A Comparative Meta-Analysis of Rorschach and MMPI Validity

Jordan B. Hiller
Harvard University

Robert Rosenthal
University of California, Riverside

Robert F. Bornstein
Gettysburg College

David T. R. Berry
University of Kentucky

Sherrie Brunell-Neuleib
University of Oregon

Two previous meta-analyses concluded that average validity coefficients for the Rorschach and the MMPI have similar magnitudes (L. Atkinson, 1986; K. C. H. Parker, R. K. Hanson, & J. Hunsley, 1988), but methodological problems in both meta-analyses may have impeded acceptance of these results (H. N. Garb, C. M. Florio, & W. M. Grove, 1998). We conducted a new meta-analysis comparing criterion-related validity evidence for the Rorschach and the MMPI. The unweighted mean validity coefficients ($\bar{r}$s) were .30 for MMPI and .29 for Rorschach, and they were not reliably different ($p = .76$ under fixed-effects model, $p = .89$ under random-effects model). The MMPI had larger validity coefficients than the Rorschach for studies using psychiatric diagnoses and self-report measures as criterion variables, whereas the Rorschach had larger validity coefficients than the MMPI for studies using objective criterion variables.

The Rorschach Inkblot Method and the Minnesota Multiphasic Personality Inventory (MMPI) are the two most widely used instruments for the assessment of personality and psychopathology (Lubin, Larsen, & Matarazzo, 1984; Piotrowski & Keller, 1992). They come from divergent traditions within personality research—the MMPI having been developed using empirical criterion keying and the Rorschach owing its birth to the clinical method. The Rorschach in particular has long been the subject of controversy, with early critics noting its poor interrater reliability, the dearth of adequate validation studies, and the absence of population norms (Eysenck, 1959; Jensen, 1965; Zubin, Eron, & Schumer, 1965). Since these critiques were written, two developments have given the Rorschach a more favorable outlook. First came the development of Exner's Comprehensive System for the Rorschach (Exner, 1974, 1978), which systematized scoring procedures for the Rorschach and reported findings from a large archive of Rorschach protocols of normal adults as well as from various psychiatric

groups. Second, two meta-analyses were published, both comparing criterion-related validity evidence for the Rorschach to that of its chief rival, the MMPI (Atkinson, 1986; Parker, Hanson, & Hunsley, 1988), and both concluding that validity evidence is roughly equivalent for both instruments.

In spite of these developments, the debate about the reliability and validity of the Rorschach has continued (Dawes, 1994; Exner, 1996; Meyer, 1997a, 1997b; Weiner, 1996; Wood, Nezworski, & Stejskal, 1996a, 1996b, 1997). In these recent exchanges, the meta-analytic data have been mostly ignored by Rorschach detractors, while being cited as proof positive of Rorschach validity by its supporters. One reason that the two meta-analyses may not have been more widely accepted is that they both suffer from limitations due to methodological problems. Garb, Florio, and Grove (1998) have recently illuminated some of the problems in the meta-analysis by Parker et al. (1998), and they offered a reanalysis of the Parker data showing greater validity for MMPI than for Rorschach. Unfortunately, the reanalysis by Garb et al. (1998) is itself limited by methodological problems. Because the issue of Rorschach validity is so contentious, the methodological features of the two original meta-analyses and the reanalysis deserve careful scrutiny. We consider each of these studies in detail and describe how the problems encountered by previous investigators have influenced our own design for a new meta-analysis of Rorschach and MMPI validity.

Atkinson's (1986) meta-analysis used all conceptual Rorschach and MMPI studies (i.e., those guided by a priori hypotheses) listed in *Psychological Abstracts* for the years 1960, 1965, 1970, 1975, and 1980. Two strategies were used to evaluate validity evidence from these studies. When enough information was provided in study reports, effect sizes were computed (276 Rorschach effect

sizes and 237 MMPI effect sizes). When effect sizes were not calculable, Atkinson calculated a ratio for each study of the number of statistical tests that were significant at $p < .05$ to the total number of statistical tests performed (39 Rorschach studies and 29 MMPI studies). These ratios, then, reflected the proportion of significant findings out of all tests computed in each study. Limitations of analyses using such "box-score" approaches are well known (Bushman, 1994), so analyses using the ratios will not be discussed further. An analysis comparing effect sizes for Rorschach and MMPI studies found that MMPI effect sizes were slightly larger, but Atkinson dismissed the magnitude of this difference as trivial.

Problems with general meta-analytic technique were evident in the Atkinson (1986) meta-analysis. The effect sizes used (r^2, ω^2, Cramér's V) were not satisfactory, because they only assume positive values, and they cannot indicate whether the direction of a validity result is consistent with or opposite to the predicted association.[1] Thus, two studies in this meta-analysis with exactly contradictory results would nevertheless yield identical effect sizes. Measures such as ω^2 and Cramér's V are also not appropriate for meta-analysis because they can be computed from unfocused significance tests (F with more than one df in the numerator and χ^2 with more than one df, respectively). It is not clear whether such unfocused effect sizes entered the meta-analysis, but the use of Cramér's V rather than the simple ϕ coefficient suggests that this is the case. Also, Atkinson sometimes extracted several effect sizes from individual studies and treated them as if they were independent. This procedure, although useful for some purposes, violates the assumption of independence among effect sizes, a violation that can lead to serious errors in the computation of significance levels (Rosenthal, 1991).

The meta-analysis by Parker et al. (1988) concerned reliability evidence as well as validity evidence, and it examined the Wechsler Adult Intelligence Scale as well as the Rorschach and MMPI. We restrict our discussion to Parker et al.'s treatment of the convergent validity data for the Rorschach and the MMPI. The sample of studies used in the Parker et al. meta-analysis was subject to a number of restrictions. Validity studies on the Rorschach and MMPI were selected if they appeared in the *Journal of Personality Assessment* and the *Journal of Clinical Psychology* between 1970 and 1981. The potential for selection bias is great when only well-known journals are used, as these journals may contain studies with effect sizes larger than those found in other less prestigious journals. In particular, the possibility of editorial bias must be noted for the *Journal of Personality Assessment*, given its historical association with the Rorschach. The sample was also restricted to studies examining 9 Exner Comprehensive System variables for the Rorschach or the 13 basic clinical and validity scales from the MMPI. Additionally, studies using t, F, or χ^2 were excluded from the main analyses, leaving only those studies in which a correlation coefficient was directly reported. These restrictions reduced the overall number of convergent validity studies that were meta-analytically compared to 30 MMPI studies and only 5 Rorschach studies. No significant difference was detected between Rorschach ($r = .41$) and MMPI ($r = .46$) validity effect sizes. It is remarkable that the main findings from this meta-analysis, often cited in support of Rorschach validity, are based on only 5 Rorschach studies.

A subsidiary table in Parker et al. (1988) was presented for studies analyzed with t or F, using ω^2 as an effect size measure (studies using χ^2 were excluded from consideration because they were rare). These studies were set aside from the correlational studies in the main analysis, because the authors correctly observed that ω^2 and r are not directly comparable.[2] We believe that not only is ω^2 not comparable to r, ω^2 is altogether unsuitable for meta-analytic work. As we noted earlier, ω^2 does not take negative values, therefore it cannot represent evidence that argues against the meta-analytic hypothesis; we also noted that ω^2 can be computed from unfocused analyses, that is, F tests with more than one df in the numerator, which cannot correctly contribute information to a meta-analysis. Many of the effect sizes in this subsidiary table did indeed come from such unfocused results (K. C. H. Parker, personal communication, November 24, 1998). And although Parker et al. stated that they assigned appropriate signs to effect sizes according to whether the results were consistent with or contrary to test validity, this strategy could not have been properly applied to effect sizes computed from unfocused results.[3] A further problem with ω^2 is that it consistently underestimates the magnitude of results, especially when sample sizes are low. This is due to a correction factor in the formula for ω^2, which is meant to adjust the statistic for chance levels of association. The effect of this correction factor is that even when r is .50 or higher, the value of ω^2 can still be zero, when sample sizes are modest; ω^2 prevents results from properly contributing information to a meta-analysis. This effectively defeats one of the primary purposes of meta-analytic work, which is to aggregate effects accurately across studies. Thus, even if ω^2 had been used only for focused studies with appropriate signs, it would still be inappropriate for meta-analytic use.

Garb et al. (1998) described other problems in the Parker et al. (1988) meta-analysis. Garb and colleagues concluded that some of the effect sizes used by Parker et al. tended toward zero when the validity of the test is supported (e.g., a near-zero effect size reflecting the fact that the number of responses on a Rorschach

---

[1] Cramér's V was mistakenly identified as θ in the article by Atkinson (1986). It is clear that Cramér's V was used because Atkinson specifically cited a formula for its calculation in Hays (1973). Contrary to Atkinson's assertion, and unlike the other two effect sizes, Cramér's V is not technically a measure of percentage of explainable variance. However, it is subject to the same limitations as r^2 and ω^2.

[2] It would have been preferable to use r rather than ω^2 as an effect size measure for results based on t or F. Simple techniques to compute r from t (or from F with 1 df in the numerator) were available at the time the meta-analysis was written (Cohen, 1965; Rosenthal & Rosnow, 1984), although they might not have been known by Parker et al. (1988).

[3] An example will illustrate the problem. If three groups of subjects are compared on a certain MMPI scale, then ω^2 could be computed from the F test with 2 df in the numerator. How should the sign be determined for the effect size? Assuming that ties are impossible, there are six possible configurations of high, middle, and low mean scores for three groups. Allocation of two signs (+ or −) to the six patterns is not feasible or sensible. If one of the six patterns is truly indicative of test validity, then a more appropriate procedure is to compute the effect size from a contrast analysis with 1 df, and the sign of the t (and the effect size) would reflect the degree to which the outcome agreed or disagreed with the pattern specified by the contrast weights (Rosenthal & Rosnow, 1985).

protocol is unaffected by verbal reinforcement).[4] Also, Garb et al. noted that in the Parker et al. meta-analysis, sometimes Rorschach or MMPI scales were validated against criterion measures that came from the same instruments, raising the possibility that shared method variance resulted in inflated effect sizes. A final criticism leveled at Parker et al. by Garb et al. was that the subsidiary ω^2 effect sizes computed from t and F tests were not pooled with the r effect sizes and used in the main analyses. These problems led Garb et al. to reanalyze the data set used by Parker et al., excluding the problematic effect sizes and pooling the ω^2 effect sizes with the others. They found a significant advantage for MMPI convergent validity studies over Rorschach. However, these findings are questionable for two reasons. First, many of the ω^2 effect sizes were based on unfocused significance tests, violating meta-analytic principles. Second, even the ω^2 effect sizes based on t or F with a single degree of freedom in the numerator were not properly converted to units of r before they were combined with the other effect sizes and are, therefore, too small.[5] The problematic treatment of the ω^2 effect sizes renders the pooled effect size estimates presented by Garb et al. very difficult to interpret.

A conceptual issue that affects both of the original meta-analyses as well as the Garb et al. (1998) reanalysis concerns a basic distinction in the literature on these tests, between exploratory and confirmatory studies.[6] Studies testing a specific prediction made about one or two scales, either based on sound theory or on prior empirical research, can be rightfully called confirmatory. Confirmatory studies generate effect sizes that are usually suitable for inclusion in a meta-analysis of validity evidence. In the exploratory scenario, several groups are compared across a host of MMPI variables or Rorschach variables without any a priori predictions, to see whether any interesting differences turn up. Exploratory studies pose a particularly vexing problem for meta-analysts attempting to examine the overall validity of the Rorschach or the MMPI, because it is difficult to decide which effect sizes from exploratory studies should be considered validity evidence, and should therefore enter the meta-analysis. The previous meta-analysts who have examined Rorschach and MMPI validity have dealt with exploratory studies by segregating them from the confirmatory studies, and either reporting their results separately (Parker et al., 1988) or discarding them altogether (Atkinson, 1986).[7] This strategy of segregating exploratory studies is problematic, because it depends on the authors of individual studies for determinations of what constitutes relevant validity data. Validity evidence is validity evidence, regardless of whether an author made an a priori prediction or not. This strategy runs the risk of excluding relevant validity evidence, simply because an author failed to make a reasonable prediction; conversely, it runs the risk of including irrelevant or misleading evidence when study authors falsely claim to have made a priori predictions concerning post hoc discoveries.

