gender, age, educational, and ethnic influences. At present, it is not certain how these factors may affect scores, although gender, age, and education effects have been observed by others (e.g., Ames, Métraux, Rodel, & Walker, 1973/1995b; Pires, 2000). In general, older and less well-educated adults produce more constricted Rorschach records.

Finally, Exner (2001) presented relatively large reference groups, including 600 nonpatient adults, 1,390 nonpatient children, 279 depressed inpatients, 328 schizophrenic inpatients, and 535 outpatients. While impressive, the large samples do not serve a clear psychometric purpose. For instance, the Wechsler scales derive IQ scores from 75 to 200 participants at each age level, and it has been suggested that 50 participants per cell provides a reasonable lower boundary for developing neuropsychological norms (Mitrushina, Boone, & D'Elia, 1999). Because CS scoring rules have regularly evolved (Meyer, 2001) and will undoubtedly continue to do so, whenever scoring modifications are introduced it is essential to rescore the reference protocols (Meyer & Richardson, 2001). Depending on the complexity of changes that may need to be made, rescoring the current reference samples would take about 1,000 to 3,000 hr, which constitutes a rather daunting task. Consequently, it would be prudent to ensure that CS reference samples are sufficiently large to provide stable data, yet not so large as to discourage regularly updated scoring.

Reliability and Adequacy of Test Administration

Scoring reliability has received considerable attention in this Special Series (Garb et al., 2001; Hunsley & Bailey, 1999, 2001; Viglione & Hilsenroth, 2001; Weiner, 2001), and we believe it has been adequately addressed (also see Meyer et al., in press). However, the consistency of test administration and inquiry across examiners has received little attention (Meyer, 1996a, 2001). Examiners may fail to establish a good working alliance, overlook key words in a response and thereby fail to inquire sufficiently, ask too many or too few questions, ask inappropriate questions, fail to record responses verbatim and thereby omit relevant information or include faulty reconstructions, and so on. These factors are not corrected through reliable scoring, yet they can have a notable impact on many scores (e.g., R, W, Dd, color, shading, Lambda, EA, es, D score, Adjusted D score, EB styles; see Exner, 1993, for an explanation of CS scores).

Clear evidence demonstrates the importance of inquiry for certain types of scores (e.g., Ritzler & Nalesnik, 1990). Nonetheless, the issue may be best illustrated with a vivid example. Two years ago, Adriana Lis and her colleagues presented adult (N = 212) and adolescent (N = 99) nonpatient data for Italy (Erdberg & Shaffer, 1999) but, after doing so, realized they had not used standard administration and inquiry procedures during data collection. Consequently, they discarded their records, obtained consultation, and began normative data collection all over again. Preliminary results for their new adult (N = 101) and adolescent (N = 51) samples were presented this year (Erdberg & Shaffer, 2001). The original adult records had Lambda values (M = 2.09) that differed markedly from Exner's (M = 0.60), though the revised sample did not (M = 0.69). Similar Lambda changes were seen in the initial and revised adolescent samples (M = 2.43 vs. 0.57, respectively). Not only does Lis's decision to throw out a large amount of hard-won data provide a model of good science to emulate, but the findings demonstrate the importance of systematically monitoring the quality of administration and inquiry during data collection.

Temporal Stability

Exner (e.g., 1993) has conducted many studies addressing CS retest reliability. These studies have examined children, adolescents, and adults; retest intervals ranging from several days to 8 years; and the impact of standard instructions versus instructions to provide new responses on the retest. Based on the available data, stability has been substantial for adults. Scores for children have been less stable, though retest coefficients increase as children mature, a finding also observed in non-CS longitudinal data sets (e.g., Ames, Métraux, Rodel, & Walker, 1974/1995a).

Garb et al. (2001) indicated that Exner has reported higher stability coefficients than other investigators. However, both Garb et al. and Viglione and Hilsenroth (2001) noted that the studies by others had methodological features that made it difficult to directly compare results (e.g., lengthy or unspecified intervals, interventions during the retest interval). Nonetheless, in an effort to more clearly understand possible investigator differences, Table 5 presents data from six adult samples. These are all the samples with reasonably complete data that have an average retest interval of 1 month or less. The samples are quite diverse methodologically, though they share a common problem, which is relatively small sample sizes. Three samples were collected by Exner and three by different investigators. Schwartz, Mebane, and Malony (1990) had 24 deaf college students take the Rorschach under two unusual formats: administration and inquiry in written form or in sign language.[3] Erstad (1995/1996) studied two samples: 11 adults age 18–55 years and 17 adults age 60–95 years. The average retest interval for the first group was 24 days. It was not specified for the older group, although presumably a similar interval was used. Haller and Exner (1985) studied two samples of inpatients with depression who received either standard instructions on retest or instructions to give different answers. Finally, Exner (1993) provided data on 35 nonpatients.

