JOURNAL OF PERSONALITY ASSESSMENT, 1997, *69*(1), 1–38

The Ability of the Rorschach to Predict Subsequent Outcome: A Meta-Analysis of the Rorschach Prognostic Rating Scale

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage

Leonard Handler
Department of Psychology
University of Tennessee, Knoxville

To evaluate the ability of the Rorschach to predict subsequent outcome, the journal literature on the Rorschach Prognostic Rating Scale (RPRS) was reviewed and a meta-analysis was conducted on 20 statistics derived from a combined sample of 752. Using outcome criteria obtained an average of 1 year after initial testing, the uncorrected population correlation between RPRS scores and outcome was found to be $\rho = .443$ (95% confidence interval = .39 to .50). After making corrections to determine the validity of the RPRS when all subjects, all RPRS scores, and all outcome scores were included in the final statistical analysis, the estimated parameter increased to $\rho = .560$ (90% credibility value = .53). We are not aware of any other personality scale that uniformly demonstrates such high predictive validity. To flesh out the meaning of these results, they were placed in the context of other predictor–criterion relations drawn from various fields of study. The disparity between this strong evidence for validity and the chronic criticisms leveled against the Rorschach is discussed and suggestions are made for future research.

As with all personality measures, the predictive validity of the Rorschach is difficult to evaluate for two reasons. First, most research is not cumulative. There is very little replication from one study to the next, resulting in an extensive list of potential predictor–criterion relations to consider. More important, however, there has been relatively little theorizing about the Rorschach constructs that should predict various kinds of outcomes. As a consequence, the literature contains many exploratory studies that use a large array of variables and small samples to generate completely empirical predictive equations—equations for who will drop out of therapy, for who will improve in therapy, and so on. The problem with this approach is that these equations are very sample-dependent. Most often, they fail to work in subsequent studies. In combination, these factors make it difficult to summarize the general effectiveness of the Rorschach as a method for obtaining prognostic information.

There is at least one exception to this general state of affairs, however. In 1951 Bruno Klopfer introduced the Rorschach Prognostic Rating Scale (RPRS). Unlike many other prognostic indices, Klopfer used his clinical experience with the Rorschach to generate theoretical notions about the types of scores that should predict treatment outcome. Subsequently, these clinically derived expectations were refined through preliminary research. Fundamentally, Klopfer envisioned the RPRS to be a measure of ego strength that quantified "the concept of ego-strength in its most important components: reality testing, emotional integration, self-realization, and mastery of reality situations" (Klopfer, Kirkner, Wisham, & Baker, 1951, p. 425). Importantly, Klopfer believed that the RPRS tapped both the patient's "available ego

TABLE 1
Components of Klopfer's Rorschach Prognostic Rating Scale

Variable	*Component*
Human movement	Amount of movement in space
	Freedom in seeing movement
	Cultural distance
	Form quality of *M*
Animal movement	Amount of movement in space
	Freedom in seeing movement
	Cultural distance
	Form quality of *FM*
Inanimate movement	Natural and mechanical forces
	Abstract forces
	Form quality of *m*
Shading	
Texture	Form dominant versus form secondary versus formless/minus form quality
	Warm, soft, or transparent surface versus hard or cold surface versus shading as color versus shading in a diseased organ
Vista	Form dominant versus form secondary/formless versus minus form quality
Shading use problems	Shading evasion, shading insensitivity
Color	Form dominant versus form secondary versus formless/minus form quality
Color use problems	Color description/color denial/symbolic color (euphoric)/color comments versus forced or arbitrary use of color versus symbolic color(dysphoric)/color in a diseased organ versus color naming/color contamination
Form quality	Averaged across protocol[a]

[a] Klopfer's approach to scoring form quality is quite cumbersome and incorporates qualities that are coded as Developmental Quality and Cognitive Special Scores in the Comprehensive System.

strength" as well as his or her "potential ego strength." He felt the former was directly tied to patients' general adjustment status or diagnostic level of functioning. However, he conceptualized the latter component as the ego functions that may become mobilized during the course of psychotherapy. It was this element that he felt reflected therapeutic "promise" or the ability to make treatment gains, regardless of diagnosis or general functional capacity. Thus, as a measure of overt and tacit ego resources, Klopfer believed the RPRS should help identify those people who would profit most from treatment. Table 1 lists the major elements that comprise the RPRS (see Klopfer, Ainsworth, Klopfer, & Holt, 1954; Klopfer et al., 1951).

Klopfer's system for scoring the Rorschach is now infrequently taught (Hilsenroth & Handler, 1995) and the RPRS was not incorporated into Exner's (1993) Comprehensive System. Probably because of these factors, the RPRS has not been subjected to empirical validation since 1983. Nonetheless, for continuity with current scoring practices, it is worthwhile to note there is some overlap between components of the RPRS and components of ego functioning scales that have been recently derived from the Comprehensive System. For example, the criteria associated with Klopfer's human and animal movement variables make them similar to the object representational measures included in Perry and Viglione's (1991) Ego Impairment Index. Klopfer's shading and color variables are also similar to the form dominance variables included in Resnick and Meyer's (1995; Meyer & Resnick, 1996) Conceptual Ego Strength Index. Finally, Klopfer's criteria for scoring form quality take into account disruptions in logical thought processes, making this variable related to both the form quality scores and the cognitive special scores contained in the more recent ego strength scales.

Although it will be valuable to empirically document the association between the RPRS and these more recently developed scales of ego functioning, the RPRS has value for another reason. Not only was it developed with an underlying rationale for why it should predict outcome, but in the 25 years that followed its development it became the focus of a small body of research examining its ability to predict outcome.[1] Thus, the RPRS is an optimal scale to evaluate with meta-analytic procedures in order to assess the prognostic utility of the Rorschach method.

[1] There have been relatively few studies exploring other aspects of Rorschach Prognostic Rating Scale (RPRS) validity. Seidel (1960) found the RPRS had a moderate association with case history ratings of premorbid adjustment ($r = .30$), whereas Williams, Monder, and Rychlack (1967) found the RPRS had strong correlations with psychiatrist-generated ratings of prognosis ($M\ r = .547$) but not with social worker-generated ratings ($M\ r = .05$). In a small sample of prisoners, Edinger and Bogan (1976) found the RPRS had correlations in the expected direction with concurrent ratings of group therapy adjustment ($r = .31$) and work adjustment ($r = .37$), but not with dormitory adjustment ($r = -.23$). Several studies found no association with the Minnesota Multiphasic Personality Inventory Ego Strength Scale (e.g., Endicott & Endicott, 1964; Newmark et al., 1974) or measured intelligence (e.g., Mindess, 1957; Williams et al., 1967). Lundin and Schpoont (1953) presented a case study where RPRS scores obtained from repeated assessments were conceptually linked to aspects of the therapy.

Meta-analysis has several goals. Its primary goal is to summarize a body of research in a cohesive fashion to derive an estimate of the relation between two variables. For our purposes, this relationship, known as the *effect size,* quantifies the extent of association between RPRS scores and subsequent outcome. The effect sizes from each primary study are aggregated across studies in order to generate an estimate of the effect size in the population. Once an overall effect size has been computed, three additional questions become pertinent. First, is the effect size constant across studies? That is, even though findings may differ from one study to another, do all studies appear to be targeting the same underlying statistical parameter? If so, then confidence can be placed in the summary effect size. Second, is the magnitude of the overall effect statistically significant? That is, does the average effect size differ from zero—the effect that would be expected if there were no real relationship between the RPRS and outcome? The third and final question is whether there are substantive moderating variables that cause the effect size to vary from one study to the next. This question is particularly important to address when it seems likely that all the primary studies are not targeting the same statistical parameter. For our purposes, substantive moderators could include certain types of patients for whom the RPRS is more or less effective, certain types of treatment for which the RPRS is more or less effective, certain types of outcome which are easier to predict, and so on.

In this article we address each of the preceding questions by focusing on studies that have used the RPRS in longitudinal research to predict subsequent outcome. Previous narrative reviews of the RPRS have often been positive (Garwood, 1977; Goldfried, Stricker, & Weiner, 1971), although this has not been a universal sentiment, as other reviewers have concluded the RPRS has questionable utility (Frank, 1967; Garfield, 1994; Shields, 1978). However, narrative reviews have generally relied on "vote counting" procedures (Rosenthal, 1991), where the number of studies with and without statistically significant results are tallied to reach a general conclusion about test validity. There are many dangers associated with such a procedure (Hunter & Schmidt, 1990; Rosenthal, 1991; Schmidt, 1996), the most obvious being that small samples will result in many Type II statistical errors (i.e., the test will be seen as invalid when actually it is not). Meta-analytic procedures rectify this problem and allow for more decisive conclusions to be drawn about RPRS validity.

METHOD

Literature Search

The studies identified for this analysis were obtained as part of a larger, ongoing meta-analysis designed to evaluate the prognostic utility of personality tests.[2] As

[2] This meta-analysis is sponsored jointly by the Society for Personality Assessment and Rorschach Workshops and it is supported by the University of Alaska, Anchorage and the University of Tennessee, Knoxville.

part of the larger project, a broad and extensive search of longitudinal personality and psychotherapy research was conducted using the PsychLit database covering the years 1974–1995. This search identified slightly more than 13,000 potentially relevant studies. Several of these studies employed the RPRS. Additional citations were then obtained by working backwards from these sources of information and by consulting other relevant review articles (e.g., Baekeland & Lundwall, 1975; Frank, 1967, 1993; Fulkerson & Barry, 1961; Garfield, 1971, 1978, 1986, 1994; Garwood, 1977; Goldfried et al., 1971; Luborsky, Chandler, Auerbach, Cohen, & Bachrach, 1971; Windle, 1952; Zubin & Windle, 1954). In total, we found 22 investigations that used the RPRS as a baseline measure to predict longitudinal outcome.[3]

Four of the 22 studies were excluded from consideration. Sheehan and Tanaka (1983) conducted a logistic regression analysis with the RPRS and its subcomponents. Although significant findings were obtained, the authors only reported statistics related to the combined logistic model. They did not report the univariate statistics needed to aggregate data across studies. A second study was excluded because it was a dissertation. Rockberger (1953; cited in Goldfried et al., 1971) observed a robust association ($r = .57$; $N = 36$) between RPRS scores and therapists' ratings of improvement. Although this is the only dissertation we are aware of, we made no effort to systematically review unpublished dissertations. Thus, this study was excluded and the review was limited to data published in English-language journals. Another study (Whiteley & Blaine, 1967) was excluded because the authors omitted the RPRS color and form quality variables, using only the movement and shading variables as independent predictors. The final study (Cooper, Adams, & Gibby, 1969) was also excluded because the authors did not use the complete RPRS. Instead, they utilized Cartwright's "short-form" of the RPRS, which employs human movement, color, and form quality, but omits shading and the other kinds of movement variables.