For all these reasons, lingering questions remain concerning the meta-analytic data bearing on Rorschach and MMPI validity. The present meta-analysis of criterion-related validity evidence for the Rorschach and the MMPI was undertaken in an effort to address some of the problems of the other meta-analyses. We used a random sample of studies from the MMPI and Rorschach literature published between 1977 and 1997, and we asked expert judges to select appropriate validity evidence from Rorschach and MMPI investigations, enabling us to include data from both exploratory

and confirmatory studies. Furthermore, we conducted several moderator analyses to shed light on the circumstances under which Rorschach and MMPI variables might prove to have greater or lesser validity.

Method

Literature Search

PsycLIT searches were used to identify potentially relevant studies published between January 1977 and December 1997. The start of this period was chosen in accordance with the focus of this Special Section on the research literature concerning the Rorschach and MMPI published since 1977; the end of the period reflected the most recent information available in PsycLIT at the time the literature search was conducted. MMPI articles were identified by searching for the terms "MMPI or (Minnesota and Multiphasic)" in article titles and abstracts; Rorschach articles were identified with the single search term "Rorschach." These searches yielded 4,378 MMPI and 1,793 Rorschach articles.[8]

In addition to the published literature on the MMPI and Rorschach, we attempted to obtain unpublished studies in this area. Using letters, e-mail, phone, and fax, we attempted to contact 115 researchers who had presented research on the MMPI or the Rorschach at the Society for Personality Assessment between 1993 and 1997, asking them to send us unpublished studies conducted by them or by their colleagues.[9] Additionally, an appeal for unpublished studies was made on the SSCPnet, an e-mail discussion group sponsored by the Society for a Science of Clinical Psychology, which is monitored by many clinical psychology researchers. There was only one response to the message posted to the SSCPnet, from Gregory Meyer, the editor of the present Special Section. Altogether these efforts yielded two unpublished MMPI studies and eight unpublished Rorschach studies.

Studies were selected for inclusion in the sample in a two-step procedure. First, Jordan B. Hiller screened the studies to determine whether they

---

[4] It should be noted that this point has been disputed by Parker, Hunsley, and Hanson (in press). But see also the reply by Garb, Florio, and Grove (in press).

[5] Garb et al. (1998) took the square root of each ω^2 effect size, to place it on a scale comparable to r, before combining the effect sizes. The new effect sizes yielded by this procedure are still smaller than they should be, when they are calculated properly as r.

[6] Parker et al. (1988) call these studies "unknown validity" and "convergent validity" studies, while Atkinson (1986) refers to them as "undirected" versus "conceptual" studies. We adopt Garb et al.'s (1998) more transparent terminology of "exploratory" and "confirmatory."

[7] Parker et al. (1988) summarized effect size estimates for exploratory studies, but they did not explain how direction of effect was assigned to effect sizes in the absence of a priori hypotheses. Results from their category of "unknown validity" studies are therefore questionable, because it is unclear whether any evidence against test validity (i.e., having a negative sign for the effect size) could have been properly represented.

[8] Many of the articles identified in the PsycLIT searches were case reports, literature reviews, or theoretical papers, and they were not suitable for inclusion.

[9] Researchers were identified from printed programs of the annual meetings of the Society for Personality Assessment between 1993 and 1997, and contact information was gleaned from a recent membership directory. We attempted to contact most first authors and many second authors of presentations made in this time period, although contact information was often unavailable. We cannot say with certainty how many of the authors received our solicitation message.

met several basic eligibility criteria. Studies were eligible for inclusion if they were written in English and contained data relating one or more Rorschach indices or MMPI scales to at least one external criterion variable. Studies that used short forms of the MMPI were excluded, as were studies for which Rorschach or MMPI results were part of an assessment battery and were not reported separately. Among the studies meeting these criteria, a further distinction was made. Some studies were deemed to contain clear, unambiguous validity coefficients (e.g., a correlation between the Beck Depression Inventory and the MMPI Depression scale), and these were immediately included in the sample. Other studies contained correlations involving Rorschach or MMPI variables that were not as clearly indicative of test validity. These studies were submitted to judges who were experts on the two tests, for further consideration.

Two Rorschach experts independently coded the Rorschach studies, and the MMPI studies were likewise coded by 2 separate MMPI judges.[10] The judges were furnished only with synopses of the methodology of each study, so that their decisions would not be contaminated by the authors' predictions or by the results. The judges were asked to indicate whether each effect size for the relationship between a Rorschach or MMPI variable and a criterion variable constituted validity evidence—that is, whether an effect could reasonably be expected to be "significant," given the nature of the test, the sample, and the criterion variable.

Reliability was computed for the first set of five studies considered by both pairs of judges. These studies contained 60 and 281 individual effect sizes for Rorschach and MMPI, respectively. Reliability calculations were made by considering each effect size as an individual observation, disregarding the nesting of effect sizes within studies. Interrater reliability was .35 for the Rorschach judges and .39 for the MMPI judges, as indexed by the ϕ coefficient. Effective reliability for each pair of judges was .51 for the Rorschach and .57 for the MMPI, as calculated by the Spearman–Brown formula.

Only effect sizes that both judges agreed were validity coefficients were extracted from the studies and used in the meta-analysis. Thus, studies evaluated by the judges were included in the meta-analysis as long as they contained at least one effect size that was deemed appropriate for inclusion by both judges. Studies for which judges did not agree about any effect sizes were excluded, as were studies that both judges agreed did not contain any appropriate effect sizes.

Further information about the number of studies considered at each step of the selection procedure is contained in Table 1. We had decided to obtain a sample of 30 published Rorschach studies and 30 published MMPI

Table 1
Summary of Literature Search and Review of Studies for Inclusion

Categories	Published		Unpublished	
	MMPI	Rorschach	MMPI	Rorschach
Screening				
Excluded	71	96	0	4
Included	13	10	1	1
Submitted to judges	32	29	1	3
Total	116	135	2	8
Judges				
Agreed to include	17	20	0	3
Agreed to exclude	8	3	0	0
Disagreed (excluded)	7	6	1	0
Total	32	29	1	3
Included in meta-analysis				
From screening	13	10	1	1
From judges	17	20	0	3
Total	30	30	1	4

studies for the meta-analysis, and therefore studies were sampled randomly from the pools of published studies identified by PsycLIT until this sample size was achieved. In an effort to obtain the sample of 60 published articles used here, it was necessary to evaluate 135 randomly selected Rorschach articles and 116 randomly selected articles on the MMPI. We had initially hoped to include 10 each of unpublished Rorschach and MMPI studies, but we were unable to do so. Of the two unpublished MMPI studies and eight unpublished Rorschach studies that we obtained, only one MMPI study and four Rorschach studies were ultimately included in the sample. Because of their importance for estimating the potential magnitude of publication bias, even these few studies were retained for use in the meta-analysis.

Coding Procedure

For both Rorschach and MMPI studies, several variables reflecting study characteristics were coded as follows: (a) the year of publication; (b) a dichotomous variable reflecting whether the study appeared in one of five core journals that regularly publish research concerning these instruments (*Assessment, Journal of Clinical Psychology, Journal of Consulting and Clinical Psychology, Journal of Personal Assessment,* and *Psychological Assessment*) or in a different outlet; (c) a dichotomous variable indicating whether the study was included in the meta-analysis during the initial screening or later, after consideration by the judges; and (d) a code for whether the analytic method used in the study was t or F, Pearson's r, or some different method. Furthermore, a categorical variable with six levels was used to reflect the nature of the criterion variables used to validate MMPI or Rorschach measures (groups based on psychiatric diagnoses, objective outcomes such as suicide or hospitalization, ratings made by observers or judges, self-report questionnaires or scales, "projective" measures, or a combination of the preceding criterion types).[11]

Rorschach studies were categorized by the different types of Rorschach predictors used in the studies (dichotomous signs such as the presence or absence of space responses, sums or ratios reflecting the absolute number or proportion of certain types of responses in a protocol, scales or composites of multiple elements, or a combination of several types of predictors). A dichotomous variable was used to code whether the Exner Comprehensive System or some other Rorschach coding system was used. Also, we noted whether the Rorschach predictors used in each study were based on structural features of responses (such as the shape or color of the inkblot area identified in the response), content of responses (i.e., characteristics of the percept itself, such as whether it is an animal, a type of food, a household item, etc.), or both. MMPI studies were categorized with a different scheme for predictor type (basic validity and clinical scales, supplemental or research scales, 2- or 3-point codetypes, or a combination of MMPI predictor types), and a dichotomous variable reflected whether the MMPI or MMPI-2 was used.

Meta-Analytic Techniques

Information extracted for meta-analytic calculations included validity coefficients (i.e., effect sizes), their significance levels, and the number of sampling units used in each study. When results were reported as "significant at $p < .05$" or "not significant," conservative estimates of significance and effect size were obtained by assuming one-tailed significance

---

[10] Robert Bornstein and Mark Hilsenroth were the judges for Rorschach studies; David Berry and Radhika Krishnamurthy evaluated the MMPI studies.

[11] In this article, we use the term "projective" in the historical sense, in reference to tests using ambiguous stimuli and eliciting open-ended responses. They are contrasted with self-report measures (historically known as "objective" instruments), which usually have a true–false or multiple-choice format.

levels of .05 or .50, respectively. When unfocused F or χ^2 statistics were reported in the original studies, contrast analyses were conducted in order to extract meaningful effect sizes from focused comparisons. When more than one effect size was available within a study, the data were combined into a single estimate according to the methods described by Rosenthal and Rubin (1986; for tests with multiple dependent measures) and Hayes (1998; for validity data involving several pairs of independent and dependent measures). These procedures typically provide more powerful, less conservative estimates than other common procedures, such as taking the median or mean effect size from each study.[12]

Effect size calculation was usually straightforward, but sometimes alternative procedures were necessary. On some occasions, test scores for a single group were compared to appropriate published norms, such as the MMPI-2 standardization sample (Kornfeld, 1995) or Exner's normative data (Cruz, Brier, & Reznikoff, 1997; Exner, Colligan, Boll, Stischer, & Hillman, 1996; Kaser-Boyd, 1993; Zimmerman & Dillard, 1994). In these instances, the large normative samples were considered to be populations with known means and standard deviations, and the effect sizes were derived from Z values. For one MMPI study (Fals-Stewart & Schafer, 1993) and one Rorschach study (Sheehan & Tanaka, 1983) where multivariate analyses were conducted without presenting sufficient information to calculate univariate effect sizes (i.e., zero-order correlations), partial r effect sizes were computed.[13] For one study in which hundreds of effect sizes were computed but only a small subset of the very largest ones were reported (Ben-Porath, Hostetler, Butcher, & Graham, 1989), a Bonferroni-type procedure was used to adjust the largest effect sizes for the number of effect sizes that were computed (Rosenthal, 1991, pp. 30–31).

The meta-analytic procedures used here are those described by Rosenthal (1991). The effect sizes used in calculations were Fisher-transformed Z_rs, but results were transformed back to units of r for discussion, to facilitate interpretation. Effect size calculation and other meta-analytic calculations were performed using Microsoft Excel spreadsheets developed specifically for this purpose.[14]

Two statistical models are commonly distinguished in meta-analytic work: fixed-effects models and random-effects models. The more common fixed-effects model effectively uses participants from the constituent studies in a meta-analysis as sampling units, whereas random-effects analyses use entire studies as sampling units. The chief interpretive difference between these two meta-analytic models concerns the population to which results may be generalized. For fixed-effects models, results are technically generalizable only to the populations examined in the particular studies entering the meta-analysis. Random-effects models allow generalization to populations of relevant studies (existing or hypothetical) that were not included in the meta-analysis. The price of the greater generalizability of the random-effects model is reduced statistical power.[15] There are good reasons for using each kind of analysis, but the random-effects approach may be particularly relevant for the current meta-analysis, given that the studies used here are truly a sample from a larger population of studies to which we would like to generalize. However, random-effects analyses are underpowered for some of the comparisons between smaller subsets of the sample. In this investigation, we generally use random-effects analyses for comparisons among sets of studies, and for the major analyses, fixed-effects calculations are presented as well. We note that although some sophisticated random-effects techniques have recently been developed (Raudenbush, 1994; Shadish & Haddock, 1994), here we use the simple and intuitive approach of applying standard procedures for statistical analysis to the effect sizes obtained from the studies.