Table 5 displays results for the 23 scores used in all six samples. The last two columns provide summary information for each score, whereas the last three rows provide summary data for each sample. Examining the latter, it can be seen that Exner's (1993) samples produced slightly higher coefficients overall. However, observed coefficients often varied considerably from sample to sample. At times, one of the non-Exner samples produced unusually low results (e.g., Schwartz et al., 1990, for CF + C, Texture, Vista, Lambda, and the Affective Ratio; Erstad, 1995/1996, for Achromatic Color), but at other times it was Exner's samples that produced the unusually low findings (e.g., Animal Movement, Passive Movement, Pure Form, Popular).

Disregarding some of the substantial methodological differences across samples, Table 5 highlights problems with sampling error. As indicated by the last column, variability across samples is often quite large because 11 to 35 participants do not provide generalizable estimates of stability, particularly for those CS scores that occur infrequently. As Viglione and Hilsenroth (2001) recommended, to adequately determine retest reliability, researchers will need to obtain data from relatively large samples that are sufficiently heterogeneous in clinical characteristics to appropriately mimic the kinds of people seen in applied practice.[4]

A limitation associated with Table 5 has to do with the scores that are not included (Garb et al., 2001). Although retest findings are available for a substantially larger number of scores than those

[3] This study can be contrasted with a similar MMPI study. Brauer (1992) had 35 deaf participants respond to MMPI critical items. One videotaped individual communicated in sign language at baseline, whereas a different person was used at retest. Over the 30-min retest interval, median reliability was .475 across 34 items (range = −.06 to 1.0) and about .53 for the sum of all items.

[4] The median stability across these Rorschach scales and studies (r = .69, N = 137) is almost identical to the median 1-month stability for the Wechsler Memory Scale—III (WMS–III; r = .70, N = 297; see Wechsler, 1997). Just as significantly, the WMS manual does not report correlations for 10 subtests that have skewed distributions and narrow raw-score ranges. Because these distributions (which also occur for many Rorschach scales) may produce artificially low retest correlations, the WMS manual instead reports the percentage of people who stay in the same interpretive range on test and retest.

Table 5
The Stability of 23 Comprehensive System Scores Across Six Studies With a Retest Interval of 1 Month or Less

	Non-Exner samples			Exner samples			Row summary	
Variable	SMM	Er-1	Er-2	HE-1	HE-2	Ex	Weighted *M*	*SD*
N	24	11	17	25	25	35	137	137
Score *M* retest	1 week	24 day	?	4 day	4 day	3 week	2 week	2 week
Responses	.70	.93	.90	.72	.77	.84	.80	.10
Human Movement	.66	.75	.75	.78	.75	.83	.76	.06
Animal Movement	.62	.55	.68	.72	.28	.72	.60	.17
Inanimate Movement	.64	.73	.45	.34	.84	.34	.53	.21
Active Movement	.64	.53	.72	.78	.71	.87	.74	.12
Passive Movement	.67	.75	.75	.68	.51	.85	.70	.11
Form-Color	.61	.80	.08	.79	.70	.92	.69	.30
Color-Form (CF)	.33	.31	.42	.33	.51	.68	.46	.14
Pure Color (C)	.28	.44	.95	.43	.50	.59	.52	.23
CF + C	.26	.48	.62	.62	.74	.83	.62	.20
Sum Color	.40	.62	.75	.69	.68	.83	.67	.15
Sum Texture	−.07	.92	.70	.84	.86	.96	.70	.39
Sum Achromatic Color	.46	.15	.67	.69	.77	.67	.61	.23
Sum Diffuse Shading	.40	.07	.40	.57	.69	.41	.46	.21
Sum Vista	.00	.82	.57	.86	.82	.89	.67	.34
Pure Form	.80	.80	.80	.69	.78	.76	.77	.04
Lambda	.12	.26	.46	.40	.82	.76	.52	.28
Popular	.76	.40	.77	.37	.76	.81	.67	.20
X+%	.83	.57	.63	.72	.82	.87	.77	.12
Affective Ratio	.24	.78	.48	.67	.57	.85	.61	.22
Egocentricity	.79	.74	.60	.70	.74	.90	.76	.10
EA	.64	.61	.73	.76	.71	.84	.73	.08
es	.66	.63	.80	.78	.69	.59	.68	.08
M	.50	.59	.64	.65	.70	.77	.66	.18
Mdn	.62	.62	.68	.69	.74	.83	.69	.17
SD	.26	.24	.19	.16	.14	.16	.10	.09

Note. SMM = Schwartz, Mebane, and Malony's (1990) deaf college students tested using written versus signed administration; Er-1 = Erstad's (1995/1996) adults age 18–55; Er-2 = Erstad's elderly adults age 60–95; HE-1 = Haller and Exner's (1985) depressed inpatients instructed to give different answers on retest; HE-2 = Haller and Exner's depressed inpatients retested under standard instructions; Ex = Exner's (1993) nonpatient adults; Weighted *M* = weighted mean across samples; *SD* = standard deviation across samples; X+% = percent good form quality; EA = sum of human movement and weighted color; es = sum of non-human movement, all shading, and achromatic color.

reported in Table 5, it is still the case that evidentiary holes exist in the literature (Garb et al., 2001; Viglione & Hilsenroth, 2001). The CS contained a more limited set of core variables when Exner's stability studies were first reported. However, as the system has evolved, newer scores have emerged and become central to interpretation. Published retest findings have not kept pace. Thus, researchers should make it a priority to collect psychometrically sound retest data sets (i.e., from sufficiently large and appropriately heterogeneous samples) and report findings for all the scores found on a standard structural summary.