[3] Several reviews have erroneously considered Lessing's (1960) study to be an investigation into the Rorschach Prognostic Rating Scale (RPRS; Frank, 1967; Garfield, 1971, 1978, 1986, 1994). It is not. The problem of this perpetuated error is compounded because this study has been cited as one of the few that did not find evidence supporting the validity of the RPRS. However, not only did Lessing never use the RPRS, but she did not even use the Klopfer system of scoring. Instead, Lessing had psychologists make ratings of predicted therapy success after considering a social history, extensive referral information, and a battery of psychological tests, which always included a Rorschach protocol scored by the Beck system. Frank (1993) has also erroneously reported that Williams et al. (1967) conducted a prognostic treatment outcome study. They did not. The design was one of concurrent validity (as indicated in the title), even though data collection spanned a 1-year period. Finally, Frank (1993) also erred with respect to a study by Davids and Talmadge (1964). Although the latter found strong associations between pretreatment Rorschach data and eventual outcome, they did not use the RPRS. Instead, Davids and Talmadge made use of a number of other Rorschach scores.

The 18 remaining studies contained data from 20 independent samples. In addition, two studies used the RPRS to predict two distinct outcome criteria. Mindess (1953) used the RPRS to predict patients' level of functioning following treatment, and also presented data to predict change that occurred as a function of treatment. Sheehan, Frederick, Rosevear, and Spiegelman (1954) used the RPRS to predict improvement as a function of treatment, as well as to predict attrition from treatment. Although the criteria in these two studies share subjects and are not statistically independent, the coefficients were treated separately because the criteria in each case were conceptually distinct. Thus, the initial meta-analysis was conducted on 22 individual statistics.[4]

Table 2 provides some of the pertinent information for each of the 22 statistics included in the meta-analysis. The table includes the citation, the type of sample used in the research, the average number of days between baseline testing and outcome assessment, the outcome variable, the source(s) of the outcome information, whether outcome ratings were made while blind to Rorschach data, the number of independent sampling units in the study (generally, the number of subjects, unless a paired comparison design was used), and the magnitude of the relation between the predictor and the outcome criterion expressed as the correlation coefficient (r).

Aggregation Procedures

If data were available for more than one follow-up period, only information from the longest follow-up was used. If outcome criteria were hierarchically organized into a composite score and its subcomponents, the effect size was derived from the composite criterion. When a study utilized more than one dependent variable (DV) to assess the same or a similar construct (e.g., clinical improvement), all the data for that study were aggregated into a single effect size estimate. In this context, raw or residual gain scores were treated as equivalent to rated benefit scores.

Special note should be made about several studies. Fiske, Cartwright, and Kirtner (1964) presented initial outcome data from the Chicago Counseling Center Project. Subsequently, these data were reanalyzed by Luborsky, Mintz, and Christoph (1979)

[4] Violating the assumption of independence by using two effect sizes from the same study does not have any systematic effect on measures of central tendency in a meta-analysis. However, it does affect estimates of sampling error variance, particularly when a large number of nonindependent effects are derived from a large proportion of the studies contributing to the meta-analysis (which is not the case here). Specifically, violating the assumption of independence results in an underestimate of sampling error. Because sampling error is subtracted from the observed variance in effect sizes to estimate true effect size variance, an underestimate of sampling error results in an overestimate of true effect size variance (see Hunter & Schmidt, 1990, p. 451–454). Thus, to the extent that using more than one statistic from a single study may generate problems, the problems result in conservative decisions regarding effect size homogeneity.

TABLE 2
General Study Information

Study	*Sample Type*	*Mean Follow-Up*	*Outcome Variable*	*Outcome Source*	*Blind Ratings*	*n*	*r*
1. Bloom, 1956	Adult OP		Therapy improvement	R & T	?	25	.46
2. Bloom, 1956	Adult OP		Therapy improvement	R & T	?	21	–.00
3. Brawer & Cohen, 1966	Beginning teachers	365	Adjustment as a teacher	Supervisor	Yes	20	.37
4. Cartwright, 1958	Adult OP	182	Therapy improvement	T	?	13	.51
5. Endicott & Endicott, 1964	Adult OP	183	Therapy improvement	T	Yes	21	.43
6. Endicott & Endicott, 1964	Adult OP	194	Wait list improvement	T	Yes	40	.38
7. Filmer-Bennett, 1955	Adult IP	730	Improved functioning	R	Yes	11	.36
8. Fiske et al., 1964[a]	Adult OP	182	Therapy improvement	Pt & T & R	?	40	–.14
9. Johnson, 1953	Child residential	730	Therapy improvement	T & Teacher	?	21	.53
10. Kirkner et al., 1953	Adult OP		Therapy improvement	R & T	?	40	.58
11. Luborsky et al., 1979[a]	Adult OP	287	Therapy improvement	Pt & T & R	?	42	.16
12. Mindess, 1953	Adult MP, 1/2 CR	183	Level of functioning	T	Yes	80	.81
13. Mindess, 1953	Adult MP, 1/2 CR	183	Therapy improvement	T	Yes	79	.52
14. Mindess, 1957	Student nurses	547	Success in training	Exams & Supervisor	?	68	.41
15. ~~Newmark et al., 1973~~	~~Adult OP~~	~~128~~	~~Therapy improvement~~	~~2 of 3 (Pt, T, R)~~	?	27	.55
16. Newmark et al., 1974	Adult OP	215	Therapy improvement	2 of 3 (Pt, T, R)	?	26	.48
17. Newmark et al., 1979	Adult IP	267	Remitted thought disorder	1 of 3 (Pt, R, R)	?	74	.38
18. Novick, 1962[a]	Child OP	140	Therapy improvement	Staff & Parent	?	22	.42
19. Schulman, 1963	Adult OP	196	Therapy improvement	T	Yes	20	.29
20. Seidel, 1960	Adult IP	1095	Recovered or still in hospital	T	Yes	63	.40
21. Sheehan et al., 1954	Adult OP		Leave versus stay in treatment	T	?	35	.45
22. Sheehan et al., 1954	Adult OP		Therapy improvement	T	?	35	.45

Note. OP = outpatient psychiatric; IP = inpatient psychiatric; MP = mixed in and outpatient psychiatric population; CR = court referred; R = researcher's perspective; T = therapist's perspective; Pt = patient's perspective.

[a] Relevant information appeared in reports other than the primary citation (see Cartwright et al., 1963; Goldfried et al., 1971; Luborsky et al., 1980; Mintz et al., 1979).

after they generated alternative DVs that paralleled the outcome measures used in the Penn Psychotherapy Project. Because slightly different outcome data were available in two publications, the final effect size for the Chicago Project was calculated as the average of the coefficients derived from both analyses. Second, although the authors did not indicate it, some of the data presented in Newmark, Finkelstein, and Frerking (1974) duplicated data that were presented by Newmark, Hetzel, Walker, Holstein, and Finkelstein (1973). Specifically, Newmark et al.'s (1973) data are the same as the "Group A" data in their later publication (Newmark et al., 1974). Consequently, Group A was excluded from Newmark et al. (1974) and the data for this citation were limited to Group B. Third, Endicott and Endicott (1964) presented data on both a treatment and a wait-list control group. They have a separate publication (Endicott & Endicott, 1963) that only discusses their control group. The latter article was not used in the meta-analysis. Fourth, Novick's (1962) findings were published as an abstract, which led us to debate whether it was appropriate to include it in the analysis. Ultimately, because the abstract presented the basic information we required (supplemented by Goldfried et al., 1971), we decided it should be included. However, Novick only reported that his findings were statistically significant. By necessity, we assumed that his results just reached the .05 level of significance. Using the formula presented by Rosenthal (1991, Equation 2.18), the standard normal deviate for this probability level was transformed to a Pearson correlation. Finally, in one of Bloom's (1956) samples, a nonsignificant finding was obtained in the direction opposite of that predicted. Because this study did not contain enough information to calculate a point estimate of the effect size (group means of 2.2 and 2.3, but no standard deviations), it was assigned a value of $r = -.00$.

Eight studies presented raw data for each subject. In seven of these studies (Cartwright, 1958; Johnson, 1953; Kirkner, Wisham, & Giedt, 1953; Mindess, 1953; Newmark et al., 1974; Newmark et al., 1973; Schulman, 1963), the raw data were used to calculate correlations.[5] The data in the remaining study (Filmer-Bennett, 1955) were used to generate a matched-pairs t test, with the t statistic then transformed to the Pearson correlation (Rosenthal, 1991).

A primary concern in a meta-analysis is deciding on the most accurate way to quantify effect sizes. Although we selected the correlation as our effect size metric

[5] In several instances, this resulted in slightly different coefficients being used for the meta-analysis than those reported in the initial studies. The value used for Cartwright (1958) was .51 rather than tau = .52, the value for Kirkner et al. (1953) was .58 rather than phi = .67, the value for Nemark et al. (1973) was .55 rather than r_{pb} = .41, and the value for Schulman (1963) was .29 rather than rho = .32. Also, the table in Mindess (1953) only presented data on 79 of his 80 participants. Because the computed correlation between Rorschach Prognostic Rating Scale (RPRS) scores and functioning was the same as that reported in the initial article, N was left at 80 for this coefficient. However, the initial article did not calculate a correlation between RPRS scores and therapy improvement. Consequently, this coefficient was based on $N = 79$.

(such that other statistics in the literature were translated into correlations), there is some disagreement among meta-analysts about whether the raw correlations are the best statistics to use when combining information across studies. Although Hunter and Schmidt (1990) disagree, both Rosenthal (1991) and Hedges and Olkin (1985) believe it is best to transform the raw correlations prior to aggregation in order to correct for the nonlinear downward bias that is found in averaged correlations. Johnson, Mullen, and Salas (1995) indicated that Hunter and Schmidt's failure to transform correlations results in a slight but consistent underestimate of mean effect sizes, particularly when raw correlations are large or variable across studies. Hunter and Schmidt acknowledge the bias that is present in correlations and acknowledge that the aggregation of raw correlations produces a slight underestimate of true correlations. However, they maintain that the alternative, which is to transform *r* to Fisher's *Z* prior to aggregation, actually introduces an upward bias in the final correlation. They maintain that the upward bias caused by Fisher's transformation is actually larger than the downward bias found when the coefficients are untransformed. Thus, they are opposed to transforming raw correlations prior to aggregating them across studies. For this study, we adopted a conservative approach and did not use Fisher's transformation procedure. We recognized this would cause an underestimate of the true effect size, particularly if the validity coefficients for the RPRS were large or variable across studies.

Identifying and Correcting Statistical Artifacts

Hunter and Schmidt (1990) identified 11 methodological and statistical sources of bias that, when unchecked, will cause errors in meta-analytic data. These artifacts will cause mean effect size estimates to be erroneously high or low and/or will make it seem like effect sizes vary from study to study when in actuality they do not. Study artifacts include: sampling error, random measurement error in the independent variable (IV) or DV, false dichotomization of a continuous IV or DV, range variation in the IV or DV, deviation from perfect construct validity in the IV or DV, reporting or transcription error, and variance due to extraneous factors. Not all of these artifacts can be corrected, particularly the last two. Nonetheless, Hunter and Schmidt vigorously encourage meta-analytic researchers to correct as many of these sources of bias as possible. In the present analysis, each primary study was coded in a manner that would allow us to identify and potentially correct the first seven artifacts. Each will be discussed in turn.