Results

Information about characteristics of individual studies is presented in Table 2; Table 3 shows a back-to-back stem and leaf plot of effect sizes for MMPI and Rorschach studies. Both distributions appear approximately normal.

Comparisons Between Published and Unpublished Studies

Because the studies examined here are from two very different sources—a randomly selected sample of published studies and a small convenience sample of unpublished studies—the first analyses were conducted to examine the differences between published and unpublished studies. Only a small number of unpublished studies were available for these analyses, so fixed-effects comparisons were conducted to maximize their power. Effect sizes from unpublished Rorschach studies ($\bar{r} = .29$) were practically identical in magnitude to those obtained from published studies ($\bar{r} = .29$, $Z = .13$, $p = .90$, two-tailed).[16] However, the single unpublished MMPI study had an effect size ($r = .74$) that was greater than those from published MMPI studies ($\bar{r} = .30$, $Z = 5.88$, $p = 4 \times 10^{-9}$, two-tailed). Given that unpublished studies are generally expected to have lower effect sizes than published ones, the unpublished MMPI study we were able to obtain was probably not representative of the population of unpublished studies. We therefore omitted both Rorschach and MMPI unpublished studies from the remaining analyses, recognizing that the results reported here are generalizable only to the population of published studies.

Effect Sizes and Overall Significance

Table 4 describes the distributions of effect sizes for the MMPI and for the Rorschach. Median effect sizes were .22 for the MMPI and .29 for the Rorschach. The unweighted mean effect sizes for the two tests were quite similar, .30 for the MMPI and .29 for the Rorschach. When weighted by the degrees of freedom from each study, the mean effect sizes were somewhat less similar, .37 for the MMPI and .26 for the Rorschach. We inspected the effect sizes to account for the differences among these three measures of central tendency. The median, unweighted mean, and weighted mean were all quite similar for the Rorschach, but the discrepancy among

---

[12] Both the Hayes procedure and the Rosenthal and Rubin procedure require estimates of the typical intercorrelation among dependent variables (and among independent variables as well, for the Hayes procedure). On rare occasions, these intercorrelations were obtainable from the study manuscript. When they were not available, estimates were generated using test reference materials if possible (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), or conservative estimates were generated on the basis of prior research. Note that for this application, "conservative" estimates are *larger* than the true intercorrelations, because larger intercorrelations lead to smaller composite effect sizes for both procedures (see p. 403 of Rosenthal & Rubin, 1986).

[13] Because "unpartialled" effect sizes are generally preferable for meta-analytic purposes, we later reconducted several of the key analyses, excluding both of the partial r effect sizes. Both of the partial r effect sizes were quite close to the mean effect sizes for their respective instruments, therefore the results were virtually unchanged when excluding these studies. The unweighted mean effect sizes for each instrument were identical to two decimal places, and they were still not detectably different from each other under fixed effects ($Z = .25$, $p = .80$, two-tailed) or random effects models ($t(56) = 0.12$, $p = .91$, two-tailed, $r = .02$).

[14] Available on request from Jordan B. Hiller.

[15] For a fuller discussion of the differences between fixed-effects and random-effects models, see Hedges (1994b; see also Hedges, 1994a, and Raudenbush, 1994).

[16] All reported p values are one-tailed unless otherwise specified.

Table 2
Study Characteristics

Study	Predictor variables	Criterion variables	Moderator coding[a]	N	r	Z
		MMPI studies				
Aaronson, Dent, & Kline (1996)	MacAndrew Alcoholism Scale—Revised, Addiction Potential Scale, & Negative Treatment Indicators	Discharge status (regular vs. irregular) for veterans in a domiciliary	C/i/J/r/M/o/2	335	.08	1.52
Bagby, Buis, & Nicholson (1995)	L, Obvious–Subtle index, Positive Malingering	Undergraduates with "fake good" instructions vs. standard undergraduate controls	C/i/S/t/B/m/2	470	.69	16.91
	F, Obvious–Subtle index, Dissimulation–Revised	Undergraduates with "fake bad" instructions vs. nonforensic psychiatric inpatients				
Ben-Porath, Hostetler, Butcher, & Graham (1989)	Content-homogeneous subscales for *Si*	Behavioral ratings made by spouses	C/n/S/r/S/r/2	1,644	.35	14.07
Bloom, Shelton, & Michaels (1978)†	*Hs, D,* & *Hy*	Dysmenorrheac group vs. normal control group	C/n/J/t/B/o/1	48	.16	1.05
Bowman, Bennett, & Welsh (1981)	*Si*	Introversion–Extroversion from the Strong–Campbell Interest Inventory, & Affiliation from the Personality Research Form	N/n/S/r/B/s/1	143	.47	5.79
Brown & Gutsch (1985)	*Pd*	Behavioral test of ability to delay gratification	N/n/J/m/B/o/1	53	.12	0.88
Chaney, Roszell, & Blaes (1982)†	*D* & *Pt*	Drug counselors' ratings of client treatment readiness	N/o/J/m/B/r/1	78	0	0
Cooper & Holmstrom (1984)	Alexithymia scale, Repression–Sensitization scale	Cornell Medical Index	N/n/J/r/S/s/1	123	−.26	−2.90
Fals-Stewart & Schafer (1993)	*Pd, Pa, Sc*[b]	Treatment attendance for outpatient behavioral therapy for obsessive–compulsive disorder	N/o/J/m/B/o/1	169	.32	4.04
Faull & Meyer (1993)	*D* & MMPI-2 Depression Content scale	Beck Depression Inventory	U/o/S/r/M/s/2	87	.74	8.27
Fontana & Rees (1982)	*Hs, D,* & *Hy*	Dysmenorrheac group vs. normal control group	N/n/J/t/B/o/1	48	.15	0.99
Friedman, Gleser, Smeltzer, Wakefield, & Schwartz (1983)	Psychotic Overlap Scale, Neurotic Overlap Scale, & Maladjustment Overlap Scale	Psychiatric inpatients & psychiatric outpatients vs. medical controls & undergraduate controls	C/i/S/t/S/d/1	300	.63	12.01
Gas (1987)	MacAndrew Alcoholism Scale	Alcoholic inpatients vs. normal controls	N/i/S/t/S/d/1	60	.59	4.99
Gayton, Burchstead, & Matthews (1986)	Posttraumatic Stress Disorder (PTSD) subscale (*PK*)	Veterans with PTSD vs. veterans without PTSD	C/o/S/t/S/d/1	59	.26	1.97
Goshtautas & Rugevitchius (1983)†	*Hs, D,* & *Hy*	Patients with neurocirculatory asthenia and angina pectoris vs. normal controls	N/o/J/t/B/o/1	150	.32	4.01
Gottschalk, Stein, & Shapiro (1997)†	*D, Pd, Pa, Pt,* & *Sc*	Content analysis of speech samples for anxiety, hostility directed outward, social alienation, and depression	C/o/J/r/B/r/2	25	.12	0.57
Hirschfeld et al. (1983)	Ego resiliency	Inpatients and outpatients whose depressive symptoms remitted vs. patients with unremitting symptoms	N/i/J/t/S/r/1	57	.22	1.64
Kornfeld (1995)	*L* & *K*	Police officer applicants vs. population norms	C/n/J/m/B/o/2	84	.13	1.18
Labott, Preisman, Torosian, Popovich, & Ianuzzi (1996)†	*Hs* & *Hy*	Somatizing patients vs. patients with medical diagnoses	N/o/J/t/B/o/2	29	0	0
Laudeman (1977)†	*Pd* & *Ma*	Young adults arrested for alcohol-related offenses vs. normal controls	N/n/J/t/B/o/1	27	.20	1.01
Morrison, Edwards, & Weissman (1994)	Classification into psychiatric categories based on MMPI-2 codetypes	Clinical diagnoses	C/o/S/r/C/d/2	154	.21	2.68
Morton-Page & Wheeler (1997)	*D*	Beck Depression Inventory	N/n/S/r/B/s/1	200	.76	13.05
Perconte & Goreczny (1990)	*F, K, Hs, D, Hy, Pd, Pa, Pt, Sc,* & *Ma*	Veterans with PTSD vs. veterans with fabricated symptoms of PTSD	N/o/J/t/B/d/1	39	−.02	−0.10

(table continues)

Table 2 (continued)

Study	Predictor variables	Criterion variables	Moderator coding[a]	N	r	Z
		MMPI studies (continued)				
Perconte, Griger, & Bellucci (1989)	F, D, Hy, Pt, Heterosexual Discomfort, Introspective Critical, Manifest Anxiety, Anxiety 1st Factor, General Maladjustment, Social Maladjustment, Neurosis, Neurotic Overcontrol, Overcontrolled Hosility	Veterans diagnosed with biogenic erectile dysfunction vs. veterans diagnosed with psychogenic erectile dysfunction	N/o/J/t/M/o/1	46	−.11	−0.74
Schmidt & Miller (1983)	D	Outpatient psychotherapy for depression (measured post-treatment) vs. wait-list control group	N/o/S/t/B/o/1	56	.69	5.66
Shealy, Lowe, & Ritzler (1980)†	Hs, D, Hy, & Pt	Sleep onset insomniacs vs. normal control group	C/n/J/t/B/o/1	80	.37	3.37
Skolnick & Zuckerman (1979)†	Pd	Pre-post differences after drug treatment in a therapeutic community	C/i/J/t/B/o/1	96	0	0
Sloan, Arsenault, Hilsenroth, & Harvill (1996)	PTSD scales (PK & PS)	Number of symptoms of acute posttraumatic stress reported by Gulf War veterans	C/n/S/r/S/d/2	57	.44	3.38
Snyder & Regts (1990)	Arnold Sign Indicator of marital distress	Couples presenting for marital counseling vs. normal control couples	N/o/S/t/M/o/1	60	.26	2.43
Turner & Romano (1984)	D	Beck Depression Inventory, Zung Self-Rating Depression Scale, and clinical diagnosis of depression	C/o/S/m/B/m/1	40	.73	5.20
Watson et al. (1987)	Welsh R, Eichman Repression Factor, Haan Repression, Haan Denial, Byrne et al. Repression-Sensitization, Little–Fisher Denial	Projective Repression Instrument	C/i/S/r/M/p/1	190	.20	2.73
		Rorschach Studies				
Aebischer (1994)†	Sexual Anatomy content	Physically abused children vs. sexually abused children	U/o/J/t/R/o/N/c	28	0	0
Berg (1990)	Afr, Lambda, Grandiose content & Splitting content	Outpatients with borderline personality disorder vs. outpatients with narcissistic personality disorder	C/o/J/t/R/d/M/b	76	.15	1.27
Bornstein, Leone, & Galley (1988)	Oral content	Personality ratings made from written self-descriptions	C/n/S/r/R/s/N/c	161	−.03	−0.44
Bornstein, Rossner, Hill, & Stepanian (1994)	Oral content	Interpersonal Dependency Inventory	C/n/J/r/R/p/N/c	125	.51	5.94
Campo, Dow, & Tuset (1988)	Global rating of improvement in form level and determinants	Pre-post differences after 1 to 2 years of psychoanalytic psychotherapy	N/o/J/t/R/o/N/b	30	−.17	−0.90
Cruz, Brier, & Reznikoff (1997)	X+%, X−%, & Xu%	Learning disabled adolescents vs. population norms	N/o/J/m/R/o/E/s	44	.43	2.92
Ellis & Zahn (1985)	P & F+%	Patients with closed head injury vs. normal controls	C/i/J/t/R/o/E/b	71	.65	6.15
Exner, Colligan, Boll, Stischer, & Hillman (1996)	P, F+%, & Sum 6 Sp	Patients with closed head injury vs. population norms	C/i/J/m/R/o/E/b	60	.08	0.64
Fowler, Hilsenroth, & Handler (1995)	Mutuality of Autonomy, Holt Primary Process Aggression, & Secondary Process Aggression	Complexity of Representations and Affect Tone scored from early memories	N/o/J/r/C/p/N/c	29	.43	1.95
Frank, Tuber, Slade, & Garrod (1994)	Holt Primary Process Integration from mother's protocol	Infant attachment style, assessed using the Strange Situation paradigm	N/n/S/t/C/r/N/b	25	.52	2.58
Gordon & Oshman (1981)	Reaction time to chromatic cards	Hyperactive boys vs. normal controls	N/o/J/t/R/d/N/b	40	−.16	−0.96
Harper & Scott (1990)	FQ+, FQo, FQu, FQ−, & X−%	Learning disabled children vs. normal controls	N/n/J/t/R/o/E/s	25	.25	1.18
Karp & Gernert (1995)	Harm Scale from Karp Inkblot Response Questionnaire (KIRQ)[c] Leadership scale from KIRQ	Anxiety sensitivity index, Beck Depression Inventory, state and trait anxiety from the State–Trait Anxiety Inventory Leadership questionnaire	U/n/J/m/R/m/N/c	273	.28	4.29
Kaser-Boyd (1993)	X+% & X−%	Battered women who killed their spouses vs. population norms	C/n/J/m/R/o/E/s	22	.44	2.03
Kiran Kumar & Thimmappa (1982)	Witkin's indices of field dependence	Embedded Figures Test	N/n/S/r/R/s/N/s	50	.10	0.64