Understudied Variables

For whatever reason, some variables capture the fancy of researchers more than others. Within the CS, the SCZI and DEPI have probably generated the most research over the past decade. Yet many variables given fairly substantial interpretive emphasis have received little or no attention (Weiner, 2001). These include the Coping Deficit Index, Obsessive Style Index, Hypervigilance Index, active-to-passive movement ratio, D-score, food content, anatomy and X-ray content, Intellectualization Index, and Isolation Index. Recently, Exner (2001) described the Perceptual Thinking Index (PTI) as a replacement for the SCZI. Both scales share variables and have obvious empirical overlap. By deemphasizing a diagnostic label, the PTI is more appropriately aligned with the type of data that actually can be obtained from a Rorschach and a contemporary dimensional view of thought disorder (Viglione & Hilsenroth, 2001). Nonetheless, because the SCZI was the most studied and well-validated CS scale, replacing it with a new and essentially unknown commodity invites criticism. To address the latter, as well as other understudied variables, researchers should begin to systematically catalog the evidence base supporting each CS score. When data are sparse or lacking, or when conflicting findings emerge from reasonably designed studies, an organized and coordinated effort to obtain the necessary construct validation data should be implemented.

Test-Taking Styles

This issue covers all of the nontargeted factors that systematically influence test scores. On the Rorschach, the major confounding influence is associated with limited and simplistic responding versus frequent and complex responding. This kind of task engagement has been recognized as a moderator that affects how one should interpret formal test scores for decades (e.g., Rappaport, Gill, & Schafer, 1968). The major confounding influence on the MMPI is associated with a failure to recognize or report problems as opposed to a hypersensitivity to, or excessive reporting of, problems. This test-taking dimension has also been recognized as a confounding moderator for decades (e.g., Block, 1965; Edwards, 1957; Jackson, Fraboni, & Helmes, 1997). With respect to intelligence tests, only in the last 10 years has evidence accumulated regarding the confounding impact of test session behaviors on the measurement of IQ (e.g., Glutting, Youngstrom, Oakland, & Watkins, 1996; Konold, Maller, & Glutting, 1998). Avoidant, inattentive, and uncooperative behaviors are emerging as important contextual moderators; they are substantially correlated with IQ scores obtained during the testing session (average $r \approx -.34$) but not with cognitive performance outside of the test setting.

For the MMPI and Rorschach, test-taking styles are part and parcel of each test's first principal component, which typically accounts for about 50% of the total scale score variance on the MMPI and about 25% of the total variance on the Rorschach CS.[5] We do not know of studies that jointly factored IQ scales and indices of in-session testing behaviors, though they would indicate the extent to which the first factor from IQ tests (typically viewed as Spearman's *g;* see Jensen, 1998) is also intrinsically intertwined with forms of bias. Regardless of the latter, even though a few studies have explored some implications of these test-taking styles (e.g., Glutting et al., 1996; Meyer et al., 2000), their substantial association with observed test scores and potential to produce biased elevations and suppressions mandate more intensive research to understand their parameters and contend with their effects so as to maximize accurate clinical interpretation.

Unpublished Citations

Rorschach, MMPI, and Wechsler scales possess a vast published literature. Simultaneously, research that is described only in test manuals forms part of the empirical foundation for these tests (e.g., Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Exner, 1993; Wechsler, 1997) and many others (e.g., McGrew & Woodcock, 2001; Millon, 1994; Morey, 1991; Reynolds & Kamphaus, 1998).

The value of otherwise unpublished studies could certainly be debated, and the issue has generated heat in recent discussions of the Rorschach (Garb et al., 2001; Wood, Nezworski, & Stejskal, 1996a, 1996b). Exner (1996) suggested that his current textbooks list 41 otherwise unpublished studies, although Wood et al. (1996b) believed 76 may be the more accurate number. As a potentially helpful point of reference, we counted at least 68 briefly described and otherwise unpublished studies described in the WAIS–III and WMS–III manual (Wechsler, 1997).