Sampling error. Statistical sampling error causes observed correlations to differ from one another even though the true correlation in the population remains fixed. In other words, even though there is only one population parameter for the effect size, individual studies will observe effects of different magnitude simply because each study is a sample estimate drawn from the overall population. Variation from study to study will be greatest when samples are small and when the true effect size is small. Hunter and Schmidt (1990) believe statistical sampling error is the largest single source of bias affecting research. They convincingly demonstrated how the failure to remove this artifactual source of variance from the observed effect size variance will make it appear like studies disagree with each other when in fact they do not. Sampling error is corrected by statistical formulas that take into account the size of the sample and the size of the effect in each of the primary studies.

Range variation. When the variance in an IV or DV is unusually large (e.g., the sample only includes patients rated as the "most" and "least" improved and omits those in the middle portion of the continuum), correlations will be artificially inflated. Alternatively, if IV or DV variance is attenuated (e.g., only people with scores above a certain cutoff on the IV are included; the DV of success in therapy is only calculated on people who complete a set number of sessions, not on those who drop out; etc.), correlations will be artificially deflated. The extent of bias in a correlation can be identified when a reference variance is available for the IV and DV. Correlations obtained from samples that are more variable than the reference standard are larger than they should be and must be reduced in order to be accurate. Correlations obtained from samples that are less variable than the reference standard are smaller than they should be and must be increased in order to be accurate. Although it is easiest to make a determination about range variation when sound normative information is available for each IV and DV, Hunter and Schmidt (1990) also provided formulas to estimate range variation by using the proportion of subjects excluded from an analysis and the normal distribution as the reference standard. Although this approach can be quite effective, it does not identify all instances of range variation (e.g., that which may be a function of excessive subject heterogeneity or homogeneity).

Table 3 provides information related to statistical artifacts. The first two columns present information on range variation in the IV. To assess IV range variation, we recorded the mean and standard deviation of the RPRS in each study. Theoretically, the RPRS is a 30 point scale, ranging from –12 to 17. Given this, it is surprising that the standard deviations generally fall in a rather narrow range, between 2 and 4. This suggests that each sample may have had rather homogeneous scores about its mean. To assess whether sample variance was restricted in each of these studies it would be optimal to have normative information on RPRS variance. Unfortunately, sound normative data are not available. Barring this, however, the central limit theorem can be used to estimate expected sample variance.

Specifically, we used the central limit theorem to generate two estimates of the same variance. Because the variance of the sampling distribution of means (σ^2_M) is

TABLE 3
Artifact Information for Each Study

	Range Variation				*False Dependant Variable Dichotomization*		*Reliability*						
	RPRS		*Outcome*				*RPRS*		*Outcome*				
Study	*M*	*SD*	*a/e*	*wgt*	*PS*	*a Factor and wgt*	*rxx′*	*wgt*	*rxx′*	*wgt*	*Compound wgt*	*n*	*r*
1. Bloom, 1956	3.18	4.20	1.19[a]	.84	.52	.80		.5		1.0	.34	25	.46
2. Bloom, 1956	2.25		1.24[a]	.81	.52	.80		.5		1.0	.32	21	–.00
3. Brawer & Cohen, 1966	7.48		1.00	1.00	—	1.00		.5		.5	.25	20	.37
4. Cartwright, 1958	6.25	2.77	1.00	1.00	.62	.78		.5		.5	.20	13	.51
5. Endicott & Endicott, 1964	5.44	2.44	.83	.83	.52	.80	.86	1.0		.5	.33	21	.43
6. Endicott & Endicott, 1964	4.74	2.30	1.00	1.00	.40	.79	.86	1.0		.5	.40	40	.38
7. Filmer-Bennett, 1955	1.66	4.21	1.21[a]	.83	.50	.80		.5		.5	.17	11	.36
8. Fiske et al., 1964			.84	.84	—	1.00		.5		1.0	.42	40	–.14
9. Johnson, 1953	1.75	2.60	1.00	1.00	.71	.75		.5		1.0	.38	21	.53
10. Kirkner et al., 1953	5.67	2.95	1.00	1.00	.65	.78		1.0		1.0	.78	40	.58
11. Luborsky et al., 1979			.77	.77	—	1.00		.5	.75	1.0	.39	42	.16
12. Mindess, 1953	3.91	3.88	1.00	1.00	—	1.00		.5		.5	.25	80	.81
13. Mindess, 1953	3.91	3.88	1.00	1.00	—	1.00		.5		.5	.25	79	.52
14. Mindess, 1957	6.25	2.58	.84	.84	—	1.00		.5		1.0	.42	68	.41
15. Newmark et al., 1973	6.22	2.65	.86	.86	.63	.78	.83	1.0		1.0	.67	27	.55
16. Newmark et al., 1974	6.85	2.32	.80	.80	.69	.76	.75	1.0		1.0	.61	26	.48
17. Newmark et al., 1979	2.37	1.77	1.00	1.00	.58	.79	.80	1.0	.83	1.0	.79	74	.38
18. Novick, 1962	1.10		1.00	1.00	—	1.00		.5		1.0	.40[b]	22	.42
19. Schulman, 1963	3.78	1.64	1.00	1.00	—	1.00		1.0		.5	.50	20	.29
20. Seidel, 1960			1.20	.83	.51	.80		.5		.5	.17	63	.40
21. Sheehan et al., 1954	6.07	2.65	1.00	1.00	—	1.00	.80	1.0		.5	.50	35	.45
22. Sheehan et al., 1954	6.07	2.65	1.00	1.00	.54	.80	.80	1.0		.5	.40	35	.45

Note. RPRS = Rorschach Prognostic Rating Scale; a/e = attenuation/enhancement factor; wgt = weight assigned to study for each artifact factor; PS = proportion of patients assigned to the falsely dichotomized "success" category; a Factor = attenuation factor; rxx′ = reliability coefficient.

estimated by the observed sample variance divided by sample size (σ^2_O / n), the observed sample variances can be used to estimate the variance of the sampling distribution of means. In addition, the observed RPRS means can be used to calculate an independent estimate of the sampling distribution variance. If the variance of the sampling distribution that is generated from sample means is much larger than the estimate of the same variance generated from observed scale variances, it would suggest that the sample variances found in Table 3 are actually smaller than would typically be expected. Using this procedure assumes that each sample is randomly selected and is targeting a common population mean. Because of these assumptions, only the nine independent studies that used adult outpatient samples and provided both mean and variance data on the RPRS were used for this analysis.

The variance of the sampling distribution estimated by the means of the nine studies was found to be 1.497. The average variance of the sampling distribution estimated by dividing each observed variance by its sample size was .303. This ratio, 1.497:.303, indicates the sampling variance estimated by observed means is 4.94 times larger than the sampling variance estimated by observed scale variance, which is a statistically significant difference, $F(8, 26) = 4.94$; $p = .00086$, with *df* for the denominator estimated by the average $n - 1$. Another way to consider the same relation is to calculate an estimate of the average sample standard deviation that would generate a sampling distribution variance of 1.497 (for studies with an average sample size of 27). Using the formula $\sigma^2_M = \sigma^2_O / n$, rearrangement shows that the average sample standard deviation ought to be about 6.36, rather than the observed values between 2 and 4.

Thus, it appears that the samples in the meta-analysis may have restricted variance on the RPRS. It is not clear why this would be the case. However, assuming the preceding analysis is founded on correct assumptions, the findings suggest that the observed validity coefficients for the RPRS are underestimates of the typical correlations that would be found with the scale. As a result, they could legitimately be enlarged using the Hunter and Schmidt (1990) formulas. However, a more conservative strategy was adopted for this analysis. Rather than correcting for range restriction in the RPRS and increasing the magnitude of the validity coefficients, this potential source of bias was simply ignored in subsequent calculations.

To quantify range variation in the DV, we assumed range restriction was operative whenever subjects from the initial sample were excluded from the final analysis because they did not complete the intervention. We assumed that these participants would have been rated as least successful had they been included in the final analysis (i.e., range restriction due to attrition). As indicated in the third column of Table 3, six studies suffered from range restriction (indicated by a/e values less than 1.0). The authors of five studies deleted those subjects who failed to continue treatment for a minimum number of sessions, thereby excluding those who theoretically should have improved the least. The author of the remaining study omitted student nurses who dropped out of training, thereby excluding those who should have been rated the least successful. Each of these studies provided exact information about the number of subjects excluded from their analysis and, therefore, exact corrections for attrition could be made.

We assumed range enhancement was present whenever patients scoring in the middle section of the DV were excluded from the final analysis. Using formulas presented by Hunter and Schmidt (1990), range variation on the DV was then estimated using the normal distribution as the reference standard. As indicated in Table 3, four studies suffered from range enhancement (a/e values greater than 1.0). In these studies the final analysis was limited to participants drawn from the ends of the outcome distribution. Only Seidel (1960) provided sufficient information to determine an exact correction for range enhancement. This study excluded 37% of the subjects from the center portion of the outcome distribution.

When a meta-analysis is conducted on some studies that have range attenuation and other studies that have range enhancement, it is best to correct for range variation at the level of individual studies rather than rely on artifact distributions to make the corrections. The reason for this is simply because opposing sources of bias will tend to cancel each other within the artifact distributions. Because it is best to correct the individual studies for range variation under these circumstances, we estimated range enhancement for the remaining three studies. Although we do not have data attesting to this point, we suspect most studies would not exclude more than 35% to 40% of the subjects falling in the middle region of the outcome distribution. Therefore, the extent of range enhancement found in Seidel's study probably reflects a fairly generous assumption about the proportion of subjects that would be excluded from a typical analysis. Given this, the three other studies suffering from range enhancement were estimated to have excluded the middle 37% of their outcome distribution. Using Seidel's value to correct for range enhancement is likely to be a conservative procedure that is more prone to over-correct (i.e., excessively reduce) than under-correct correlations.

Note that range variation in the DV is only partially corrected by the aforementioned procedures. The DV could still have an unusually large or small degree of variance. Unfortunately, it is impossible to correct for this additional form of range variation without normative information for each DV.

False dichotomization. Studies that falsely dichotomize a continuous IV or DV create validity coefficients that are attenuated. For instance, if the continuum "extent of change in psychotherapy" is treated as a dichotomous outcome variable consisting of either "successful" or "unsuccessful" change, the validity coefficient generated from the dichotomous outcome variable will be smaller than it would have been had the continuous outcome variable been used. To correct for this source of bias, the proportion of subjects in each of the dichotomous groups must be identified.

Thirteen of the studies artificially dichotomized a continuous or multi-step outcome variable (no corrections were applied to genuine DV dichotomies). Table 3 lists the proportion of people classified as "successful" or "improved" on the outcome variable, as well as the attenuation factor (and equivalent attenuation weight) that results from such a split. Only one study split their sample on the RPRS (Novick, 1962). This sample was split at the RPRS median, forming two groups of equal size. Because this was the only study that dichotomized the RPRS, this information was not presented in Table 3. However, a median split results in an attenuation factor (and weight) of .80. Because specific information was available from each study for this artifact, each study was individually corrected for false dichotomization.