Table 2 (continued)

Study	Predictor variables	Criterion variables	Moderator coding[a]	N	r	Z
		Rorschach Studies (continued)				
Leavitt & Labott (1996)	Oppressive content, sexual content, texture responses, color dominated responses	Adult women with childhood sexual abuse vs. normal controls	N/o/S/m/D/o/N/b	114	.53	6.09
Meloy (1984)†	Thought disorder index	Intensity of primary process content in dreams of parents of schizophrenia patients	N/n/J/r/C/p/N/c	14	0	0
Meyer & Resnick (1996)	Ego impairment index, & Conceptual ego strength index	Ego deficit rating based on clinical diagnoses	U/i/S/r/C/d/E/b	232	.44	7.06
Meyer et al. (1993)	X−% & W	Developmental Test of Visual-Motor Intergration, Bender Visual Motor Gestalt Test, Rey–Osterrieth Complex Figure	U/o/J/r/R/o/E/s	30	.43	2.34
Perry et al. (1995)	m	Effects of amphetamine	C/n/J/t/R/o/E/s	22	.47	2.19
Pierloot & Houben (1978)†	Barrier index	Overestimation of body size in anorexia nervosa	N/i/J/r/R/o/N/c	19	0	0
Regmi (1986)	F+%	Level of acculturation in indigenous Nepalese participants	N/n/J/t/R/o/N/s	71	−.05	−0.44
Ridley (1987)	Developmental Level & Developmental Quality	WISC–R Verbal, Performance, and Full Scale IQ	C/o/S/r/R/s/M/s	134	.39	4.62
Sah (1989)	F+%, CF, C, A, Ad, & Reaction time to chromatic cards	Reckless railway drivers vs. safe railway drivers	N/n/J/t/R/o/N/b	65	.33	2.64
Salyer, Holmstrom, & Noshpitz (1991)	X+%	Learning disabled children vs. normal controls	N/o/S/t/M/d/E/b	47	.35	2.39
Schlesinger & Fox (1980)	Number of achromatic responses	Depressed inpatients vs. nondepressed psychiatric patients	N/i/S/t/R/d/N/s	40	.35	2.17
Sheehan & Tanaka (1983)	Rorschach Prognostic Rating scale & Form Level[d]	Improvement after psychotherapy for stuttering	C/o/J/m/M/r/N/s	50	.26	1.72
Simon (1985)	Egocentricity index	Coopersmith Self-Esteem Inventory	C/n/S/r/R/s/E/c	60	−.05	−0.41
Singh (1983)	Hostile content in children's protocols	Strict childrearing attitudes in mothers	N/n/J/t/R/s/N/c	60	.32	2.47
Steiner, Martin, Wallace, & Goldman (1984)	Affective vs. schizoptypal classification based on content	Dexamethasone Suppression Test in a group of patients with comorbid depression and borderline personality disorder	N/i/S/m/R/o/N/c	21	.30	1.32
Tegtmeyer & Gordon (1983)†	S	Social-withdrawal, Aggression, Delinquency, and Overall Behavior Problems scales from Child Behavior Checklist	N/n/J/t/D/r/E/s	38	0	0
Watson & Pantle (1993)	Reflection responses	Scales 5 (Confident/Narcissistic) and 6 (Forceful/Antisocial) of the Millon Adolescent Personality Inventory	N/i/S/t/D/s/E/c	112	.06	0.64
Yanovski, Menduke, & Albertson (1995)	Bizarre content	Visual imagery reactivity in psychotherapy	N/o/J/m/R/r/N/c	80	.29	2.58
Zimmerman & Dillard (1994)	Weighted Sum 6, X−%, DEPI, F+%, & D	Sexually abused children in residential treatment vs. population norms	N/i/J/m/M/o/E/b	8	.91	2.95

Note. Studies in which effect sizes were estimated using approximate significance levels are denoted by a dagger (†) following the year of publication. MMPI = Minnesota Multiphasic Personality Inventory; L = Lie; F = Infrequency; Si = Social Introversion; Hs = Hypochondriasis; D = Depression; Hy = Hysteria; Pd = Psychopathic Deviate; Pt = Psychasthenia; Pa = Paranoia; Sc = Schizophrenia; PK = Keane PTSD; K = Correction; Ma = Hypomania; PS = Post-Traumatic Stress Disorder

[a] Key to moderator coding for MMPI studies: Core journal (C = core; N = not core; U = unpublished)/Sample (i = inpatient, o = outpatient; n = nonpatient)/Inclusion (J = judges; S = screening)/Analytic method (t = t or F; r = r, m = mixed/other)/Predictor type (B = basic clinical/validity scale; S = supplemental; C = codetype; M = multiple types)/Criterion type (o = objective; d = psychiatric diagnosis; r = observer rating; s = self-report; p = "projective"; m = multiple types)/Version (1 = MMPI; 2 = MMPI-2). Key to moderator coding for Rorschach studies: Core journal (C = core; N = not core; U = unpublished)/Sample (i = inpatient; o = outpatient; n = nonpatient)/Inclusion (J = judges; S = screening)/Analytic method (t = t or F; r = r, m = mixed/other)/Predictor type (D = dichotomous sign; R = ratio or sum; C = constellation or scale; M = multiple types)/Criterion type (o = objective; d = psychiatric diagnosis; r = observer rating; s = self-report; p = "projective"; m = multiple types)/Coding system (E = Exner system; N = non-Exner system; M = mixed)/Rorschach variable type (s = structural; c = content; b = both). [b] These predictors were included as part of a multiple regression equation using all 13 basic clinical and validity scales. [c] The Karp Inkblot Response Questionnaire is a pencil and paper questionnaire that is given to Rorschach respondents following standard individual or group administration of the Rorschach. Likert-type scales are used to rate various aspects of response content. Although an argument could be made that Rorschach data based on self-report of test takers should be excluded from the meta-analysis, the point is moot because this study was unpublished and therefore did not contribute to the main analyses. [d] These predictors were included as part of a multiple logistic regression equation that also used M, FM, Shading, and Color responses as predictors. The Rorschach Prognostic Rating Scale is a composite derived from the other six predictors that were used in the equation.

Table 3
Stem and Leaf Plot of Effect Sizes (r) From MMPI and Rorschach Studies

MMPI		Rorschach
	.9	1
	.8	
6*4*3	.7	
993	.6	5
9	.5	123
74	.4	*333447*
7522	.3	023559
662100	.2	56*8*9
65322	.1	05
8000	.0	*000*068
2	−.0	355
1	−.1	67
6	−.2	

Note. Effect sizes from four unpublished Rorschach studies and one unpublished Minnesota Multiphasic Personality Inventory (MMPI) study are italicized.

these three measures for the MMPI suggests that one or more MMPI studies with large sample sizes have large effect sizes as well. This effect appears to be caused by several studies rather than a single outlier. Four of the five MMPI studies with the largest sample sizes (between 200 and 1,644 subjects) have effect sizes that are greater than the unweighted mean. Closer examination of the characteristics of these four studies (Bagby, Buis, & Nicholson, 1995; Ben-Porath et al., 1989; Friedman, Gleser, Smeltzer, Wakefield, & Schwartz, 1983; Morton-Page & Wheeler, 1997) revealed that they were all chosen for inclusion at the screening stage rather than being submitted to the judges, suggesting that the predictor–criterion pairs were obviously well matched for these particular studies.

Compared with the correlational results reported in Table 1 of Parker et al. (1988; weighted means of .46 for the MMPI and .41

Table 4
Meta-Analytic Summary of MMPI and Rorschach Studies

Statistic	MMPI (n = 30)	Rorschach (n = 30)
Central tendency (r)		
Unweighted M	.30	.29
Weighted M	.37	.26
Mdn	.22	.29
Significance		
Stouffer's Z	19.60	9.85
One sample t	5.22	4.70
Variability (r)		
Range	1.02	1.09
s	.26	.26
χ^2 for heterogeneity	630.86	112.68
Confidence interval for r[a]		
95%	.19–.40	.17–.39
99%	.15–.43	.13–.43
99.9%	.11–.46	.09–.46

Note. MMPI = Minnesota Multiphasic Personality Inventory.
[a] Confidence intervals are based on the number of studies, not the number of participants.

Table 5
Binomial Effect-Size Display for the Unweighted Mean Effect Size of MMPI Studies (r = .30)

Test indicator	Theoretically relevant criterion		Total
	Positive	Negative	
Positive	65	35	100
Negative	35	65	100
Total	100	100	200

for the Rorschach), the validity coefficients/effect sizes from the present meta-analysis are somewhat smaller; however, their magnitudes are still substantial. To enhance interpretation of the results and demonstrate their implications for test validity, we present binomial effect size displays (BESDs; Rosenthal & Rubin, 1982) for the unweighted mean effect sizes of the MMPI (Table 5) and the Rorschach (Table 6). The BESDs show the agreement of hypothetical dichotomous MMPI and Rorschach indicators with some theoretically relevant dichotomous criterion. When the data are cast in the standardized format of the BESD, the MMPI indicator agrees with the criterion 65% of the time, and the Rorschach indicator is in agreement with the criterion 64.5% of the time.

Overall, the results were significant for both MMPI ($Z = 19.60$, $p = 7 \times 10^{-86}$) and for Rorschach ($Z = 9.85$, $p = 3 \times 10^{-23}$). The stronger significance for MMPI reflects that MMPI studies typically had more participants ($N = 4,920$) than Rorschach studies ($N = 1,713$). When a random-effects view is adopted (considering the studies to be the units of analysis rather than the participants within individual studies), their significance is quite similar [$t(29) = 5.22$, $p = 7 \times 10^{-6}$ for MMPI vs. $t(29) = 4.70$, $p = 3 \times 10^{-5}$ for the Rorschach].

The variability of effect sizes for both tests is quite striking. Effect sizes ranged between −.26 and .76 for the MMPI ($SD = 0.26$) and between −.17 and .91 for the Rorschach ($SD = 0.26$). The stem and leaf plot in Table 3 illustrates that the effect sizes are spread relatively evenly across a broad range of values, rather than clustered tightly around a central value. Heterogeneity tests were significant for both MMPI [$\chi^2(29) = 630.86$, $p = 3 \times 10^{-114}$] and Rorschach [$\chi^2(29) = 112.68$, $p = 8 \times$

Table 6
Binomial Effect Size Display for the Unweighted Mean Effect Size of Rorschach Studies (r = .29)

Test indicator	Theoretically relevant criterion		Total
	Positive	Negative	
Positive	64.5	35.5	100
Negative	35.5	65.5	100
Total	100	100	200

Table 7
MMPI Moderator Analyses

	n			Mean r^a		
Category	Participants	Studies	Z	Unweighted	Weighted	SD of r
Core journal						
No	1,338	16	10.18	.26	.34	0.29
Yes	3,582	14	17.81	.34	.38	0.23
Sample						
Inpatient	1,508	7	15.05	.38	.46	0.29
Outpatient	905	12	7.42	.26	.27	0.27
Nonpatient	2,507	11	12.62	.29	.36	0.26
Inclusion						
During screening	3,433	13	25.20	.51	.47	0.21
By judges	1,487	17	4.00	.11	.12	0.16
Analytic method						
F or t	1,625	16	13.80	.30	.49	0.26
r	2,871	9	13.63	.29	.32	0.29
Mixed/other	424	5	5.05	.30	.25	0.29
Predictor type						
Codetype	154	1	2.68	.21	.21	—
Clinical/validity	1,835	18	14.99	.32	.47	0.27
Supplemental	2,300	7	13.29	.34	.37	0.30
Multiple types	631	4	2.97	.11	.12	0.16
Criterion type						
Objective	1,281	14	6.79	.20	.20	0.20
Psychiatric diagnosis	669	6	10.18	.37	.47	0.25
Observer rating	1,804	4	8.14	.17	.33	0.15
Self-report	466	3	9.20	.39	.48	0.53
"Projective" measure	190	1	2.73	.20	.20	—
Multiple types	510	2	15.64	.71	.70	0.03
MMPI version:						
MMPI	2,122	22	14.30	.31	.37	0.28
MMPI-2	2,798	8	14.25	.27	.37	0.23

Note. MMPI = Minnesota Multiphasic Personality Inventory. Dash indicates value was not computed due to insufficient sample size.
^{a} Means computed using Fisher-transformed Z_r, then transformed back to units of r.