We offer three recommendations on this issue, all of which address the broader point of ensuring a cumulative science in assessment research. Undoubtedly, some Rorschach findings will continue to be published in manuals, books, book chapters, or technical reports, and this will continue to happen for other kinds of tests as well. This is not inherently problematic, and a carefully constructed study described briefly may be more useful than the lengthy description of a poorly conceptualized project. Nonetheless, critical studies, no matter where they are published or described, must be presented with sufficient methodological and descriptive detail so that others can understand the procedures and sample and undertake replication efforts. Second, studies that form foundational pillars for understanding a variable should optimally undergo peer review and emerge in the published literature. This may be particularly important for Rorschach research, given the historical debates associated with this assessment technique. For example, candidate CS studies that appear to fit this bill include several Exner (1993) reported for the reflection and texture variables. Third, if contacted for additional details by others with legitimate interests, Rorschach researchers must strive to provide helpful and clarifying information. At times, this may even include generating new analyses or offering raw data (e.g., Exner, 1996).

Finally, it appears that a substantial portion of the Rorschach debate on this topic could have been avoided. Because of the format Exner (e.g., 1993) used to cite unpublished works (i.e., author[s], year, title, and study reference number), it is easy to assume that these citations refer to written documents. Even though we appreciate Exner's (1996) stated desire to use this format as a way to recognize the people who collaboratively helped with his research, it may be optimal to devise an alternative method for bestowing this recognition (e.g., a footnote).

Noncumulative Research

Hunsley and Bailey (1999, 2001) pointed out how much of the assessment literature (Rorschach and otherwise) is not programmatic and cumulative. Lykken (1991) argued that this sad disarray applies to almost all of psychology, such that there are few true cutting edges in our field. In many research domains, one can get by with consulting the prior literature at the point of writing an introduction to an already completed study, rather than at the planning stage when the historical literature would serve as a required guide for designing a study that would clearly fit within but incrementally extend an existing tradition of cumulative knowledge. We do not pretend to have any unique ideas for how to transform this state of affairs for assessment research, or for psychology more generally. However, we share the view with many others that cumulative research is essential.

[5] Lilienfeld et al. (2000) quoted Meyer's (1989/1991, p. 229) dissertation to argue that 70% of the variance in Rorschach scores could be attributed to error from response frequency (R). However, Meyer was discussing the first two principal components, not just the primary one. Meyer also wrongly equated R with each component, even though R had loadings of .54 on one dimension and .76 on the other, and neither were the largest loadings for either component. Finally, there is a critical difference between common variance and total variance. Meyer was referring to common variance, which is a restricted aspect of the data that only considers the degree of overlap among scores. It does not consider qualities that make one score unique from another (e.g., location scores distinct from use of color). The more important index is total variance because it includes what is common among the variables and also what is unique to each score. As described in the text, the first principal component—which encompasses much more than R and is by far the largest source of variance in the Rorschach—accounts for about 25% of the total variance in CS scores. Whatever measurement problems may legitimately be attributed to this dimension, the evidence indicates the problem is about twice as large with self-report scales like the MMPI–2 and the Millon Clinical Multiaxial Inventory—II (e.g., Jackson et al., 1997; Meyer, 1997; Meyer et al., 2000).

One way to enhance this is through focused meta-analyses that organize what is empirically known on a topic. Classifying criterion measures can be helpful (e.g., Hiller et al., 1999), though for construct validation it will be most optimal when the criterion stratification is organized around a predictor scale or a correlated class of predictor scales (e.g., the Rorschach Thought Disorder Index, Ego Impairment Index, and SCZI as predictor variables in relation to various types of psychotic criterion variables). A second and equally important way to ensure a cumulative and organized science is for editors, reviewers, academic mentors, and researchers themselves to insist that new studies meet this kind of standard. We should begin to expect that the introduction section of an article contains evidence that the relevant literature was systematically surveyed and that prior findings were cumulatively organized, preferably by effect size, not simply statistical significance.

Cross-Cultural Applications

Garb et al. (2001) and Lilienfeld, Wood, and Garb (2000) believe that evidence does not support the Rorschach's cross-cultural use. In contrast, Viglione (1999; Viglione & Hilsenroth, 2001) and Weiner (2001) believe that the preponderance of evidence does support its use, though further research is still needed. Perhaps because ethnicity is known to have a large impact on cognitive tests, ethnic bias has been studied extensively in this area. Even though there are large mean differences in IQ across ethnic groups, the tests themselves are unbiased and equally valid across ethnic backgrounds (e.g., Neisser et al., 1996). Although personality tests do not produce the large ethnic differences seen with IQ tests, well-designed studies examining the differential validity of personality tests have similarly not observed evidence of bias (e.g., Greene, 2000; Kline & Lachar, 1992; McNulty, Graham, Ben-Porath, & Stein, 1997). The only study directly addressing differential validity with the Rorschach also observed no evidence of ethnic bias (Meyer, in press-b). Thus, the existing literature does not lead to an expectation of ethnic bias in psychological tests generally, nor the Rorschach specifically. Nonetheless, this is an understudied issue with the Rorschach. Research that examines differential validity across relatively large ethnic samples would be quite valuable.