Measurement error. Random measurement error impacts the magnitude of effect sizes as well as their variance across studies. To the extent that measurement error is present, validity coefficients are smaller than they should be. In addition, to the extent that there is more measurement error in some studies and less in others, validity coefficients will fluctuate across studies. Because it is impossible to measure any psychological construct without measurement error, validity coefficients that are corrected for measurement unreliability do not reflect phenomena as they are actually quantified. Rather, these corrected coefficients provide information about the theoretically "true" relation between variables. For the sake of completeness (see Schmidt & Hunter, 1996), the following analysis will present validity coefficients corrected for measurement error. However, little emphasis will be placed on these values because our primary interest is in the clinical application of the RPRS. In clinical applications, RPRS scores always contain random measurement error.

As can be seen in Table 3, reliability information for the RPRS and the outcome variables were reported much more sporadically than the other artifacts. By necessity then, artifact distributions were used to correct for these sources of bias (Hunter & Schmidt, 1990).

Study Weights

Assuming that salient study characteristics are not correlated with sample size, all meta-analytic approaches recommend that the effect obtained from each primary study be weighted by sample size to ensure that the most stable values are given the greatest emphasis. In addition, Hunter and Schmidt (1990) recommended weighting individual effect sizes by the amount of bias they are likely to contain, with lower weights given to those studies that contain greater bias. Specifically, they advocate for each study to be weighted by the product of sample size and the squared "compound attenuation factor." The compound attenuation factor is the product of all the individual sources of bias affecting a study correlation. Although each source of bias either falsely attenuates or enhances validity coefficients, to calculate study weights the extent of each kind of bias must be expressed as an attenuation weight (a value less than 1.0 if the bias is present and a value of 1.0 if it is not) to ensure that studies with enhancing biases are not given undue emphasis. Thus, if range enhancement in a study causes an observed correlation to be 1.3 times larger than it should be, the attenuation weight must be the inverse of this value, .77, to accurately weight the degree of error in the obtained correlation. Of course, the attenuation factor itself would remain at 1.3 because the observed correlation must still be divided by this amount to correct enhancement bias (i.e., reduce the correlation) and accurately estimate the population parameter.

As indicated in Table 3, each study was assigned weights according to the extent of artifact bias within the study. Because the reliability of any variable is a precursor to validity, and because IV and DV reliability data were not available for many studies, an alternative convention was adopted for assigning reliability weights to each study. Typically, each study would be assigned an IV reliability weight that is equal to the square root of the IV reliability in that study; DV weights would be assigned in a similar fashion. In the absence of reliability data, we decided to assign a weight of .50 to the IV or DV in each study when no reliability was reported or when ratings were not determined by more than one person. Alternatively, when reliabilities were reported or when ratings were determined by more than one person, the variable was assigned a unit weight. We are not aware of other studies employing such a weighting scheme. However, this approach reflects our rough judgment that half as much confidence should be placed in potentially unreliable data. As indicated by the reliability weights shown in Table 3, a number of studies used methods that should have improved reliability, even though specific reliability figures were not calculated or reported.

Finally, Table 3 provides the compound attenuation weight assigned to each study. This value is the product of the attenuation weights assigned for range variation in the outcome variable, false dichotomization of the IVs and DVs, reliability of the RPRS, and reliability of the outcome variable. Studies that contain more psychometric error have weights that approach zero (e.g., Studies 7 and 20), whereas studies with less error have weights that approach one (e.g., Studies 10 and 17).

RESULTS

Preliminary Meta-Analysis of All Studies

Table 4 provides a statistical summary of the data drawn from all the RPRS studies. In terms of general study characteristics, it can be seen that the effect size calculations were carried out on 22 statistics (*K*) generated from a total of 823 participants (*N*). On average, each study contained about 37 subjects (range = 11 to 80). Outcome status was determined an average of about 11 months (342 days)

TABLE 4
Summary Variables for the Rorschach Prognostic Rating Scale Meta-Analysis: All Studies

Variable	*Uncorrected*	*Corrected–1*[a]	*Corrected–2*[b]
General study characteristics			
K (# of samples)	22.0		
N (# of participants)	823.0		
M n per study	37.4		
Mdn n per study	31.0		
M follow-up interval (days)	341.6		
Mdn follow-up interval (days)	196.0		
Percentage of studies contributing data	77.3		
Central tendency			
Unweighted *M* effect size	.400	.484	
Unweighted *Mdn* effect size	.424	.488	
Weighted *M* effect size (ρ)	.418	.528	.658
Variability			
Unweighted *SD r*	.1975	.2490	
Weighted *SD r*	.1700	.2247	
Weighted *SD* error	.1327	.1619	
Weighted *SD* ρ	.1062	.1558	.1924
90% credibility value	.282	.329	.412
Standard error (for heterogeneous studies)	.040		
95% confidence interval about weighted *M* effect size	.339 – .497		
Maximum effect size	.810	.823	
Q_3 (75th percentile)	.511	.668	
Q_2 (50th percentile)	.424	.488	
Q_1 (25th percentile)	.367	.371	
Minimum effect size	–.140	–.167	
Normal-based *SD* $(.75[Q_3 - Q_1])$	.108	.223	
Do all studies estimate the same parameter?			
Heterogeneity of effect (chi-square $df = 21$)	45.489		
Probability effect is not constant	.0015		

[a]Corrected for dependent variable range variation and false independent and dependent variable dichotomy. [b]Corrected for the preceeding and for rater unreliability.

after baseline testing with the Rorschach, although this distribution was positively skewed and about 23% of the studies did not provide this information.

The columns reflect increasing levels of psychometric correction. The first column presents statistics on the raw coefficients, uncorrected for any type of artifact. The second column presents statistics on the coefficients after they have been individually corrected for range variation in the DV and artificial dichotomization of the IVs and DVs. The third column presents statistics on the coefficients after artifact distributions were used to further correct for unreliability (interrater) in the predictor and criterion measures. The information in this column indicates the expected validity parameter once DV range variation, IV false dichotomization, DV false dichotomization, IV rater unreliability, and DV rater unreliability are corrected. Range variation (restriction) in the IV and other forms of unreliability (e.g., instability) remain uncorrected in these figures.

All three columns also contain a correction for sampling error. This correction does not change the magnitude or dispersion of the validity coefficients. Instead, it estimates the amount of variance in the correlations at each level of correction that results from using samples to measure a parameter. Because sampling error is the only form of error considered in the first column of data, the value for "Weighted *SD* error" in the first column indicates the extent of sampling error in the raw correlations. At each level of correction, a refined estimate of true variation among correlations (Weighted *SD r*) is obtained, either directly by formula (column 3) or by subtracting weighted error variance from the observed variance among correlations (Weighted *SD* r^2).

The central tendency section of the table lists the unweighted mean and median effect size, as well as the weighted mean effect size. All of these values are quite substantial. Focusing on just the first column of uncorrected data, the weighted mean effect size of $r = .418$ indicates the RPRS has a strong association with subsequent outcome across studies. The 90% credibility value, located in the variability section of the table, indicates the point above which fall 90% of the estimated true validity coefficients. Thus, for the uncorrected coefficients, 90% of the true validity coefficients are estimated to be above a value of .282. Given that this value is well above zero, it can be concluded that RPRS validity should readily generalize to new situations and settings.

However, there is also considerable variance in the observed effect sizes. The variance section of the table indicates the uncorrected effect sizes have a fairly narrow interquartile range of .144 (derived as Q3 – Q1), with a standard deviation based on this range of .108 (normal-based standard deviation). However, the general standard deviation is nearly twice as large (Unweighted *SD* $r = .198$). This pattern suggests there may be statistical outliers within the data. Indeed, even after correcting for sampling error, there is considerable variance remaining among the uncorrected correlations (weighted *SD* $r = .106$).

This variability can be seen readily in Figure 1, which is a boxplot of the uncorrected effect sizes. Boxplots have several pertinent characteristics. First, the center band indicates the median value of the distribution ($r = .424$). The box itself indicates the interquartile range, denoting where the central 50% of the validity coefficients fall. The top edge of the box indicates the 75th percentile ($r = .511$), whereas the bottom edge of the box indicates the 25th percentile ($r = .367$). The "Ts" or "whiskers" that are drawn from the top and the bottom of the box extend to the maximum and minimum effect sizes, excluding statistical outliers. Outliers falling 1.5 to 3 interquartile ranges from the box edge are plotted as squares, whereas one extreme outlier falling more than 3 interquartile ranges from the box edge is plotted as a triangle.

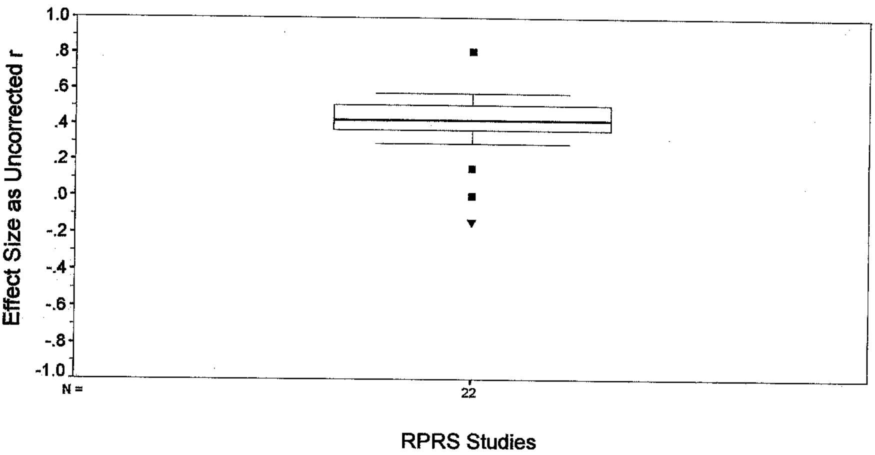

FIGURE 1 The distribution of uncorrected Rorschach Prognostic Rating Scale effect sizes across studies.

Using the normal-based standard deviation, a measure of typical deviation from central tendency that is less sensitive to extreme values, it can be determined that one outlier falls 3.6 *SD* units above the median, one coefficient falls 2.4 *SD* units below the median, another falls 3.9 *SD* units below the median, and a final coefficient falls 5.2 *SD* units below the median. Clearly there are several anomalous statistics within the data. The variability among the validity coefficients can be tested with a chi-square statistic that addresses the null hypothesis of no variability across studies (Hunter & Schmidt, 1990, pp. 111–112). With 21 degrees of freedom, this test is underpowered and capable of detecting only large effects. Given that the chi-square value of 45.5 is statistically significant (p = .0015), it is clear the coefficients derived from these studies are not targeting the same underlying parameter.