10^{-12}].[17] Perhaps this variability should not be surprising, considering the great variation in study populations, test predictors, and criterion variables among the studies that were included in the meta-analysis. Despite this great variability, the 95% confidence intervals around the mean effect sizes (using the number of studies, not the number of participants) are reasonably narrow (.19–.40 and .17–.39 for the MMPI and Rorschach, respectively).

Comparisons Within Instruments

Summaries of study characteristics are found in Table 7 for MMPI studies and Table 8 for Rorschach studies. Random-effects style moderator analyses were conducted to compare studies with different characteristics, separately for MMPI and Rorschach.

Publication outlet. Studies published in one of the five core journals were compared with studies published in other journals. Differences were not significant for Rorschach [$t(28) = 0.22, p = .83, r = .04$] or MMPI [$t(28) = 0.71, p = .48, r = .13$]. This suggests that the Parker et al. (1998) meta-analysis, which only included studies from two of the five journals examined here, may not have been affected adversely by journal bias.

Study population. Linear contrasts were computed to examine whether effect sizes were larger for clinical populations than for studies using normal participants only (weights were −1 for non-

patient samples, 0 for outpatient samples, and +1 for inpatient samples). The contrast approached traditional significance levels for Rorschach studies ($t(27) = 1.37, p = .09, r = .26$), but was not significant for the MMPI studies ($t(27) = 0.66, p = .26, r = .13$).

Inclusion. We compared studies that were deemed appropriate for inclusion during initial screening with studies that were submitted to and approved by the judges. This moderator variable serves as an indicator of the degree of conceptual fit between predictors and criterion variables. Studies approved at screening used criterion variables that had obvious relationships with test predictors, whereas the studies chosen by the judges had validity tests that were more subtle or tenuous. Although the Rorschach effect sizes from these two categories were not reliably different ($t(28) = 0.39, p = .35, r = .07$), there was a strong tendency for MMPI effect sizes to be higher for the studies approved at initial screening than for studies approved by judges ($t(28) = 5.39, p = 5 \times 10^{-6}, r = .71$).

---

[17] Although the significant chi-square tests indicate that it is unlikely that the effect sizes are estimating unitary underlying population effect sizes for MMPI and for Rorschach, this does not prohibit us from making inferences concerning average effect sizes for each test.

Table 8
Rorschach Moderator Analyses

	n			Mean r[a]		
Category	Participants	Studies	Z	Unweighted	Weighted	SD of r
Core journal						
No	932	20	6.76	.28	.24	0.27
Yes	781	10	7.50	.30	.28	0.24
Sample						
Inpatient	331	7	5.24	.42	.29	0.34
Outpatient	644	10	6.86	.26	.32	0.24
Nonpatient	738	13	5.10	.23	.19	0.23
Inclusion						
During screening	728	10	8.22	.32	.34	0.21
By judges	985	20	6.25	.27	.20	0.28
Analytic method						
F or t	722	14	5.71	.23	.22	0.25
r	592	8	4.35	.18	.23	0.23
Mixed/other	399	8	7.16	.47	.38	0.25
Predictor type						
Dichotomous sign	264	3	3.89	.22	.28	0.29
Ratio or sum	1,276	21	7.77	.23	.24	0.23
Constellation or scale	68	3	2.62	.33	.39	0.28
Multiple types	105	3	4.08	.62	.36	0.35
Criterion type						
Objective	572	13	7.43	.37	.35	0.30
Psychiatric diagnosis	203	4	2.44	.18	.18	0.24
Observer rating	193	4	3.44	.28	.26	0.21
Self-report	541	6	5.68	.23	.28	0.22
"Projective" measure	204	3	0.87	.14	.03	0.26
Multiple types	0	0	—	—	—	—
Coding System						
Exner	509	11	6.24	.39	.26	.30
Other	994	17	6.64	.22	.25	.24
Mixed	210	2	4.17	.27	.31	.17
Rorschach variable type						
Structural	496	10	5.38	.27	.26	0.19
Content	681	10	4.44	.19	.19	0.21
Both	536	10	7.23	.39	.34	0.35

Note. Dash indicates value was not computed due to insufficient sample size.
[a] Means computed using Fisher-transformed Z_r, then transformed back to units of r.

Analytic method. Parker and colleagues (1988; in press) suggested that effect sizes based on Pearson correlations between continuous variables should be on average greater than those computed from t and F statistics, due to the greater precision in measurement associated with the former analyses. The data from the present investigation do not support this hypothesis, either for the MMPI [$t(27) = -0.09, p = .53, r = -.02$] or the Rorschach [$t(27) = -0.40, p = .65, r = -.08$].

Predictor type. For the Rorschach, a linear contrast was computed testing the association between measurement precision in the predictor variables and effect size magnitude (weights were -1 for dichotomous sign indicators, 0 for sums or ratios, and $+1$ for composite variables or scales). This contrast was not significant [$t(26) = 0.47, p = .32, r = .10$]. For the MMPI, studies using the 13 basic clinical and validity scales were compared with studies using supplemental or research scales. Although we expected studies using the original scales to have higher effect sizes, this contrast also failed to achieve significance [$t(26) = -0.11, p = .54, r = -.02$].

Criterion variable type. Analyses for these categories focused on the issue of method variance. We hypothesized that studies in

which predictor variables shared method variance with criterion variables (monomethod studies) would have larger effect sizes than studies using predictors and criterion variables measured using different methods (heteromethod studies). In the case of the MMPI, studies with self-report criterion variables were compared with all other studies. This contrast was not significant [$t(28) = 0.61, p = .27, r = .11$]. For the Rorschach, studies using "projective" measures as criterion variables were compared with other studies. This comparison likewise did not find meaningful differences [$t(28) = -0.81, p = .79, r = -.15$].

Other analyses. The correlation between effect size and year of publication was not significant for the MMPI ($r = .07, p = .35$), but this relationship was quite strong for Rorschach studies ($r = .45, p = .007$). We also compared Rorschach studies that used different scoring methods, in part to investigate whether the finding that more recent Rorschach studies have larger effect sizes was related to the widespread use of the Exner Comprehensive System in recent years. If the reliability of scoring under the Comprehensive System is greater than the reliability of other scoring methods, then validity effect sizes would be expected to be greater for studies using the Comprehensive System. Studies scored using the

Exner Comprehensive System had higher effect sizes than those scored using other methods, and this comparison approached traditional significance levels [$t(27)$ = 1.39, p = .09, r = .26]. However, it should be noted that the weighted mean effect sizes for these two types of Rorschach studies were essentially indistinguishable ($\bar{r}$ = .26 for studies using the Comprehensive System and $\bar{r}$ = .25 for studies using other scoring methods).

Rorschach studies that use predictors concerning structural features of percepts had effect sizes that were not significantly different from those using response content [$t(27)$ = 0.56, two-tailed p = .58, r = .13]. A comparison of studies using the original MMPI with those using the MMPI-2 also was not significant [$t(28)$ = -0.27, p = .61, r = -.05].

Comparisons Between Instruments

Table 9 contains comparisons between the two full sets of MMPI and Rorschach studies, and also between various subsets of the samples. For each comparison, the following two analyses were conducted: (a) a fixed-effects contrast, which considers the number of participants contributing to each effect size; and (b) a random-effects contrast, which uses the individual studies as the units of analysis.

As noted previously, the overall effect sizes for the full sample of studies are quite similar for MMPI and Rorschach. They are not significantly different from each other (p = .76, two-tailed under the fixed-effects model, p = .89, two-tailed under the random-effects model). We also investigated whether method variance affects this comparison, by focusing on a subset consisting of heteromethod validity studies only (excluding MMPI studies with self-report criterion variables and Rorschach studies with "projective" criterion variables). This subset was not substantially different from the full sample, and the mean effect sizes for MMPI and Rorschach still did not differ from each other. Meyer (1996) has argued that psychiatric diagnoses obtained using highly structured interviews make use of predominantly self-report data, and therefore studies using diagnosis to validate the MMPI have shared method variance. Furthermore, studies using psychiatric diagnosis as a criterion are subject to criterion contamination, if test data are used in making diagnostic decisions (Anastasi & Urbina, 1997). When MMPI and Rorschach studies using psychiatric diagnoses as

criterion variables are removed from the heteromethod subset, a small but nonsignificant advantage for the Rorschach emerges.

Four further comparisons were made, examining differences between Rorschach and MMPI studies within criterion variable types. First, we examined studies with objective criterion variables, involving the prediction of unambiguous outcomes (e.g., arrest or hospitalization) or behavior (e.g., treatment attendance), or the discrimination of objectively different groups (e.g., patients with closed-head injury vs. normal controls). The Rorschach was superior to the MMPI in such studies under the fixed-effects model (p = .007, two-tailed), and although the contrast did not achieve traditional significance levels in the random-effects analysis (p = .18, two-tailed), the effect size associated with the random-effects analysis was substantial (r = .27). Another comparison examined differences between the MMPI and the Rorschach in studies using psychiatric diagnoses as criterion variables. Considering that the discrimination of psychiatric groups is the purpose for which the MMPI was originally developed, it is not surprising that MMPI effect sizes are higher than Rorschach effect sizes for these studies (the method variance issue discussed above should also be kept in mind). This comparison also achieved traditional significance levels in the fixed-effects analysis (p = .02, two-tailed) but not in the random-effects analysis (p = .27, two-tailed), despite having a noteworthy effect size (r = .39). A third comparison between Rorschach and MMPI studies, using observer ratings as criterion variables, did not yield statistically reliable differences under either model. The fourth comparison, examining only studies using self-report measures, showed that MMPI studies had significantly higher effect sizes than Rorschach studies under the fixed-effects model (p = .008, two-tailed), but the random-effects analysis did not yield a statistically reliable difference (p = .54, two-tailed).

In a final random-effects analysis, we conducted a 2 (tests) × 4 (criterion types) analysis of variance (ANOVA) on the effect sizes of those studies using psychiatric diagnoses, objective criteria, observer ratings, or self-report measures as criterion variables (other criterion variable types were not numerous enough for inclusion in this analysis). The main effects did not indicate reliable differences between tests, $F(1, 46)$ = 0.05, p = .82, r = .03, or among criterion types, $F(3, 46)$ = 0.13, p = .95. The interaction effect was also nonsignificant, $F(3, 46)$ = 1.24, p =

Table 9
Comparisons Between MMPI and Rorschach Studies

Comparison	MMPI No. of studies	MMPI n	Rorschach No. of studies	Rorschach n	Unweighted mean effect size MMPI	Unweighted mean effect size Rorschach	Fixed-effects analysis Z^a	Fixed-effects analysis p^b	Random-effects analysis t^a	Random-effects analysis p^b	Random-effects analysis r^a
Global	30	4,920	30	1,713	.30	.29	0.30	.76	0.14	.89	.02
Excluding monomethod studies	27	4,454	27	1,509	.29	.30	-0.39	.70	-0.19	.85	-.03
Excluding monomethod studies and psychiatric diagnoses	21	3,785	23	1,306	.26	.32	-1.39	.16	-0.68	.50	-.10
Objective criterion variables only	14	1,281	13	572	.20	.37	-2.68	.007	-1.39	.18	-.27
Psychiatric diagnoses only	6	669	4	203	.37	.18	2.33	.02	1.18	.27	.39
Observer ratings only	4	1,804	4	193	.17	.28	-1.01	.31	-0.78	.46	-.30
Self-report measures only	3	466	6	416	.39	.23	2.67	.008	0.64	.54	.24

Note. MMPI = Minnesota Multiphasic Personality Inventory.
[a] Positive values indicate larger effect sizes for MMPI studies, whereas negative values indicate larger effect sizes for Rorschach studies. [b] Two-tailed.