Incremental Validity

Several contributors to this Special Series indicated Rorschach scales should demonstrate incremental validity to justify use of the test (Garb et al., 2001; Hunsley & Bailey, 2001; also see Lilienfeld et al., 2000). Viglione and Hilsenroth (2001) reviewed some of the nomothetic evidence on this topic, Dawes (1999) described methods for evaluating incremental validity, and Weiner (2001) and Stricker and Gold (1999) articulated some of the distinct complexities that emerge when considering incremental validity at the idiographic level of understanding patients. With these contributions in mind, we offer several thoughts.

First, documenting the extent to which one test has incremental validity over another helpfully contributes to a differentiated understanding of test strengths and limitations. At the same time, requiring a test to demonstrate incremental validity to justify its use would form an unusually high standard at this point in the history of our science. For instance, when granting approval to a new medication, the Food and Drug Administration does not impose this standard. The new drug must be better than placebo, not an alternative drug already on the market. Similarly, the standards for empirically supported therapy do not require that a treatment be demonstrably better than a proven alternative treatment (Chambless & Ollendick, 2001).

By drawing these comparisons, we do not mean to imply it is inappropriate to consider an incremental validity standard for psychological tests. However, we should recognize that it would chart new territory for applied practice. Furthermore, this standard would have to apply across tests and assessment procedures, not just to the Rorschach, such that each would have to show it provides unique information not available from shorter and less expensive alternatives. The MMPI–2 would have to show incremental validity over a shorter alternative like the PAI, which in turn would have to show uniqueness over options like the Symptom Checklist–90, which would have to prove its worth over alternatives like the Beck Depression Inventory, which in turn would have to display incremental validity over an inexpensive alternative like the Center for Epidemiologic Studies Depression Scale or a brief interview. Similarly, the WAIS–III would have to demonstrate incremental validity over shorter and less costly alternatives, like the Kaufman Brief Intelligence Test, which in turn would have to show incremental gain over the Shipley or a brief mental status exam. As we think is obvious, this could quickly become a thorny problem.

In part, it is a thorny problem because any test can be used for multiple purposes. The MMPI may not show incremental yield over the Beck Depression Inventory when predicting depression-related criteria, but it should when predicting criteria related to psychosis. Relatedly, incremental validity is much easier to resolve when criterion validity is the central consideration (i.e., when the goal is to predict a single fixed outcome, such as successful job applicants). In this context, it is relatively straightforward to compute costs and determine which test performs most optimally. For applied clinical practice, however, construct validity is generally the central consideration (i.e., does the scale measure the construct it purports to) because the goal of assessment is descriptive; one hopes to gain a more complete understanding of the patient from the pattern of scores observed across various tests and methods (Meyer et al., 2001). In a construct validity paradigm, test scales need to be evaluated against many different types of criteria. Depending on the criterion being considered, one test may outperform another, although this relative superiority may be reversed when a different criterion is considered (see Greene, 2000, for a discussion related to MMPI subtle scales). Additionally, questions of incremental validity ignore the clinical value that may accrue when two distinct assessment methods converge on a common conclusion. Given these issues, as well as those described by Weiner (2001) and Viglione and Hilsenroth (2001), a different array of cost–benefit and incremental validity considerations comes to the forefront when the goal is accurate description in clinical practice rather than the classic nomothetic paradigm of predicting a single fixed criterion.

Simultaneously, there are still two compelling reasons to vigorously pursue incremental validity research. First, the Rorschach is a time-intensive test. Evidence demonstrating its unique clinical value would greatly help justify the time required. Second, there is a clear link between incremental validity research and understanding the Rorschach's locus of effectiveness. Advances with the latter should lead to direct tests of incremental validity (e.g., Bornstein, 1998a), and results from incremental validity analyses should directly refine our knowledge about the type of information that the Rorschach can and cannot provide.

Documenting Clinical Utility

Hunsley and Bailey (1999, 2001) advocated that clinical utility be the critical standard for judging the merits of a test. Although no one disagrees with the importance of this information, these data are absent for all educational, neuropsychological, and personality tests (e.g., Hunsley & Bailey, 1999, 2001; Viglione & Hilsenroth, 2001). We believe it is essential to collect data on the practical value of assessment clinicians (see Meyer et al., 2001) and offer two recommendations predicated on the belief that initial work in this area should closely parallel what is done in the actual practice of clinical assessment. First, at the outset, clinicians should use whatever tests they believe are indicated to address referral questions. Subsequently, to examine the unique contribution of a data source, researchers could attempt to dismantle various elements of the clinical evaluation by ensuring that certain tests (e.g., the Rorschach) are used or excluded. Second, it is essential for utility studies to use appropriate criterion measures. Because personality assessment is designed to provide helpful clinical information, criteria should focus on the extent to which patients and referral sources gain insight and have their questions answered. Although a sound assessment may help make treatment cheaper, shorter, or better than usual, the central purpose is to provide descriptive information, so criteria evaluating utility should maintain this emphasis.