Although the most extreme coefficients could be deleted from the analysis on psychometric grounds (Huffcutt & Arthur, 1995[6]; Hunter & Schmidt, 1990), it

[6] Huffcutt and Arthur (1995) proposed a new statistic for detecting outliers in a meta-analysis. One problem with the traditional boxplot approach is that each study is treated as if it were analogous to a single subject, which means that all studies receive the same weight in the final calculations. The procedure proposed by Huffcutt and Arthur is designed to circumvent this problem. However, their procedure involves removing effect sizes one-by-one and iteratively calculating a weighted estimate of variance for the remaining distribution of effect sizes. The problem with this approach emerges when there is more than one outlier in the data. Under such conditions, even though one outlier may be removed from the variance calculation, the remaining outliers continue to artificially inflate the variance of the effect size distribution. Thus, although there are disadvantages to identifying outliers with the traditional unit-weighted approach, using the normal-based *SD* from this approach still provides a more accurate picture of atypical deviation when there is more than one outlier.

would be optimal to have some rationale that would account for their deviance. The likely reason for two of these outliers is apparent when study characteristics are considered. The unusually high coefficient ($r = .81$) is likely to have been the result of two interacting factors (see Mindess, 1953). First, it was obtained with an atypical sample that combined inpatients and outpatients and spanned the diagnostic continuum from schizophrenia to "anxiety hysteria" and "homosexuality." The study by Mindess was the only one incorporating such a broad range of psychopathology/health. Second, the DV in this analysis was patients' absolute level of functioning 6 months after the start of treatment. Given the large range of psychiatric problems in this sample, this particular DV probably had a highly inflated variance. As a result, the correlation would be much larger than expected. A separate coefficient obtained from the same study assessed the relation between RPRS scores and actual change in treatment—the typical DV considered in the meta-analysis. Understandably, the second coefficient ($r = .52$) was much more congruent with the other values generally reported in the literature. Because the high outlier was obtained with an atypical sample and DV combination, it was excluded from the analysis. However, Mindess recognized the problem caused by including such an extensive range of subjects in his analysis. To correct for this, he reported the coefficient between RPRS scores and outcome after omitting the schizophrenic patients ($r = .66$; $n = 70$). Our meta-analysis proceeded with these values, rather than the values listed in Tables 2 and 3.

A second outlier ($r = -.00$) was obtained in one of Bloom's (1956) two samples. This coefficient was derived from Rorschach protocols that had 10 or fewer responses and at least one card rejection. By current standards (Exner, 1993), these protocols would be considered invalid and would never be scored, much less interpreted, precisely because it is impossible to obtain useful information from patients who are so resistant to the task. Thus, for statistical and methodological reasons, this coefficient was dropped from the analysis.

There is not an obvious methodological explanation that would account for the most extreme outlier. Fiske et al. (1964) were surprised by the lack of effectiveness of the RPRS in their study ($r = -.14$), particularly because the RPRS was not correlated with any variables in their analysis. These findings struck them as even more anomalous because the RPRS had worked quite well in one of their previous investigations (Cartwright, 1958). Unlike their previous study, however, Fiske et al. (1964) had just one unsupervised person score all of their Rorschach data. Consequently, they speculated that the negative performance of the RPRS may have been a function of "unique biases" in that person's scoring procedures. Given that the data are statistically quite deviant when considered in light of other research, and given that there is reason to doubt the accuracy of RPRS scoring in this study, the coefficient from Fiske et al.'s investigation was dropped from the remaining analyses.

The final outlier ($r = .155$) fell 2.4 *SD* units below the median. Luborsky et al. (1980; Luborsky et al., 1979) and Mintz, Luborsky, and Christoph (1979) provided no information about the procedures that were used to calculate RPRS scores in their study. Thus, although there was no reason to place particular confidence in their validity coefficient, there was also no particular reason to doubt their methodology. Because the coefficient obtained in this study is also a less extreme outlier than the other three coefficients, it was retained for the final analysis.

Despite the statistical and conceptual reasons for excluding or modifying data, Table 4 still presents the full meta-analytic findings for all studies. This is done for the sake of completeness and it allows the interested reader to make comparisons with the subsequent analysis.

Final Meta-Analysis Excluding Outliers

Table 5 presents the statistical summary of the RPRS studies with outliers omitted. For this analysis, effect size calculations were carried out on 20 statistics (*K*) generated from a total of 752 participants (*N*). On average, each study contained about 38 subjects (range 11 to 79). In general, outcome was determined about 1 year (352 days) after baseline testing, although the follow-up duration was again positively skewed.

The data in Table 5 indicate that removing statistical outliers has a trivial impact on measures of central tendency. For instance, the average uncorrected validity coefficient weighted by sample size and psychometric quality is now $\rho = .443$, rather than the $\rho = .418$ value reported in Table 4, and the unweighted median effect is now $\rho = .438$, rather than $\rho = .424$.

As would be expected, removing outliers has a more pronounced effect on the variance of the validity coefficients. The chi-square statistic evaluating effect size heterogeneity now indicates that all of the uncorrected coefficients are likely to be targeting the same underlying parameter ($\chi^2 = 13.85$, $p = .84$). In fact, all of the variance in observed correlations is likely to be a function of sampling error. Because sampling error variance is slightly larger than observed score variance, the data indicate the uncorrected predictive validity of the RPRS is $\rho = .443$ with no true variance around this parameter. In other words, even though the observed coefficients range from a high of .660 to a low of .155, this extent of variability should be expected by chance alone when a number of relatively small samples are independently drawn from a single population.

Given that all of these RPRS studies appear to be targeting a single population parameter, it is appropriate to consider two additional questions: (a) How likely is it that the effect observed across all studies is really zero? and (b) How robust are the data?

The first question addresses the statistical significance of the average effect size. It asks how likely it is that a validity coefficient of .443 would be observed in this

TABLE 5
Summary Variables for the Rorschach Prognostic
Rating Scale Meta-Analysis: Excluding Outliers

Variable	*Uncorrected*	*Corrected–1*[a]	*Corrected–2*[b]
General study characteristics			
K (# of samples)	20.0		
N (# of participants)	752.0		
M n per study	37.6		
Mdn n per study	31.0		
M follow-up interval (days)	351.6		
Mdn follow-up interval (days)	206.0		
Percentage of studies contributing data	80.0		
Central tendency			
Unweighted *M* effect size	.439	.533	
Unweighted *Mdn* effect size	.438	.504	
Weighted *M* effect size (ρ)	.443	.560	.699
Variability			
Unweighted *SD r*	.1095	.1673	
Weighted *SD r*	.1040	.1598	
Weighted *SD* error	.1288	.1579	
Weighted *SD* ρ	.0000	.0242	.0092
90% credibility value	.443	.529	.687
Standard error (for homogeneous studies)	.030		
95% confidence interval about weighted *M* effect size	.385 – .501		
Maximum effect size	.660	.823	
Q_3 (75th percentile)	.514	.658	
Q_2 (50th percentile)	.438	.504	
Q_1 (25th percentile)	.377	.423	
Minimum effect size	.155	.201	
Normal-based *SD* (.75[$Q_3 - Q_1$])	.103	.176	
Do all studies estimate the same parameter?			
Heterogeneity of effect (chi-square *df* = 19)	13.847		
Probability effect is not constant	.838		
How statistically significant are the findings?			
Combined *z* (sample and quality weighted)	39.157	44.842	
Probability findings due to chance	< .0000001	< .0000001	
How many null studies ($r = .00$) would be needed to make the findings statistically nonsignificant?			
Failsafe *N*	1,046.0	1,653.0	

[a] Corrected for dependent variable range variation and false independent and dependent variable dichotomy. [b] Corrected for the preceeding and for rater unreliability.

meta-analysis if the RPRS was really an invalid scale. This question is tested by determining the standard normal deviate (z) associated with the combined results. The combined z value assessing the uncorrected validity of the RPRS is 39.157. To put this z value in context, it is helpful to recall that a standard normal z value of 1.645 is a typical cut-off for statistical significance because it is associated with a one-tailed probability of .05. Thus, 5 times out of 100 a z value of 1.645 will be found by chance even though true scale validity is really zero. A z value of 6.00 will occur 1 time in a billion by chance alone ($p = .000000001$). Given that the RPRS z value is 39.16, it is virtually inconceivable that the RPRS is actually invalid.

Another way to consider this issue is to evaluate the robustness of the findings. This is done by conducting what Rosenthal (1991) called a *file drawer analysis.* This analysis generates a "fail-safe *N,*" which indicates the number of studies with null results (i.e., a validity coefficient of $r = 0.0$) that would need to be found in order to bring the average uncorrected validity coefficient down to a nonsignificant level ($p > .05$). As Table 5 indicates, it would take 1,046 additional studies, either completed but not published or not yet completed, all with null results, to reduce the average prognostic validity of the RPRS to a level of statistical nonsignificance.[7]

The second column of Table 5 presents data on the RPRS after correcting the individual coefficients for range variation in the DV and for artificial dichotomization of the IV and DV. The data presented here are quite pertinent because they indicate what should be observed when research is conducted with two very simple modifications: (a) when all initial subjects are retained for the final analysis and (b) when researchers employ the full continuum of scores from the RPRS and the outcome criterion.

When researchers follow these basic principles, the average predictive validity coefficient will jump to $\rho = .560$, a substantial increase. The estimated true variance about the partially corrected mean validity coefficient is now .00058, with a *SD* around the population parameter of .0242 (i.e., weighted *SD* ρ). Employing the latter to determine the 90% credibility value, it can be seen that 90% of the estimates of true RPRS validity fall above a coefficient of .529. As a result, the RPRS should have excellent validity generalization across settings.

[7] A more demanding test of robustness determines the number of studies with null results (i.e., $r = 0.0$) that would be required to reduce the average uncorrected validity coefficient to some smaller value. For instance, it could be hypothesized that the Rorschach Prognostic Rating Scale (RPRS) must have an uncorrected validity coefficient of at least .10 in order for the test to have any practical utility. Even though such a value would be considered a small effect (Cohen, 1988), it seems an appropriate target because it meets or exceeds the magnitude of uncorrected validity coefficients that have been found for self-report personality tests predicting job performance (see Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Using procedures described in Hunter and Schmidt (1990, pp. 512–513), it can be determined that 69 additional studies finding null results (and no additional studies finding positive results) would need to be added to the 20 studies summarized in the meta-analysis in order to bring the uncorrected validity of the RPRS down to the level of a small effect size ($\rho = .10$).

Because it would be easy to conduct studies without bias due to range variation and false dichotomization, we estimated the statistical significance of the corrected coefficients as if they had been obtained in the primary studies. As expected, the combined *z* value of 44.84 indicates such partially corrected coefficients would not be attributable to chance. Furthermore, 1,653 null studies would have to be added to the current 20 to bring the mean partially corrected validity coefficient down to a level of statistical nonsignificance.[8]

The last column of Table 5 presents data on the partially corrected coefficients after they have been corrected again with artifact distributions to rectify the attenuating bias of measurement error. Thus, the weighted mean validity coefficient of $\rho = .699$ reflects the theoretical relationship between true scores on the RPRS and true scores on the outcome criterion. Although this validity coefficient may be of some interest, it is more valuable to know the relationship between RPRS scores as they are actually used in practice (i.e., with measurement error) and theoretically true outcomes (i.e., criterion measurement without error). This validity coefficient is estimated to be $\rho = .6305$.