.31. Although there were no a priori predictions concerning the pattern of findings among criterion types, we conducted a contrast analysis for the criterion type main effect, based on examination of the condition means. Contrast weights used were 0 for psychiatric diagnoses, 0 for objective criteria, −1 for observer ratings, and +1 for self-report measures. This contrast did not achieve traditional significance levels, $F(1, 46) = 0.32$, $p = .57$, $r = .08$. Another contrast was conducted in the interaction term, also based on examination of the means. The contrast weights for the MMPI studies were 1 for psychiatric diagnoses, −1 for objective criteria, −1 for observer ratings, and +1 for self-report measures; for the Rorschach studies, the sign of these weights were reversed. This contrast had a stronger significance level and a larger effect size, $F(1, 46) = 2.79$, $p = .10$, $r = .24$. However, this result must be interpreted conservatively, as any reasonable Bonferroni correction would reduce the significance of the latter contrast considerably.

Retrievability Bias

A question often asked in meta-analysis is whether the studies included may have a bias toward larger effect sizes, because studies with significant findings are more likely to be published (Greenwald, 1975; Sterling, 1959) and are therefore easier to retrieve (Rosenthal, 1979). This question is especially important for this meta-analysis, because the number of unpublished investigations using the MMPI and the Rorschach is undoubtedly large, and we were unable to obtain a representative sample of them. Although it is difficult (if not impossible) to determine the effect of retrievability bias if it exists, we use the following two common strategies for assessing the potential magnitude of the problem: funnel plots and file drawer analysis.

Funnel plots. This technique, suggested by Light and Pillemer (1984), involves generating a scatter plot of effect size against sample size for each study included in the meta-analysis. If the studies all estimate a common effect size, and no retrievability bias is present, then the graph assumes the characteristic shape of a funnel. The wide mouth occurs where studies with small sample sizes have a broad range of effect sizes, and the shape of the cloud of points narrows to a small spout, at the point where large studies converge on the population effect size value. If publication bias prevents the inclusion of studies that do not achieve statistical significance, then this is manifested as a "bite" taken out of the funnel, occurring in the region where studies with near-zero effect sizes ought to be.

Separate funnel plots for the MMPI and the Rorschach are found in Figure 1 and Figure 2, respectively. Despite the great variation in effect sizes described earlier, the funnel shape is clearly discernable for both plots. For the MMPI, there is no indication of any region where studies may be missing, whereas, the Rorschach funnel plot may have a slight indication of bias, because there are no studies with sample size less than about 25 and effect size less than zero. However, there are several small studies with effect sizes estimated to be zero because of imprecise reporting—it is possible that some of these effect sizes were, in truth, negative. Furthermore, several larger published Rorschach studies were obtained with effect sizes near zero or less than zero. Thus, it appears that studies with null results were not necessarily prevented from
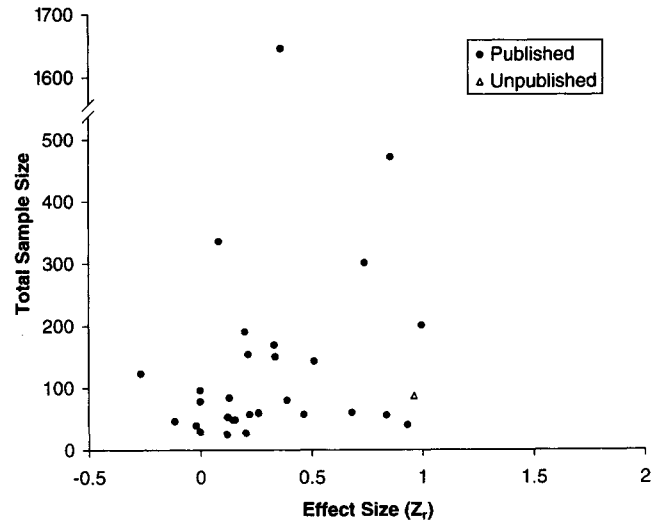

Figure 1. Funnel plot of effect sizes for MMPI studies ($n = 31$).

being published. Any bias in this meta-analysis due to unrepresented unpublished studies is probably small.

File drawer analysis. An indirect way of addressing the retrievability bias issue is to ask, "how many unobtainable studies must there be, all averaging null results, in order to make the overall significance level greater than .05?" Using a procedure described by Rosenthal (1979), we determined that 1,045 null Rorschach studies must be hidden away in file drawers in order to make the Rorschach results nonsignificant, and 4,230 null MMPI studies would be required to reduce its significance beneath the traditional threshold.[18] Although these numbers are rather large, it is certainly plausible that unpublished studies exist in quantities exceeding these limits, given the widespread use of these instruments. However, it should be emphasized that these file drawer limits are based on only a very small sample from the vast published literature, and the magnitude of these limits reflects the size of our samples more than the nature of the literature. Assuming that other published results are similar to those we obtained here, if we had sampled twice as many Rorschach studies, then the new file drawer number would be about 4,241 studies. For the MMPI, doubling the number of sampled studies would change the file drawer number to about 16,981. Although some bias may be present because of unrepresented unpublished studies, the likelihood that these unpublished studies would reduce the significance of results below traditional levels is very small.

Discussion

In a meta-analytic comparison of criterion-related validity coefficients for the MMPI and for the Rorschach, we found both instruments to have validity effect sizes of substantial magnitude (unweighted mean r of .30 and .29 for MMPI and Rorschach, respectively). Validity estimates for the MMPI and Rorschach were not reliably different from each other, even when studies in

---

[18] These calculations excluded the four unpublished Rorschach studies and the single MMPI study that was unpublished.
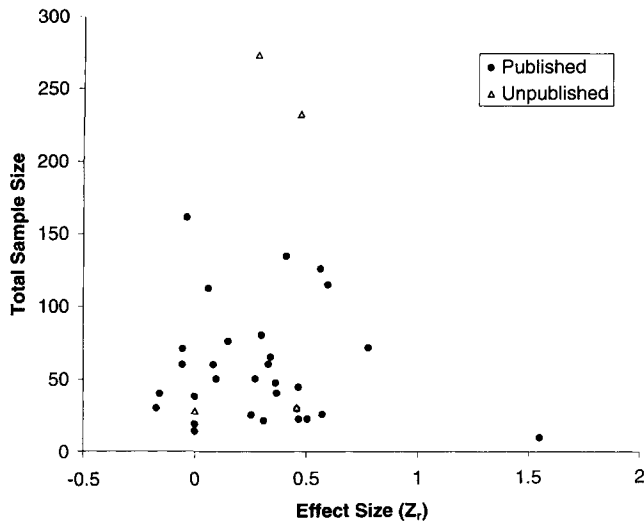
Figure 2. Funnel plot of effect sizes for Rorschach studies ($n = 34$).

which test predictors and criterion variables had common measurement methods were removed from consideration. Our finding of no global differences between these tests echoes results from two previous meta-analyses (Atkinson, 1986; Parker et al., 1988), but it contradicts the finding by Garb et al. (1998) of higher validity coefficients for the MMPI than for the Rorschach. The methodological features of this study, including random sampling from the published literature, expert judgments for inclusion of validity evidence, and the use of accepted effect size estimation techniques, lend greater credibility to these results compared with those from previous efforts.

The magnitude of validity coefficients obtained in this investigation are not directly comparable to those reported by Atkinson (1986) or Garb et al. (1998), because these authors used problematic indices of effect size. When the current results are compared to the correlational results reported in Table 1 of Parker et al. (1988), the new estimates for both Rorschach and MMPI validity were somewhat smaller than the Parker et al. estimates. Parker et al. reported mean effect sizes weighted by the number of subjects in each study, $\bar{r} = .46$ for the MMPI, and $\bar{r} = .41$ for the Rorschach. In the present investigation, weighted mean effect sizes were $\bar{r} = .37$ for the MMPI and $\bar{r} = .26$ for the Rorschach. Most likely, reduced bias in the selection of studies and in the selection of effect sizes resulted in the smaller validity figures we report. The problems in effect size estimation noted by Garb et al. (1998; in press) for the Parker et al. meta-analysis may also partially explain differences between the Parker et al. estimates and our own. These smaller revised estimates of Rorschach and MMPI criterion-related validity coefficients should not cause loss of confidence in either instrument, however, because these figures are still quite high for personality measures. As noted by Cohen (1988), "when one looks at near-maximum correlation coefficients of personality measures with . . . real-life criteria, the values one encounters fall at the order of . . . $r = .30$" (p. 81). In other words, validity for these instruments is about as good as can be expected for personality tests.

It is worth emphasizing that the global validity estimates were computed from studies with a wide variety of test predictors,

criterion variables, and study populations. Indeed, measures of personality should by definition be predictive of a broad range of behaviors in a broad range of situations. To make finer discriminations between the relative strengths of the Rorschach and the MMPI, we identified potential moderator variables and conducted subset analyses. When we examined studies with different types of criterion variables separately, we found that the MMPI was superior to the Rorschach in predicting self-report criterion variables and in making discriminations among psychiatric groups, whereas the Rorschach had higher validity coefficients in studies using objective outcomes as criterion variables. The MMPI probably has higher correlations with self-report measures than Rorschach because of shared method variance between MMPI and other self-report measures. The finding of an MMPI advantage for making diagnostic classifications is also readily understandable, given the empirical criterion-keying procedure that was used in MMPI development. It is less obvious why Rorschach validity should exceed that of the MMPI when objective criterion variables are used. Meyer (1996) has noted that self-report instruments like the MMPI may be more prone to being affected by participants' self-presentation biases than is the case for the Rorschach. Indeed, one of the studies contained in the present meta-analysis (Bornstein, Rossner, Hill, & Stepanian, 1994) showed that a Rorschach dependency measure was less susceptible to conscious manipulation by participants that was a common self-report measure of dependency. Perhaps objective criterion variables are more accurately predicted when respondents' attempts at impression management are in some way filtered out. In these instances, it may be that the Rorschach is more "objective" than so-called objective instruments.

We paid particular attention to the issue of method variance, in accordance with its important role in efforts at construct validation (Campbell & Fiske, 1959; Cronbach, 1995). When monomethod validity coefficients and heteromethod coefficients were compared within the MMPI, we found that shared method variance between self-report criterion variables and MMPI resulted in elevated monomethod coefficients. This is consistent with speculations by McClelland (1980) and Meyer (1996) that monomethod validation strategies may inflate validity coefficients for pencil and paper personality measures. However, monomethod validity coefficients for the Rorschach, obtained from studies using other "projective" measures as criterion variables, were actually nonsignificantly smaller than heteromethod Rorschach coefficients. This counterintuitive result suggests the possibility that the monomethod studies using "projective" criterion variables to validate the Rorschach may not be truly monomethod after all. The Rorschach may have psychometric properties that are different from, and perhaps superior to, those of other "projective" measures that are commonly mentioned in the same breath as the Rorschach. Nevertheless, when the comparison between MMPI and Rorschach was restricted to heteromethod validity coefficients only, removing the advantage conferred to the MMPI by monomethod validity studies, the difference between the two instruments was still negligible.

Other moderator analyses addressed unresolved issues in the literature on these tests. First, we did not find evidence to support the hypothesis by Parker et al. (1988, in press) that effect size rs that are reported directly are larger than effect sizes computed from F or t tests. Earlier findings consistent with this hypothesis (Atkinson, 1986; Parker et al., 1988) may be artifactual, because

ω^2 was used for t and F results in these investigations, and any given ω^2 is smaller than the equivalent r^2. When effect sizes were all properly represented in units of r, there were no differences among effect sizes computed from t, F, and r. We concur with Cohen's (1983) observation that dichotomization of a continuous independent variable results in reduction in power and effect size due to loss of information; however, in the context of the current investigation, the loss of information attributable to discrete measurement of independent variables was negligible in terms of overall magnitude of effect.

In another set of moderator analyses, we found a strong relationship between year of publication and effect size for Rorschach studies ($r = .45$), but not for MMPI studies ($r = .07$). To explain the trend of rising Rorschach effect sizes over time, we hypothesized that the introduction and almost universal adoption of the Exner Comprehensive System for the Rorschach may have played a role. The Comprehensive System incorporated many of the strongest features of earlier systems and might be more valid as a composite than any of the constituent scoring systems alone. It is also thought that the Comprehensive System has greater reliability than other Rorschach systems, and this might lead to improved validity as well. We note that the reliability of the Comprehensive System is an issue of some controversy, however (see Meyer, 1997a, 1997b; and Wood, Nezworski, & Stejskal, 1997). When we compared effect sizes from studies using the Comprehensive System to those from studies using other systems, we found only weak evidence suggesting that effect sizes were higher for the Comprehensive System. The rising trend in Rorschach effect sizes over time cannot be explained completely by the effect of the Comprehensive System. Perhaps the average quality of Rorschach research has increased over the past 20 years. It is also possible that journal editors have become more selective in recent years, tending to accept only manuscripts with strong, unambiguous findings.