Concluding Comments

Before closing, we offer a brief wish list for future research. Combined with sophisticated theorizing, we would like to see newer statistical procedures, such as item-response theory and taxometric analyses (e.g., Embretson & Reise, 2000; Waller & Meehl, 1998), elucidate the structure of Rorschach data. We would like to see research systematically address how each card pulls for certain types of responses (e.g., Ames et al., 1974/1995a) to develop refined scales that accommodate differential weighting by card. Relatedly, we would like to see scales conceptualized and iteratively improved by accounting for the specific constraints associated with the inkblots while following optimal strategies to ensure content and construct validity. Although exemplified with other types of tests, Clark and Watson (1995), Haynes, Richard, and Kubany (1995), and Smith and McCarthy (1995) provide helpful guidance. We would like clinicians and researchers to grapple with the moderating influence of task engagement on the interpretative process. This could include developing software that compares a patient with reference data before and after adjusting for task engagement (see Morey, 1982). We would like to see creative validation efforts that (a) deemphasize diagnoses as criteria (see Persons, 1986), (b) emphasize observable behavior as criteria (Hunsley & Bailey, 2001; Viglione & Hilsenroth, 2001), (c) examine understudied variables, and (d) provide insight into the Rorschach's unique and inherently limited locus of effectiveness (e.g., Bornstein, 1998a). We would like to see researchers begin the difficult task of devising paradigms that can adequately capture and evaluate the Rorschach as a tool for understanding the idiographic richness and complexity of an individual. Beliefs that the Rorschach provides illuminating insight will go unheeded by the larger scientific community unless this can be documented with evidence. Finally, we would like to see clinicians and researchers develop collaborative networks to generate systematic, multisite, cross-validated findings (Meyer, 1996a; Widiger & Schilling, 1980).

To summarize the empirical evidence, we believe the global and focused meta-analyses clearly indicate Rorschach scales can provide valid information. Like all tests, the Rorschach is more valid for some purposes than for others. Given this evidence and the limitations inherent to any assessment procedure (Meyer et al., 2001), there is no reason to single out the Rorschach for praise or criticism. Simultaneously, like many others (e.g., Garb et al., 2001), we wish to ensure the Rorschach is used optimally in applied practice. Thus, although we do not believe a moratorium should be imposed on the test (Garb et al., 2001), there should be a moratorium on certain kinds of test-based inferences. As knowledge of personality, psychopathology, and assessment methods continues to evolve and the puzzle pieces become more completely fit in place, greatest trust should be placed on inferences that are most directly supported by replicated empirical findings. In addition, we believe Neil Jacobson's credo provides appropriate direction for clinicians on the front line: "Don't do things that are directly contradicted by empirical evidence, especially when there are empirically supported alternatives" (cited in Beutler & Harwood, 2001, p. 47). In the context of assessment, although we do not believe this credo means *any* inference is legitimate so long as it is not contradicted by evidence, it does mean that we should not tolerate test-based inferences that the evidence indicates are wrong, particularly when there are more valid ways to derive those inferences. For instance, the evidence clearly indicates that psychologists should not use the DEPI on its own to diagnose a major depressive disorder from the *Diagnostic and Statistical Manual of Mental Disorders* (e.g., 4th ed.; American Psychiatric Association, 1994). The diagnosis should be based on more valid methods for ascertaining whether explicit criteria have been met. As an alternative example, because it is impossible to determine whether a specific historical event actually did or did not happen from Rorschach responses, clinicians should not draw positive or negative conclusions about sexual abuse from the Rorschach (see Kamphuis, Kugeares, & Finn, 2000). To optimally serve patients, no less is required. Beyond their potential to harm patients, undisciplined or incorrect inferences discredit tests and the profession more generally.

Simultaneously, to optimally serve patients as well as the training of students, clinical and academic psychologists should not withhold or ignore tests with demonstrated validity. Not every patient undergoing an evaluation will need or benefit from a Rorschach. But many will, and the test should be used for them. As with any test (Meyer et al., 2001), clinicians choosing the Rorschach should be able to articulate a rationale explaining why it is likely to be valuable with a particular patient who presents a distinctive set of referral questions to be addressed. This Special Series has articulated important limitations and strengths for the Rorschach that should assist with this determination. The same data and issues serve as a foundation for researchers to further evolve an integrative and differentiated hard science of personality assessment that includes what the Rorschach validly brings to the table.

References

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

Ames, L. B., Métraux, R., Rodel, J. L., & Walker, R. (1995a). *Child Rorschach responses: Developmental trends from 2 to 10 years* (Rev. ed.). Northvale, NJ: Jason Aronson. (Original work published 1974)

Ames, L. B., Métraux, R., Rodel, J. L., & Walker, R. (1995b). *Rorschach responses in old age* (2nd ed.). Northvale, NJ: Jason Aronson. (Original work published 1973)

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Macmillan.