Despite the homogeneity of effect sizes in this meta-analysis, we still evaluated several potential moderator variables. First, we limited the analysis to just the 17 effect sizes that were obtained from patients being seen in some form of psychotherapeutic treatment. This meta-analysis was conducted on data from 624 patients ($K = 17$) who were seen for an average of 347.6 days of treatment. Overall effect sizes were essentially the same as reported in Table 5. The weighted mean uncorrected effect size was $\rho = .449$ (rather than .443), whereas the weighted mean effect size corrected for range variation and false dichotomization was $\rho = .572$ (rather than .560).

Second, in general, it would seem to be more difficult to predict psychological change than to predict a patient's functional status, such as level of symptomatology, general effectiveness, or general level of functioning (see Mintz et al., 1979). Therefore, we classified outcome criteria on a three-point continuum ranging from measures of *genuine change* (e.g., rated benefits, residual gain, raw gain, remission of problems; Studies 1, 5, 6, 10, 11, 13, 15, 16, 17, 18, 19, and 22) to *indeterminate or other* measures (Studies 4, 7, 9, 20, and 21), to *level of functioning* measures (Studies 3, 12, and 14). Note that several studies were conservatively rated as indeterminate because the studies did not provide details about their criterion measure or because it seemed possible for raters to confound level of functioning with genuine change. The correlations between this DV classification and the observed effect sizes was .19 for the raw coefficients and –.09 for the partially corrected coefficients ($N = 20$, $p = .41$ and .70, respectively). Omitting the

[8] Ninety two null studies would have to be added to the pool of 20 to bring the partially corrected effect ($\rho = .56$)) down to the level of a small effect ($\rho = .10$).

indeterminate coefficients did not alter these associations. Although an underpowered analysis, these findings do not support the notion that the RPRS is less effective at predicting genuine psychological change or more effective at predicting level of functioning. Unfortunately, this analysis is also somewhat inconclusive because two out of the three level-of-functioning studies used samples of high functioning students rather than clinical patients.

Finally, we examined whether effect sizes were lower in studies that explicitly indicated their outcome ratings were made while the raters were blind to any Rorschach data. As indicated in Table 2, the authors of eight studies reported that outcome ratings were made by raters who had no access to Rorschach or RPRS scores, whereas no mention was made of this issue in the other 14 studies. Because this methodological factor may have resulted in statistical outliers, the association between blindedness and effect size magnitude was examined in all 22 studies. Blind ratings were no higher than "unclear" ratings for either the raw validity coefficients, blind $M = .43$, unclear $M = .37$; $t(20) = .63$, $p = .533$, or the partially corrected validity coefficients, blind $M = .47$, unclear $M = .48$; $t(20) = .11$, $p = .910$.[9]

DISCUSSION

To evaluate the ability of the Rorschach method to predict subsequent outcomes, the existing predictive literature on the RPRS was reviewed. After eliminating studies with incomplete data or methodological confounds that produced statistical outliers, a meta-analysis was conducted on 20 statistics derived from a combined sample of 752 patients. The uncorrected population correlation between RPRS scores and outcome determined about 1 year later was $\rho = .443$. After correcting for range variation and artificial dichotomies, the estimated validity of the RPRS when all subjects, all RPRS scores, and the full continuum of outcome criteria were included in the final statistical analysis was $\rho = .560$. Removing the measurement error associated with quantified outcome criteria, the theoretical validity of the RPRS increased to $\rho = .631$.

Because we made conservative decisions in the analyses, the central tendency values are likely to underestimate the actual parameters at each level of psychometric correction. Specifically, we did not transform each validity coefficient to Fisher's Z prior to aggregation in the meta-analysis. When validity coefficients are large, as they are with the RPRS, combining the raw correlations will result in a final parameter that slightly underestimates the actual value (Johnson et al., 1995). Second, our analysis of sample variances indicated that all of the studies may have

[9] When this analysis was limited to the final 20 effect sizes, blind ratings were again no higher than "unclear" ratings for either the raw validity coefficients, blind $M = .43$, unclear $M = .45$; $t(18) = .43$, $p = .676$ or the partially corrected validity coefficients, blind $M = .47$, unclear $M = .58$; $t(18) = 1.45$, $p = .165$.

had a somewhat restricted range of RPRS scores. We did not increase coefficients to reflect this range constriction. Finally, even though artifact distributions were generated from just two estimates of outcome reliability, the estimate used (rxx′ = .79) does not take into account transient error and is much higher than the .50 to .52 values that have been found in other summaries of interrater reliability (see Schmidt & Hunter, 1996).

Although central tendency may be underestimated, the effect sizes appear to be homogeneous. Validity coefficients from each study seem to be targeting the same underlying parameter, even though they were obtained from different types of patients, across different forms of intervention, over variable periods of follow-up, and across different types of outcome criteria. Hunter and Schmidt (1990) proposed a "75% rule" whereby researchers should attribute all the variance in observed effect sizes to statistical artifacts when 75% or more of the variance can be explained by known artifacts. In the uncorrected validity coefficients, sampling error accounted for 100% of the variance in the coefficients (see Table 5, column 1). In the partially corrected validity coefficients, statistical artifacts (consisting of sampling error, subject attrition or exclusion on the DV, false dichotomization of the IV, and false dichotomization of the DV) accounted for 97.7% of the variance (see Table 5, column 2). Because known artifacts account for 97.7% of the variance in these partially corrected coefficients, it is most likely that the remaining variance is also a function of uncorrected or uncontrolled artifacts, such as range variation in the RPRS, other forms of range variation in the DV, and measurement error in the IV and DV.

The available data suggest that the RPRS is equally effective when used with children or adults, with schizophrenics or healthier outpatients, with those who are court referred or those who elect treatment on their own, with those who are followed for 6 months or 36 months, and when outcome is determined by therapists or by researchers. The RPRS also seems to be an effective predictor of outcome regardless of whether outcome is measured as change over the course of treatment or functional capacity at the end of treatment. Given that one of Endicott and Endicott's (1964) samples (see Table 2) consisted of subjects on a waiting list for psychotherapy, it may be the RPRS has a generalized ability to predict growth and improvement in people suffering from psychological difficulties regardless of whether they receive treatment.

The Issue of Statistical Outliers

Because we excluded three statistical outliers from our final analysis, it is important to consider if our conclusions might be altered when considering all coefficients. With respect to effect size magnitude, the conclusions would not change. As the central tendency sections of Table 4 and 5 demonstrate, mean and median coefficients were virtually unchanged when the meta-analysis excluded statistical outliers. This is because outliers occurred at both ends of the effect size distribution.

However, the situation is different when it comes to effect size heterogeneity. With all coefficients included in the analysis, the effect sizes were not constant across studies. Because effect size heterogeneity suggests substantive moderators may be causing validity coefficients to fluctuate, it becomes important to consider what factors may contribute to effect size variability. Explicitly blind criterion ratings data did not appear to be a contributing factor. However, several other variables seemed important. Validity coefficients appeared unusually large when the patient sample contained an extensive range of symptomatology and when the criterion was ultimate level of functioning. Conversely, validity coefficients appeared unusually small when data were obtained from clinically invalid Rorschach protocols (i.e., $R < 10$ and Rejections ≥ 1) and when Rorschach protocols may have been inaccurately scored. Barring these conditions, the existing literature indicates the parameters identified in Table 5 are homogeneous across a wide range of patients, settings, interventions, and outcomes. As such, these parameters reflect stable expected values for RPRS research.

The Practical Utility of the RPRS

Following Rosenthal (1991), it is instructive to provide a context for the meta-analytic findings in order to give meaning to their magnitude. This can be done in two primary ways: first by providing an index of the practical utility of the RPRS and second by comparing the RPRS–outcome relationship to other kinds of relations. An index of practical utility can be obtained by the binomial effect size display (BESD; Rosenthal, 1991; Rosenthal & Rubin, 1982). The BESD provides a simple summary of the association between RPRS scores and subsequent outcome. To display this association, the RPRS is conceptually dichotomized into high scores and low scores, whereas outcome is also conceptually dichotomized into success and failure. Table 6 presents the BESD for $\rho = .56$. It indicates that 78% of the patients with high scores on the RPRS will have a therapeutic experience quantified as successful, whereas only 22% of the patients with low scores will be quantified

TABLE 6
Binomial Effect Size Display Quantifying the Rorschach Prognostic Rating Scale (RPRS) and Outcome Relationship

	Outcome		
RPRS Score	*% Successful*	*% Unsuccessful*	Σ
High	78	22	100
Low	22	78	100
Σ	100	100	200

as successful. Thus, the rate of successful change jumps from 22% to 78% when RPRS scores go from low to high; a difference of 56%.

Another way to evaluate the practical utility of the RPRS is to consider its use as a selection device for treatment. Assuming resources are limited and all patients seeking assistance cannot be treated, the Taylor–Russell tables (Taylor & Russell, 1939) can be used to estimate how the RPRS can increase the rate of treatment success over the base rate of success found when no selection device is employed. Using the 13 studies in Table 3 that calculated successful and unsuccessful therapy outcome, the weighted mean base rate of success is .562, indicating that 56.2% of the patients were considered successful. Rounding up to a base rate of .60 and using an estimated RPRS validity coefficient of .55, the Taylor–Russell tables can be employed to determine success rates after employing the RPRS as a screening device. If an organization has the resources to treat 80% of the people seeking services and if those with the highest RPRS scores are selected for treatment, 68% of patients would now attain successful improvement (rather than 60%). If only the top scoring 50% of patients could be treated, successful improvement now would be obtained by 78% of the patients.

RPRS Validity in Context

The uncorrected (r = .443) and partially corrected (r = .560) RPRS validity coefficients are considered "large" effect sizes (Cohen, 1988). Although such a designation is helpful for understanding the strength of the RPRS–outcome relationship, it is also illuminating to compare this relationship with other empirical relationships.

An important comparison for the RPRS is Barron's (1953) Ego Strength (*Es*) scale, derived from the Minnesota Multiphasic Personality Inventory (MMPI). Like the RPRS, the *Es* scale was developed to predict response to psychotherapy. Unlike the RPRS, the *Es* scale is a self-report measure that does not require professional time for administration and scoring—making it a potentially attractive alternative to the RPRS. It should also be noted that the *Es* scale is a good marker of the MMPI's first factor (see Graham, 1993; Greene, 1991). This factor is typically conceptualized as a dimension of general mental health versus neuroticism. Because the *Es* is a good measure of this factor, *Es* predictive validity coefficients also evaluate the relatively simple proposition that those who begin treatment reporting better mental health are able to obtain the greatest benefits from treatment.