We used judges who are experts on the Rorschach and the MMPI to make determinations of whether experimental findings appropriately addressed the validity of either instrument, and therefore warranted inclusion in the meta-analysis. This approach has the considerable advantage that it does not rely on the original author to formulate and articulate sound hypotheses in order for a study to be included. Even if a relevant finding has been stumbled on by chance, it would have been included in the present meta-analysis as long as both judges agreed it spoke to test validity. However, the potential for bias to affect validity judgments should be noted. It would be reasonable to surmise that judges expect to find high validity for the tests they are expert in. Furthermore, even though the inclusion judgments were made blindly with respect to authors' original hypotheses and study results, it is likely that judges were nevertheless familiar with at least some of the studies included. In fact, several studies included in the meta-analysis were authored by judges (Bornstein, Leone, & Galley, 1988; Bornstein et al., 1994; Fowler, Hilsenroth, & Handler, 1995). It is possible that the judges could have leaned toward inclusion of results that they remembered as significant, exerting upward bias on the meta-analytic validity estimates. Another potential source of bias in the judgments is due to the broad range of populations and applications for which both the Rorschach and the MMPI are used in the research literature. To make good decisions about validity and inclusion for all the studies they were presented with, our judges would need to be knowledgeable about a host of different research domains, including topics as diverse as dysmenorrhea (Bloom, Shelton, & Michaels, 1978), acculturation among the Nepalese (Regmi, 1986), and the effects of amphetamines (Perry et al., 1995). By requiring that both judges agree on the effect sizes that constitute validity evidence in order for studies to be included, the likelihood that esoteric or controversial findings would be included was reduced—especially considering that the interrater reliability of inclusion judgments was not very high (.35 for Rorschach and .39 for MMPI). Thus, the inclusion strategy for this meta-analysis was fairly conservative, and there may have been a tendency for more speculative findings to be excluded from the current sample. This too may have led to upward bias in validity estimates. It is also likely that the sample of studies we studied was influenced by the particular judges we used, and that different judges would have yielded a different sample, and perhaps even different results. However, we do not see a practical, principled alternative to the use of expert judgment for validity determinations. We do not believe that the low reliability among inclusion judgments is indicative of any deficiencies in our judges; rather, this may simply mean that validity, like beauty, is a subjective concept that is to some degree in the eye of the beholder. Meta-analysts examining validity of personality measures may wish to use more than two judges, in order to ensure greater reliability of inclusion decisions.

Another methodological feature of the current meta-analysis that is worthy of note is our emphasis on unweighted mean effect sizes for both descriptive and inferential statistics. Sometimes in meta-analysis, it is desirable to weight each effect size by its degrees of freedom, because effect sizes based on larger samples are estimated with more precision than are those based on smaller samples. Use of weighted effect sizes is particularly valuable when there is reason to believe that the constituent effect sizes in a meta-analysis are all estimates of a unitary, common population effect size. However, when the studies in a meta-analysis are heterogeneous in terms of populations, methods, and measures, as they are here, it is often preferable to analyze and interpret unweighted mean effect sizes. The goal of such meta-analyses is usually to evaluate the average effect across diverse conditions. In such situations, weighted mean effect sizes may place too much emphasis on one or more studies with idiosyncratic features, simply because their samples are larger than those of other studies. This seems to be the case in the current meta-analysis. When we inspected the weighted mean effect sizes, the MMPI seemed to fare better than the Rorschach (weighted mean effect sizes of .37 and .26, respectively). However, the fact that the median MMPI effect size (.22) was considerably lower suggested that a few studies with large sample sizes were exerting strong upward influence on the MMPI weighted mean. When four of these studies were identified and examined, we learned that they were all studies in which predictor variables bore obvious conceptual relationships with the criterion variables (e.g., a correlation between the Beck Depression Inventory and the MMPI Depression scale; Morton-Page & Wheeler, 1997). Thus, the validational tasks in these four studies may have been particularly easy. We elected to focus on unweighted mean effect sizes because unweighted means represent such features of studies in the proportion that they are found in the random samples of studies, rather than according to the number of participants in individual studies.

The discussion in the preceding paragraph highlights one of the limitations in our meta-analysis. Clearly, some of the studies addressing Rorschach or MMPI validity can be expected to have higher validity coefficients than others for reasons that are not directly related to the overall validity of the instrument per se. For instance, a study using the Sc scale from the MMPI to compare samples of psychotic and psychiatrically healthy persons should have a higher validity coefficient than a study using the same scale to differentiate samples of psychotic and neurotic patients. In the second study, restriction of range in the criterion variable (psychiatric status) renders the validation task more difficult, perhaps resulting in a smaller validity coefficient. Similar effects on the magnitude of validity coefficients might be expected according to the degree of conceptual overlap between predictor and criterion variables. The previously mentioned study in which the correlation between the MMPI Depression scale and the Beck Depression Inventory was examined (Morton-Page & Wheeler, 1997) is a good example of a predictor-criterion pair with considerable conceptual overlap, and unsurprisingly the effect size is rather large (.76). Consider now a hypothetical study in which the MMPI Depression scale is correlated with a questionnaire measure of anxiety. We would expect these measures to be correlated, because anxiety and depression are highly comorbid conditions, and hence this study also addresses the validity of the Depression scale. However, a measure of anxiety will be less highly correlated with a measure of depression than will two measures of depression. Such differences among effect sizes that are not directly related to test validity were not well addressed in our meta-analysis. We examined one variable that bears on this issue indirectly: a dichotomous indicator for whether studies were selected for inclusion at the screening stage because the rationale for the study and the measures used were clearly appropriate, or submitted to the judges for evaluation because the validity issues were less clear. This variable can be construed as a rough proxy for the type of problems described above. We do note one significant limitation of this variable: The reliability of the screening decisions was not assessed. Despite unknown reliability, we found that MMPI studies included during screening had significantly greater effect sizes than did MMPI studies that were approved by the judges. The parallel comparison for Rorschach studies was not significant. Future investigations may benefit from more explicit attention to the confounds described above, because if differences are found between effect sizes for different tests, it could be due to either true validity differences or to differences in the nature of the validation strategies used.

A further limitation of this meta-analysis that should be acknowledged concerns our failure to obtain a representative sample of unpublished studies, despite our vigorous efforts.[19] The seeming reluctance of many researchers to share the contents of their file drawers certainly gives us pause and may indicate that unpublished studies in this area tend to have nonsignificant findings. However, we are heartened by the funnel plots of both samples, which indicate that retrieval bias probably had a small effect on the outcomes at most. The procedure we adopted for selecting effect sizes may have mitigated the effect of publication or retrieval bias, because the effect sizes selected by the judges were not necessarily the ones emphasized by study authors. Therefore, evidence that was not supportive of validity was often available to us because it

was contained in reports that were published because other scales, irrelevant for our purposes, had significant findings.

The file drawer analyses suggest that it is unlikely that unpublished null studies exist in quantities necessary to reduce the overall significance of the published results for each test to marginal levels. However, significance is admittedly not the most appropriate basis for evaluation of test validity. Given enough studies with enough participants, even an instrument with a validity coefficient of .03 could obtain significance at $p < .05$, yet it would not be psychometrically very encouraging. We cannot tell for certain what bias has been introduced into our analysis by the underrepresentation of unpublished reports, but available evidence suggests that the bias is not great.

We would urge others to interpret the findings from this meta-analysis conservatively. It would be a mistake to accept either the MMPI or the Rorschach as universally valid on the basis of this meta-analysis. The broad variability we found among validity coefficients for both instruments suggests the possibility that some individual indices and scales may be more valid than others, in some populations, for some purposes. It would be similarly erroneous to conclude that Rorschach and MMPI are equally valid, and therefore interchangeable, in the specific domains for which both instruments are commonly used. The nonequivalence of these measures is underscored by the fact that even Rorschach and MMPI variables that are thought to be measures of similar constructs are typically poorly correlated with each other (Archer, 1997; Archer & Krishnamurthy, 1993a, 1993b).

We have attempted to answer the question, "In general, is the Rorschach as valid as the MMPI?" We conclude that on average, both tests work about equally well when used for purposes deemed appropriate by experts. Although this is important information, it comprises only a first step. Efforts to validate the MMPI or the Rorschach as a whole have only limited utility. Weiner (1996) asserted that multidimensional instruments such as the Rorschach and the MMPI can be most usefully validated at the scale level rather than at the level of the entire instrument. To raise the level of discourse regarding the relative merits of the Rorschach and the MMPI, we must move beyond the question of global validity and concentrate instead on the validity of individual scales and indices for specific purposes. The literature on both of these instruments is certainly well enough developed to support more tightly focused meta-analyses, such as the work on the Rorschach Prognostic Rating Scale by Meyer and Handler (1997). Similar meta-analyses have been conducted examining the ability to detect overreporting of symptoms using the MMPI (Berry, Baer, & Harris, 1991) and the MMPI-2 (Rogers, Sewell, & Salekin, 1994); another meta-analysis examined the detection of symptom underreporting using the MMPI (Baer, Wetter, & Berry, 1992). It is perhaps less glamorous (and certainly more labor intensive) to follow the path we suggest, making a series of limited inferences about the validity of test components rather than making grand pronouncements about the overall validity of entire instruments. However, this is the stuff of which science is made, and we should not shrink from this important task.

---

[19] We also note that we did not sample from book chapters or from the foreign language literature on these instruments.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Aaronson, A. L., Dent, O. B., & Kline, C. D. (1996). Cross-validation of MMPI and MMPI-2 predictor scales. *Journal of Clinical Psychology, 52,* 311–315.

*Aebischer, G. L. (1994, April). *Content analysis of Rorschach percepts of sexually abused and neglected/physically abused children and adolescents.* Paper presented at the meeting of the Society for Personality Assessment, Chicago, IL.

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

Archer, R. P. (1997). MMPI-Rorschach interrelationships: Proposed criteria for evaluating explanatory models. *Journal of Personality Assessment, 67,* 504–515.

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.

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

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

*Bagby, R. M., Buis, T., & Nicholson, R. A. (1995). Relative effectiveness of the standard validity scales in detecting fake-bad and fake-good responding: Replication and extension. *Psychological Assessment, 7,* 84–92.

*Ben-Porath, Y. S., Hostetler, K., Butcher, J. N., & Graham, J. R. (1989). New subscales for the MMPI-2 Social Introversion (Si) scale. *Psychological Assessment, 1,* 169–174.

*Berg, J. L. (1990). Differentiating ego functions of borderline and narcissistic personalities. *Journal of Personality Assessment, 55,* 537–548.

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

*Bloom, L. J., Shelton, J. L., & Michaels, A. C. (1978). Dysmenorrhea and personality. *Journal of Personality Assessment, 42,* 272–276.

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

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

*Bowman, M. E., Bennett, T. S., & Welsh, M. C. (1981). Comparison of three objective measures of introversion–extroversion. *Psychological Reports, 48,* 469–470.

*Brown, H. J. D., & Gutsch, K. U. (1985). Cognitions associated with a delay of gratification task: A study with psychopaths and normal prisoners. *Criminal Justice and Behavior, 12,* 453–462.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193–213). New York: Russell Sage Foundation.

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 in the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

*Campo, V., Dow, N., & Tuset, A. (1988). Rorschach, O. R. T., and follow-up. *British Journal of Projective Psychology, 33,* 31–53.

*Chaney, E. F., Roszell, D. K., & Blaes, P. (1982). Drug counselors' perceptions of clients: Demographic, personality and psychopathology correlates. *Drug & Alcohol Dependence, 9,* 243–250.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

*Cooper, D. E., & Holmstrom, R. W. (1984). Relationship between alexithymia and somatic complaints in a normal sample. *Psychotherapy and Psychosomatics, 41,* 20–24.

Cronbach, L. J. (1995). Giving method variance its due. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 145–157). Hillsdale, NJ: Erlbaum.