Archer, R. P. (1997). *MMPI–A: Assessing adolescent psychopathology* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

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

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

Archer, R. P., & Krishnamurthy, R. (1999). A reply to Meyer on the convergent validity of the MMPI and Rorschach. *Journal of Personality Assessment, 73,* 319–321.

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

Atkinson, L., Quarrington, B., Alp, I. E., & Cyr, J. J. (1986). Rorschach validity: An empirical approach to the literature. *Journal of Clinical Psychology, 42,* 360–362.

Beutler, L. E., & Harwood, T. M. (2001). Antiscientific attitudes: What happens when scientists are unscientific. *Journal of Clinical Psychology, 57,* 43–51.

Block, J. D. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI.* New York: Appleton-Century-Crofts.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117,* 187–215.

Bloom, W. (1977). Relevant MMPI norms for young adult Air Force trainees. *Journal of Personality Assessment, 41,* 505–510.

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

Bornstein, R. F. (1998b). 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.

Brauer, B. A. (1992). The signer effect on MMPI performance of deaf respondents. *Journal of Personality Assessment, 58,* 380–388.

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

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

Catts, S. V., McConaghy, N., Ward, P. B., Fox, A. M., & Hadzi-Pavlovic, D. (1993). Allusive thinking in parents of schizophrenics: Meta-analysis. *Journal of Nervous and Mental Disease, 181,* 298–302.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52,* 685–716.

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

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

Diamond, J. (1987). Soft sciences are often harder than hard sciences. *Discover, 8,* 34–39.

Diamond, J. (1999). *Guns, germs, and steel: The fates of human societies.* New York: Norton.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Erdberg, P., & Shaffer, T. W. (Chairs). (1999, July). *International symposium on Rorschach nonpatient data: Findings from around the world I, II, III.* Symposium conducted at the XVIth Congress of the International Rorschach Society, Amsterdam, The Netherlands.

Erdberg, P., & Shaffer, T. W. (Chairs). (2001, March). *An international symposium on Rorschach nonpatient data: Worldwide findings. I & II.* Symposium conducted at the annual convention of the Society for Personality Assessment, Philadelphia, PA.

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

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

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

Exner, J. E., Jr. (2001). *A Rorschach workbook for the Comprehensive System* (5th ed.). Asheville, NC: Rorschach Workshops.

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

Funder, D. C. (1997). *The personality puzzle.* New York: Norton.

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., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13,* 433–448.

Glutting, J. J., Youngstrom, E. A., Oakland, T., & Watkins, M. W. (1996). Situational specificity and generality of test behaviors for samples of normal and referred children. *School Psychology Review, 25,* 94–107.

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

Gross, K., Keyes, M. D., & Greene, R. L. (2000). Assessing depression with the MMPI and MMPI–2. *Journal of Personality Assessment, 75,* 464–477.

Haller, N., & Exner, J. E., Jr. (1985). The reliability of Rorschach variables for inpatients presenting symptoms of depression and/or helplessness. *Journal of Personality Assessment, 49,* 516–521.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

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.

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach:

Unfulfilled promises and an uncertain future. *Psychological Assessment, 11,* 266–277.

Hunsley, J., & Bailey, J. M. (2001). Whither the Rorschach? An analysis of the evidence. *Psychological Assessment, 13,* 472–485.

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

Jackson, D. N., Fraboni, M., & Helmes, E. (1997). MMPI-2 content scales: How much content do they measure? *Assessment, 4,* 111–117.

Jansak, D. M. (1997). The Rorschach Comprehensive System Depression Index, depression heterogeneity, and the role of self-schema (Doctoral dissertation, California School of Professional Psychology, San Diego, 1996). *Dissertation Abstracts International, 57,* 6576B.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Johnson, R. W. (1970). A configural scoring of the MMPI and diagnosis in counseling. *Journal of Clinical Psychology, 26,* 84–86.

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

Kagan, J. (1988). The meaning of personality predicates. *American Psychologist, 43,* 614–620.

Kamphuis, J. H., Kugeares, S. L., & Finn, S. E. (2000). Rorschach correlates of sexual abuse: Trauma content and aggression indexes. *Journal of Personality Assessment, 75,* 212–224.

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.

Konold, T. R., Maller, S. J., & Glutting, J. J. (1998). Measurement and non-measurement influences of test-session behavior on individually administered measures of intelligence. *Journal of School Psychology, 36,* 417–432.

Law, K. S. (1995). The use of Fisher's Z in Schmidt–Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics, 20,* 287–306.

Lebovits, B. Z., & Ostfeld, A. M. (1970). Personality, defensiveness and educational achievement: II. The Cattell 16 PF Questionnaire. *Journal of Clinical Psychology, 26,* 183–188.