To date, a meta-analysis of the *Es* scale's ability to predict outcome has not been published. As a result, we used the current database to generate preliminary meta-analytic findings. Six of the studies listed in Table 2 also utilized the MMPI *Es* to predict treatment outcome (Studies 5, r = .14; 6, r = .23; 8, M r = .075, N =

93; 11, *M r* = –.095, *N* = 73; 15, *r* = .03; and 16, *r* = .006).[10] Given that the MMPI is much easier to administer and score than the Rorschach, two of these studies had more subjects complete the *Es* than the RPRS. The same procedures used in the RPRS meta-analysis were used to aggregate *Es* data. The six *Es* studies obtained data from a total of 280 subjects (*M n* = 46.7) with outcome determined approximately 6 months (*M* = 198.17 days) after baseline testing. The weighted mean uncorrected effect size for the *Es* was ρ = .021, with all variance around this parameter explained by sampling error. The weighted mean effect size after correcting for range deviation and false dichotomization of continuous variables was ρ = .025. Again, all observed variance around this parameter could be explained by sampling error.

Although these findings are not comprehensive, there is no reason to suspect they present a biased picture of the utility of the MMPI *Es* to predict treatment outcome. As such, it can be concluded that the *Es* has essentially no ability to predict change as a function of psychotherapy. Furthermore, to the extent that the *Es* can be considered a proxy measure for the general dimension of self-reported mental health, these findings indicate that self-ratings of psychological health do not predict outcome and should not be used as a substitute for the RPRS.

Table 7 presents some other relevant comparisons. It can be seen that the predictive utility of the RPRS is much stronger than many other kinds of phenomena—in fact, we are aware of no other personality scale that demonstrates such consistently strong prognostic abilities. RPRS validity clearly exceeds the ability of the dexamethasone suppression test to predict response to antidepressive treatment (*r* = .00), the ability of diastolic blood pressure to predict the recurrence of a heart attack (*r* = .08), the ability of extraversion test scores to predict success in sales (*r* = .09), the ability of the cardiac stress test to predict subsequent cardiac disease (*r* = .21), and the ability of cognitive multitasking measures to predict pilot performance (*r* = .23). The relationship between RPRS scores and subsequent outcome is also stronger than the relationship between smoking and subsequent lung cancer (*r* = .08), combat exposure in Vietnam and subsequent posttraumatic stress disorder (*r* = .11), or lithium treatment and subsequent outcome in bipolar patients (*r* = .25). Furthermore, the predictive validity of the RPRS is much greater than the predictive validity of two tests that are administered each year to most American students pursuing advanced education. The RPRS is more adept at predicting outcome than the SAT at predicting college grades (*r* = .27) or the GRE at predicting graduate school grades (*r* = .28). In fact, the association between RPRS

[10] Fiske et al. (1964) did not report exact correlations for the *Es* scale in their publication. However, this information was provided in Luborsky et al. (1979). Correlations between *Es* scores and outcome in the two studies by Newmark and his colleagues (1973, 1974) were calculated from the raw data presented in these articles.

scores and subsequent outcome is slightly larger than the association between intelligence and school grades ($r = .47$) or biological sex and arm strength ($r = .52$).

Reconciling Rorschach Validity and Its Long History of Criticism

These strong meta-analytic findings for the RPRS are certainly at odds with the long history of criticism that has been leveled against the Rorschach (see Dawes, 1994, or Wood, Nezworski, & Stejskal, 1996, for recent manifestations of this tradition). However, the overall pattern of findings is quite clear and consistent. The RPRS has a powerful capacity to predict psychotherapeutic benefit and other types of longitudinal outcome. Why is there such a jarring disparity between the results of this analysis and the chronic criticism the Rorschach has received?

Undoubtedly, the answer to the preceding question is complex, having to do with many factors that relate to scientific attitudes, Rorschach theory, methodology, and psychometrics. With respect to scientific attitudes, the early, unrealistic claims that the Rorschach was an "x-ray of personality" were simply false. This excessive idealism made the Rorschach an easy target for criticism and has ultimately hampered efforts to clarify its actual scope of effectiveness (Meyer, 1996b, 1997). In addition, untempered and overzealous claims made it easy for some scientists to dismiss this whole approach to personality assessment as fanciful. Dismissive attitudes have probably also been fostered by the nature of the task itself, as having people envision objects within splotches of ink strikes many scientists unacquainted with the task as an activity akin to tea-leaf reading. Mistrust of the Rorschach may also emerge because the meaning attributed to many scores is not as obvious or self-evident as the meaning contained within the verbal items that comprise a self-report or observer-rating scale (Meyer, 1996a).

Rorschach theory, or its absence in some key areas, has also probably facilitated criticism. Unfortunately, there is not an overarching theory that specifies the locus of effectiveness for the Rorschach's various scales and scores (Meyer, 1996a). Simply stated, a cogent rationale has not yet been proffered to explain which scores should quantify overtly manifested behavior, which should quantify experiences that are consciously represented but less evident in overt behavior, and which should quantify characteristics that are generally outside the realm of conscious awareness. Matters are also complicated by the large proportion of Rorschach studies that have been conducted without a theoretical rationale linking specific scores to particular validation criteria (see Parker, Hanson, & Hunsley, 1988).

The fact that the Rorschach is an assessment method, rather than a test, has undoubtedly led to some misunderstandings also (Weiner, 1994). As a method, the Rorschach is on a par with other methods, such as self-ratings, observer ratings, and behavioral performance tasks. As such, the Rorschach is as amenable to the

TABLE 7
Sampling of Effect Sizes From Various Fields of Study

Predictor and Criterion	*Effect Size*	*N*
1. Dexamethasone Suppression Test and subsequent treatment response	.00	2,068
2. Aspirin (vs. placebo) and subsequent death by myocardial infarction	–.02	22,071
3. MMPI Ego Strength Scale scores and subsequent treatment outcome	.02	280
4. Chemotherapy and subsequent reduction in breast cancer mortality	.03	9,069
5. Ventilatory Lung Function Test and subsequent lung cancer within 25 years	.06	3,956
6. Ever smoking and subsequent incidence of lung cancer within 25 years	.08	3,956
7. Birth weight and subsequent IQ at age 7	.08	2,023
8. Diastolic blood pressure and recurrence of myocardial infarction within 9 years	.08	287
9. Alcohol use during pregnancy and subsequent premature birth	.09	741
10. Extraversion test scores and success in sales (concurrent and predictive)	.09	2,316
11. Combat exposure in Vietnam and subsequent PTSD within 18 years	.11	2,490
12. Negative emotionality test scores and subsequent heart disease	.11	(*k* = 11)
13. Conscientiousness test scores and job proficiency (concurrent and predictive)	.13	12,893
14. Electrocardiogram Stress Test scores and subsequent cardiac disease	.21	2,855
15. Cognitive Multitask Performance Test scores and subsequent pilot proficiency	.23	6,920
16. Treating bipolar disorder with lithium (vs. imipramine) and relapse within 2 years	–.25	114
17. SAT scores and subsequent college GPA[a]	.27	3,816
18. GRE verbal scores and subsequent graduate GPA[b]	.28	5,186
19. ECT for depression (vs. simulated ECT) and subsequent improvement	.31	34,714
20. Extent of parental education and IQ of biological children	.32	(*K* = 375)
21. Psychotherapy and subsequent well-being	.29	205
22. IQ scores and concurrent GPA	.47	617

(*Continued*)

TABLE 7 (Continued)

Predictor and Criterion	*Effect Size*	*N*
23. Gender and concurrent weight	.47	1,970
24. Gender and concurrent arm strength	.52	551
25. Rorschach Prognostic Rating Scale scores and subsequent outcome	.56	752
26. Neuropsychological test scores and independent diagnosis of concurrent dementia	.68	(k = 94)
27. Gender and concurrent height	.71	2,054
28. Duplex Doppler Test scores and concurrent carotid artery disease	.78	1,964

Note. 1 = Ribeiro et al. (1993); 2 = Steering Committee of the Physicians' Health Study Research Group (1988); 3 = this study, see text for details; 4 = Early Breast Cancer Trialists' Collaborative Group (1988); 5 and 6 = Islam and Schottenfeld (1994); 7 = Bouchard and Segal (1985); 8 = Jenkins et al. (1976); 9 = Data combined from Kliegman et al. (1994) and Jacobson et al. (1994); 10 = Barrick and Mount (1991); 11 = Centers for Disease Control Vietnam Experience Study (1988); 12 = Booth-Kewley and Friedman (1987), Table 7, with negative emotionality defined by anger, hostility, depression, and anxiety; 13 = Barrick and Mount (1991); 14 = Hasselblad and Hedges (1995), with N determined from the original citations given in this study; 15 = Damos (1993); 16 = Prien et al. (1984); 17 = Baron and Norman (1992); 18 = Morrison and Morrison (1995); 19 = Janick et al. (1985); 20 = Bouchard and Segal (1985); 21 = Smith and Glass (1977); 22 = Wechsler (1991); 23 and 24 = National Center for Health Statistics (1987), Tables 7 and 8 and Figure 25, respectively; 25 = this study; 26 = Christensen et al. (1991); 27 = National Center for Health Statistics (1987), Tables 13 and 14; 28 = Hasselblad and Hedges (1995). MMPI = Minnesota Multiphasic Personality Inventory; PTSD = posttraumatic stress disorder; ECT = electroconvulsive therapy; k = number of effect sizes contributing to the mean estimate; K = number of studies contributing to the mean estimate.

[a]Baron and Norman (1992) presented sufficient information to correct for range restriction, such that this coefficient is what would be expected if all students would have been admitted, regardless of Scholastic Aptitude Test (SAT) scores. Without correction, r = .199. [b]This coefficient is not corrected for range restriction. The association would be stronger if all students were admitted regardless of Graduate Record Exam (GRE) scores. The correlation between GRE quantitative scores and subsequent grade point average (GPA) is .22.

development of an infinite number of specific, operationally defined scales as any other method of assessment. When the Rorschach is erroneously viewed as a single test, this diversity can add to a sense of confusion about the published validity literature. Critics frequently have not been clear about whether they are challenging a specific, operationally defined scale derived from the method, or the Rorschach as a general device for obtaining information.

The Rorschach probably also has been viewed negatively because of a factor that is quite independent of the Rorschach. The most common method of personality assessment, the self-report method, frequently and readily yields an illusion of validity. Most often, one self-report scale is "validated" by correlating it with another self-report scale. The resulting monomethod coefficients are not genuine validity coefficients—even though they are frequently interpreted this way (see

Campbell & Fiske, 1959; McClelland, 1980; Meyer, 1996a). Because these monomethod coefficients confound method variance and trait variance, they are often dramatically larger than true heteromethod validity coefficients. As a result, it is easy for researchers to develop the false impression that self-report scales are more valid than they actually are. Simultaneously, self-report scales can erroneously appear to be more valid than Rorschach-derived scales, because the latter are almost always evaluated with more demanding heteromethod validation criteria (see Meyer, 1996b).

Another factor that is likely to have fueled Rorschach criticism is the propensity to use simple signs or single scores in research, rather than more complex aggregations of data in statistical analyses driven by theoretical considerations. When combined with researchers' propensity to ignore Rorschach method variance (Meyer, 1992, 1997), these practices makes it more difficult to demonstrate validity. Finally, clinicians less skilled in psychometrics and methodology have conducted much of the Rorschach research, whereas academically rooted psychometricians and methodologists less skilled in the clinical application of the Rorschach have generated much of the criticism. Unfortunately, these distinct realms of professional experience have seemed to foster more animosity than creative, respectful, and productive collaboration.