*Cruz, E. B., Brier, N. M., & Reznikoff, M. (1997). An examination of the relationship between form level ratings on the Rorschach and learning disability status. *Journal of Psychology, 131,* 167–174.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.

*Ellis, D. W., & Zahn, B. S. (1985). Psychological functioning after severe closed head injury. *Journal of Personality Assessment, 49,* 125–128.

Exner, J. E. (1974). *The Rorschach: A comprehensive system: Vol. 1.* New York: Wiley.

Exner, J. E. (1978). *The Rorschach: A comprehensive system: Vol. 2. Current research and advanced interpretation.* New York: Wiley.

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

*Exner, J. E., Colligan, S. C., Boll, T. J., Stischer, B., & Hillman, L. (1996). Rorschach findings concerning closed head injury patients. *Assessment, 3,* 317–326.

Eysenck, H. J. (1959). The Rorschach Inkblot Test. In O. K. Buros (Ed.), *The fifth mental measurements yearbook* (pp. 276–278). Highland Park, NJ: Gryphon Press.

*Fals-Stewart, W., & Schafer, J. (1993). MMPI correlates of psychotherapy compliance among obsessive–compulsives. *Psychopathology, 26,* 1–5.

*Faull, R., & Meyer, G. J. (1993, March). *Assessment of depression with the MMPI-2: Distinctions between Scale 2 and DEP.* Paper presented at the annual meeting of the Society for Personality Assessment, San Francisco, CA.

*Fontana, D., & Rees, V. (1982). Primary dysmenorrhea, educational performance, and cognitive and affective variables in adolescent schoolgirls. *British Journal of Educational Psychology, 52,* 199–204.

*Fowler, C., Hilsenroth, M. J., & Handler, L. (1995). Early memories: An exploration of theoretically derived queries and their clinical utility. *Bulletin of the Menninger Clinic, 59,* 79–98.

*Frank, M. A., Tuber, S. B., Slade, A., & Garrod, E. (1994). *Psychoanalytic Psychology, 11,* 475–490.

*Friedman, A. F., Gleser, G. C., Smeltzer, D. J., Wakefield, J. A., & Schwartz, M. S. (1983). MMPI overlap item scales for differentiating psychotics, neurotics, and nonpsychiatric groups. *Journal of Consulting and Clinical Psychology, 51,* 629–631.

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., Florio, C. M., & Grove, W. M. (in press). The Rorschach controversy: Reply to Parker, Hunsley, and Hanson. *Psychological Science.*

*Gas, Z. (1987). The value of the MacAndrew Alcoholism Scale for

differential diagnosis: Pilot study. *Polish Psychological Bulletin, 18,* 321–324.

*Gayton, W. F., Burchstead, G. N., & Matthews, G. R. (1986). An investigation of the utility of an MMPI posttraumatic stress disorder subscale. *Journal of Clinical Psychology, 42,* 916–917.

*Gordon, M., & Oshman, H. (1981). Rorschach indices of children classified as hyperactive. *Perceptual and Motor Skills, 52,* 703–707.

*Goshtautas, A., & Rugevitchius, M. (1983). Psychological characteristics and heart rate response to exercise and emotional challenges. *Activitas Nervosa Superior, 25,* 256–260.

*Gottschalk, L. A., Stein, M. K., & Shapiro, D. H. (1997). The application of computerized content analysis of speech to the diagnostic process in a psychiatric outpatient clinic. *Journal of Clinical Psychology, 53,* 427–441.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–12.

*Harper, G., & Scott, R. (1990). Learning disabilities: An appraisal of Rorschach response patterns. *Psychological Reports, 67,* 691–696.

Hayes, A. F. (1998). Within-study meta-analysis: Pooling the significance of doubly nonindependent (nonoverlapping) correlations. *Psychological Methods, 3,* 32–45.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.

Hedges, L. V. (1994a). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

Hedges, L. V. (1994b). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29–38). New York: Russell Sage Foundation.

*Hirschfeld, R. M. A., Klerman, G. L., Clayton, P. J., Keller, M. B., McDonald-Scott, P., & Larkin, B. H. (1983). Assessing personality: Effects of the depressive state on trait measurement. *American Journal of Psychiatry, 140,* 695–699.

Jensen, A. R. (1965). The Rorschach Inkblot Test. In O. K. Buros (Ed.), *The sixth mental measurements yearbook* (pp. 501–509). Highland Park, NJ: Gryphon Press.

*Karp, S. A., & Gemert, J. J. (1995, March). *An objective questionnaire supplement to the Rorschach: Preliminary findings.* Paper presented at the meeting of the Society for Personality Assessment, Atlanta, GA.

*Kaser-Boyd, N. (1993). Rorschachs of women who commit homicide. *Journal of Personality Assessment, 60,* 458–470.

*Kiran Kumar, S. K., & Thimmappa, M. S. (1982). Psychological differentiation and anxiety in emotionally disturbed children. *Indian Journal of Clinical Psychology, 9,* 1–5.

*Kornfeld, A. D. (1995). Police officer candidate MMPI-2 performance: Gender, ethnic, and normative factors. *Journal of Clinical Psychology, 51,* 536–540.

*Labott, S. M., Preisman, R. C., Torosian, T., Popovich, J., & Ianuzzi, M. C. (1996). Screening for somatizing patients in the pulmonary subspecialty clinic. *Psychosomatics, 37,* 327–338.

*Laudeman, K. A. (1977). Personality, drinking patterns, and problem drinking among young adult offenders. *Journal of Drug Education, 7,* 259–269.

*Leavitt, F., & Labott, S. M. (1996). Authenticity of recovered sexual abuse memories: A Rorschach study. *Journal of Traumatic Stress, 9,* 483–496.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935–1982. *American Psychologist, 39,* 451–454.

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.

*Meloy, J. R. (1984). Thought organization and primary process in the parents of schizophrenics. *British Journal of Medical Psychology, 57,* 279–281.

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

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

Meyer, G. J. (1997b). Thinking clearly about reliability: More critical corrections regarding the Rorschach comprehensive system. *Psychological Assessment, 9,* 495–498.

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., Murphy, S. G., Kiolbasa, T., Daly, M., Orput, D., Zelko, F. A. J., Remer-Osborn, J., & Zillmer, E. A. (1993, March). *Neuropsychological factors and Rorschach performance in children.* Paper presented at the meeting of the Society for Personality Assessment, San Francisco, CA.

*Meyer, G. J., & Resnick, J. D. (1996, July). *Assessing ego impairment: Do scoring procedures make a difference?* Paper presented at the Fifteenth International Rorschach Conference, Boston, MA.

*Morrison, T. L., Edwards, D. W., & Weissman, H. N. (1994). The MMPI and MMPI-2 as predictors of psychiatric diagnosis in an outpatient sample. *Journal of Personality Assessment, 62,* 17–30.

*Morton-Page, D., & Wheeler, M. S. (1997). Early loss, depression, and lifestyle themes. *Individual Psychology, 53,* 58–66.

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, K. C. H., Hunsley, J., & Hanson, R. K. (in press). Old wine from old skins sometimes tastes like vinegar: A response to Garb, Florio, & Grove. *Psychological Science.*

*Perconte, S. T., & Goreczny, A. J. (1990). Failure to detect fabricated posttraumatic stress disorder with the use of the MMPI in a clinical population. *American Journal of Psychiatry, 147,* 1057–1060.

*Perconte, S. T., Griger, M. L., & Bellucci, G. (1989). Psychometric evaluation of biogenic versus psychogenic erectile dysfunction. *Annals of Sex Research, 2,* 335–344.

*Perry, W., Sprock, J., Schaible, D., McDougall, A., Minassian, A., Jenkins, M., & Braff, D. (1995). Amphetamine on Rorschach measures in normal subjects. *Journal of Personality Assessment, 64,* 456–465.

*Pierloot, R. A., & Houben, M. E. (1978). Estimation of body dimensions in anorexia nervosa. *Psychological Medicine, 8,* 317–324.

Piotrowski, C., & Keller, J. W. (1992). Psychological testing in applied settings: A literature review from 1982–1992. *Journal of Training and Practice in Professional Psychology, 6,* 74–82.

Raudenbush, S. W. (1994). Random effect models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

*Regmi, M. P. (1986). The culture-contact and the personality structure of the Gurungs: A Rorschach study. *Psychologia, 29,* 50–55.

*Ridley, S. E. (1987). The high score approach to scoring two Rorschach measures of cognitive development. *Journal of Clinical Psychology, 43,* 390–394.

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

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

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

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research.* New York: McGraw-Hill.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis.* New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

*Sah, A. P. (1989). Personality characteristics of accident free and accident involved Indian railway drivers. *Journal of Personality and Clinical Studies, 5,* 203–206.

*Salyer, K. M., Holmstrom, R. W., & Noshpitz, J. D. (1991). Learning disabilities as a childhood manifestation of severe psychopathology. *American Journal of Orthopsychiatry, 61,* 230–240.

*Schlesinger, L. B., & Fox, C. F. (1980). Achromatic Rorschach perceptions: Some implications for the diagnosis of depression. *Perceptual and Motor Skills, 50,* 199–202.

*Schmidt, M. M., & Miller, W. R. (1983). Amount of therapist contact and outcome in a multidimensional depression treatment program. *Acta Psychiatrica Scandinavica, 67,* 319–332.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

*Shealy, R. C., Lowe, J. D., & Ritzler, B. A. (1980). Sleep onset insomnia: Personality characteristics and treatment outcome. *Journal of Consulting and Clinical Psychology, 48,* 659–661.

*Sheehan, J. G., & Tanaka, J. S. (1983). Prognostic validity of the Rorschach. *Journal of Personality Assessment, 47,* 462–465.

*Simon, M. J. (1985). The Egocentricity index and self-esteem in court-ordered psychiatric evaluations. *Journal of Personality Assessment, 49,* 437–439.

*Singh, S. (1983). Rorschach hostility content and maternal child rearing attitudes. *Journal of Social Psychology, 120,* 149–150.

*Skolnick, N. J., & Zuckerman, M. (1979). Personality change in drug abusers: A comparison of therapeutic community and prison groups. *Journal of Consulting and Clinical Psychology, 47,* 768–770.

*Sloan, P., Arsenault, L., Hilsenroth, M., & Harvill, H. (1996). Assessment of noncombat, war-related posttraumatic stress symptomatology: Validity of the PK, PS, and IES scales. *Assessment, 3,* 37–41.

*Snyder, D. K., & Regts, J. M. (1990). Personality correlates of marital dissatisfaction: A comparison of psychiatric, maritally distressed, and nonclinic samples. *Journal of Sex and Marital Therapy, 16,* 34–43.

*Steiner, M., Martin, S., Wallace, J. E., & Goldman, S. (1984). Distinguishing subtypes within the borderline domain: A combined psycho-neuroendocrine approach. *Biological Psychiatry, 19,* 907–911.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

*Tegtmeyer, P. F., & Gordon, M. (1983). Interpretation of white space responses in children's Rorschach protocols. *Perceptual and Motor Skills, 57,* 611–616.

*Turner, J. A., & Romano, J. M. (1984). Self-report screening measures for depression in chronic pain patients. *Journal of Clinical Psychology, 40,* 909–913.

*Watson, R. A., & Pantle, M. L. (1993). Reflections on the Rorschach and the Millon Adolescent Personality Inventory. *Perceptual and Motor Skills, 77,* 1138.

*Watson, C. G., Plemel, D., Vassar, P., Manifold, V., Kucala, T., & Anderson, D. (1987). The comparative validities of six MMPI repression scales. *Journal of Clinical Psychology, 43,* 472–477.

Weiner, I. B. (1996). Some observations on the validity of the Rorschach Inkblot Method. *Psychological Assessment, 8,* 206–213.

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.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1997). The reliability of the comprehensive system for the Rorschach: A comment on Meyer (1997). *Psychological Assessment, 9,* 490–494.

*Yanovski, A., Menduke, H., & Albertson, M. G. (1995). Analysis by gender and visual imagery reactivity of conventional and imagery Rorschach. *Perceptual and Motor Skills, 80,* 1319–1340.

*Zimmerman, D. P., & Dillard, J. (1994). The Rorschach assessment of sexually abused children in residential treatment: A research note. *Residential Treatment for Children and Youth, 12,* 59–72.

Zubin, J., Eron, L. D., & Schumer, F. (1965). *An experimental approach to projective techniques.* New York: Wiley.