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

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (Vol. 1, pp. 3–39). Minneapolis: University of Minnesota Press.

Martinussen, M., & Bjørnstad, J. F. (1999). Meta-analysis calculations based on independent and nonindependent cases. *Educational and Psychological Measurement, 59,* 928–950.

McClelland, D. C. (1980). Motive dispositions: The merits of operant and respondent measures. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1, pp. 10–41). Beverly Hills, CA: Sage.

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

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52,* 509–516.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock–Johnson III.* Itasca, IL: Riverside.

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

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

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

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

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

Meyer, G. J. (in press-b). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment.*

Meyer, G. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Hilsenroth, M. J., Piers, C. C., & Resnick, J. (in press). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment.*

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

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

Meyer, G. J., & Handler, L. (2000). Correction to Meyer and Handler (1997). *Journal of Personality Assessment, 74,* 504–506.

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

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.

Millon, T. (1994). *MCMI-III manual.* Minneapolis: National Computer Systems.

Mitrushina, M. N., Boone, K. B., & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment.* New York: Oxford University Press.

Morey, L. C. (1982). An adjustment for protocol length in Rorschach scoring. *Journal of Personality Assessment, 46,* 338–340.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Netter, B. E. C., & Viglione, D. J. (1994). An empirical study of malingering schizophrenia on the Rorschach. *Journal of Personality Assessment, 62,* 45–57.

Newmark, C. S., Gentry, L., Simpson, M., & Jones, T. (1978). MMPI criteria for diagnosing schizophrenia. *Journal of Personality Assessment, 42,* 366–373.

Ozer, D. J. (1999). Four principles for personality assessment. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 671–686). New York: Guilford Press.

Pallis, D. J., & Birtchnell, J. (1976). Personality and suicidal history in psychiatric patients. *Journal of Clinical Psychology, 32,* 246–253.

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

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

Persons, J. B. (1986). The advantage of studying psychological phenomena rather than psychiatric diagnoses. *American Psychologist, 41,* 1252–1260.

Pervin, L. A., & John, O. P. (1999). *Handbook of personality: Theory and research* (2nd ed.). New York: Guilford Press.

Pires, A. A. (2000). National norms for the Rorschach normative study in Portugal. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 367–392). Mahwah, NJ: Erlbaum.

Rappaport, D., Gill, M. M., & Schafer, R. (1968). *Diagnostic psychological testing* (Rev. ed.). New York: International Universities Press.

Renteria, L., & Meyer, G. J. (2001, August). *The Rorschach DEPI as a measure of cognitive depression.* Poster presented at the annual meeting of the American Psychological Association, San Francisco.

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

Ritzler, B., & Nalesnik, D. (1990). The effect of inquiry on the Exner Comprehensive System. *Journal of Personality Assessment, 55,* 647–656.

Romney, D. M. (1990). Thought disorder in the relatives of schizophrenics: A meta-analytic review of selected published studies. *Journal of Nervous and Mental Disease, 178,* 481–486.

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

Rosenthal, R., Hiller, J. B., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (2001). Meta-analytic methods, the Rorschach, and the MMPI. *Psychological Assessment, 13,* 449–451.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331–340.

Schneider, S. J. (1979). Disability payments for psychiatric patients: Is patient assessment affected? *Journal of Clinical Psychology, 35,* 259–264.

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

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

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

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

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

Tellegen, A. (1991). Personality traits: Issues of definition, evidence, and assessment. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (Vol. 2, pp. 10–35). Minneapolis: University of Minnesota Press.

Tracz, S. M., Elmore, P. B., & Pohlmann, J. T. (1992). Correlational meta-analysis: Independent and nonindependent cases. *Educational and Psychological Measurement, 52,* 879–888.

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

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13,* 452–471.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* Thousand Oaks, CA: Sage.

Wechsler, D. (1997). *WAIS–III/WMS–III technical manual.* San Antonio, TX: The Psychological Corporation.

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

Weiner, I. B. (2001). Advancing the science of psychological assessment: The Rorschach Inkblot Method as exemplar. *Psychological Assessment, 13,* 423–432.

Westen, D. (1995). A clinical–empirical model of personality: Life after the Mischelian ice age and the NEO-lithic era. *Journal of Personality, 63,* 497–524.

Widiger, T. A., & Schilling, K. M. (1980). Toward a construct validation of the Rorschach. *Journal of Personality Assessment, 44,* 450–459.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Wilson, E. O. (1998). *Consilience: The unity of knowledge.* New York: Knopf.

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

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

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

Zalewski, C. E. (1989). A meta-analytic examination of group MMPI profiles: Toward the classification and nosology of psychopathology (Doctoral dissertation, University of Virginia). *Dissertation Abstracts International, 50,* 4792B.

Received May 30, 2001
Revision received July 26, 2001
Accepted August 1, 2001 ■