This meta-analysis does not rectify or address all of the preceding issues. However, it does maximize the research yield by focusing on a theoretically expected predictor–criterion relationship, a single scoring system, a theoretically organized scale that aggregates multiple components of Rorschach behavior, a replicated body of findings, and a large combined sample size. Given these conditions, validity is unmistakable. Casting wider nets, a number of other meta-analytic reviews (Atkinson, 1986; Atkinson, Quarrington, Alp, & Cyr, 1986; Parker, 1983; Parker et al., 1988) have reached the same positive conclusion about Rorschach validity. Thus, all the available meta-analytic data indicate it is inaccurate for psychological scientists to assert that the Rorschach procedure is invalid. Such a position is simply unscientific and inaccurate (see also Weiner, 1996). Rather, the data demonstrate the Rorschach can provide information that is quite valid for certain purposes. Given that the method itself is not invalid, scientists who are genuinely interested in studying personality should focus on more differentiated questions regarding the construct validity of scales and should attempt to understand the complex ways that different methods of assessment combine to produce a more complete picture of personality (Meyer, 1996a, 1997).

Predictive Validity of Other Rorschach-Derived Scales

Although the RPRS is clearly useful for predicting outcome, it is important to consider whether the robust prognostic findings are unique to the particular scores contained within the RPRS. Although the specific qualities of Rorschach behavior that are quantified by the RPRS are undoubtedly important, it does not appear that the predictive utility of the Rorschach is limited to just the Klopfer system of scoring and the specific arrangement of variables contained within the RPRS. Instead, other scales using other scoring systems have also demonstrated prognostic utility. Earlier, we indicated that some aspects of the RPRS were similar to Perry and Viglione's (1991) Ego Impairment Index (EII). Importantly, two longitudinal studies demonstrated that the EII has strong predictive and incremental validity, for periods that extend up to 5 years (Ingham, 1996; Perry, McDougall, & Viglione, 1995; Perry & Viglione, 1991).

Research with other types of scores have also provided evidence of the Rorschach's prognostic utility (e.g., Blatt & Ritzler, 1974; Cerney & Shevrin, 1974; Exner & Wylie, 1977; Greenberg & Bornstein, 1989; Hilsenroth, Handler, Toman, & Padawer, 1995; Labarbera & Cornsweet, 1985; Munroe, 1945; Russ, 1981; Tuber, 1983). Although the scales used in these studies have not been subjected to the same degree of replication as the RPRS, the findings appear quite promising. In combination, these studies suggest that the Rorschach's ability to predict outcome in a clinically relevant manner probably extends beyond the demonstrated utility of the RPRS.

Overall, given the robust empirical validity of the RPRS, we believe it would be valuable to explore the relation of this scale to more contemporary prognostic indicators derived from the Rorschach. It will also be quite important to conduct further research on the ability of these scales to predict treatment outcome and other variables related to the therapeutic process. In particular, researchers should take into account base rate predictions (e.g., Meehl & Rosen, 1955) and seek to establish the unique clinical contributions that can be made from this source of measurement.

ACKNOWLEDGMENT

We are grateful for the very helpful comments of John Hunsley and two anonymous reviewers.

REFERENCES

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

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

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

Baekeland, F., & Lundwall, L. (1975). Dropping out of treatment: A critical review. *Psychological Bulletin, 82,* 738–783.

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

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

Barron, F. (1953). An ego-strength scale which predicts response to psychotherapy. *Journal of Consulting Psychology, 17,* 327–333.

Blatt, S. J., & Ritzler, B. A. (1974). Suicide and the representation of transparency and cross-sections on the Rorschach. *Journal of Consulting and Clinical Psychology, 42,* 280–287.

*Bloom, B. L. (1956). Prognostic significance of the underproductive Rorschach. *Journal of Projective Techniques, 20,* 366–371.

Booth-Kewley, S., & Friedman, H. S. (1987). Psychological predictors of heart disease: A quantitative review. *Psychological Bulletin, 101,* 343–362.

Bouchard, T. J., & Segal, N. L. (1985). Environment and IQ. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 391–464). New York: Wiley.

*Brawer, F. B., & Cohen, A. M. (1966). Global and sign approaches to Rorschach assessment of beginning teachers. *Journal of Projective Techniques and Personality Assessment, 30,* 536–542.

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

*Cartwright, R. D. (1958). Predicting response to client-centered therapy with the Rorschach Prognostic Rating Scale. *Journal of Counseling Psychology, 5,* 11–17.

Cartwright, D. S., Kirtner, W. L., & Fiske, D. W. (1963). Method factors in changes associated with psychotherapy. *Journal of Abnormal and Social Psychology, 66,* 164–175.

Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: I. Psychosocial characteristics. *Journal of the American Medical Association, 259,* 2701–2707.

Cerney, M., & Shevrin, H. (1974). The relations between color dominated responses on the Rorschach and explosive behavior in a hospital setting. *Bulletin of the Menninger Clinic, 38,* 430–444.

Christensen, D., Hadzi-Pavlovic, D., & Jacomb, P. (1991). The psychometric differentiation of dementia from normal aging: A meta-analysis. *Psychological Assessment, 3,* 147–155.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Davids, A., & Talmadge, M. (1964). Utility of the Rorschach in predicting movement in psychiatric casework. *Journal of Consulting Psychology, 28,* 311–316.

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

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

Edinger, J. D., & Bogan, J. B. (1976). The validity of the Rorschach Prognostic Rating Scale with incarcerated offenders. *Journal of Clinical Psychology, 32,* 877–880.

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

*Endicott, N. A., & Endicott, J. (1964). Prediction of improvement in treated and untreated patients using the Rorschach Prognostic Rating Scale. *Journal of Consulting Psychology, 26,* 342–348.

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

Exner, J. E., Jr., & Wylie, J. (1977). Some Rorschach data concerning suicide. *Journal of Personality Assessment, 41,* 339–348.

*Filmer-Bennett, G. (1955). The Rorschach as a means of predicting treatment outcome. *Journal of Consulting and Clinical Psychology, 19,* 331–334.

*Fiske, D. W., Cartwright, D. S., & Kirtner, W. L. (1964). Are psychotherapeutic changes predictable? *Journal of Abnormal and Social Psychology, 69,* 413–426.

Frank, G. (1967). A review of research with measures of ego strength derived from the MMPI and the Rorschach. *Journal of General Psychology, 77,* 183–206.

Frank, G. (1993). Use of the Rorschach to predict whether a person would benefit from psychotherapy. *Psychological Reports, 73,* 1155–1163.

Fulkerson, S. C., & Barry, J. R. (1961). Methodology and research in the prognostic use of psychological tests. *Psychological Bulletin, 58,* 177–204.

Garfield, S. L. (1971). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 271–298). New York: Wiley.

Garfield, S. L. (1978). Research on client variables in psychotherapy. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (2nd ed., pp. 191–232). New York: Wiley.

Garfield, S. L. (1986). Research on client variables in psychotherapy. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (3rd ed., pp. 213–256). New York: Wiley.

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 190–228). New York: Wiley.

Garwood, J. (1977). A guide to research on the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment, 41,* 117–119.

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

Graham, J. R. (1993). *MMPI–2: Assessing personality and psychopathology* (2nd ed.). New York: Oxford.

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

Greene, R. L. (1991). *The MMPI–2/MMPI: An interpretive manual.* Boston: Allyn and Bacon.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117,* 167–178.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic.

Hilsenroth, M. J., & Handler, L. (1995). Graduate students' experiences, interests, and attitudes about learning the Rorschach. *Journal of Personality Assessment, 64,* 243–257.

Hilsenroth, M. J., Handler, L., Toman, K. M., & Padawer, J. R. (1995). Rorschach and MMPI–2 indices of early psychotherapy termination. *Journal of Consulting and Clinical Psychology, 63,* 956–965.

Huffcutt, A. I., & Arthur, W., Jr. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology, 80,* 327–334.

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

Ingham, M. (1996, March). *The Ego Impairment Index as a predictor of reactions to spousal separation.* Paper presented at the meeting of the Society for Personality Assessment, Denver, CO.

Islam, S. S., & Schottenfeld, D. (1994). Declining FEV$_1$ and chronic productive cough in cigarette smokers: A 25-year prospective study of lung cancer incidence in Tecumseh, Michigan. *Cancer Epidemiology, Biomarkers & Prevention, 3,* 289–298.

Jacobson, J. L., Jacobson, S. W., Sokal, R. J., Martier, S. S., Ager, J. W., & Shankaran, S. (1994). Effects of alcohol use, smoking, and illicit drug use on fetal growth in black infants. *The Journal of Pediatrics, 124,* 757–764.

Janick, P. G., Davis, J. M., Gibbons, R. D., Ericksen, S., Chang, S., & Gallagher, P. (1985). Efficacy of ECT: A meta-analysis. *American Journal of Psychiatry, 142,* 297–302.

Jenkins, C. D., Zyzanski, S. J., & Rosenman, R. H. (1976). Risk of new myocardial infarction in middle-aged men with manifest coronary heart disease. *Circulation, 53,* 342–347.

*Johnson, E. (1953). Klopfer's Prognostic Rating Scale used with Raven's Progressive Matrices in play therapy prognosis. *Journal of Projective Techniques, 17,* 465–470.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80,* 94–106.

*Kirkner, F., Wisham, W., & Giedt, F. (1953). A report on the validity of the Rorschach Prognostic Rating Scale. *Journal of Projective Techniques and Personality Assessment, 17,* 465–470.

Kliegman, R. M., Madura, D., Kiwi, R., Eisenberg, I., & Yamashita, T. (1994). Relation of maternal cocaine use to the risks of prematurity and low Birth weight. *The Journal of Pediatrics, 124,* 751–756.

Klopfer, B., Ainsworth, M., Klopfer, W., & Holt, R. (1954). *Development in the Rorschach technique* (Vol. 1). New York: World Book.

Klopfer, B., Kirkner, F., Wisham, W., & Baker, G. (1951). Rorschach Prognostic Rating Scale. *Journal of Projective Techniques and Personality Assessment, 15,* 425–428.

LaBarbera, J. D., & Cornsweet, C. (1985). Rorschach predictors of therapeutic outcome in a child psychiatric inpatient service. *Journal of Personality Assessment, 49,* 120–124.

Lessing, E. E. (1960). Prognostic value of the Rorschach in a child guidance clinic. *Journal of Projective Techniques, 24,* 310–321.

Luborsky, L., Chandler, M., Auerbach, A. H., Cohen, J., & Bachrach, H. (1971). Factors influencing the outcome of psychotherapy: A review of the quantitative research. *Psychological Bulletin, 75,* 145–185.

Luborsky, L., Mintz, J., Auerbach, A., Christoph, P., Bachrach, H., Todd, T., Johnson, M., Cohen, M., & O'Brien, C. P. (1980). Predicting the outcome of psychotherapy: Findings of the Penn Psychotherapy Project. *Archives of General Psychiatry, 37,* 471–481.

*Luborsky, L., Mintz, J., & Christoph, P. (1979). Are psychotherapeutic changes predictable? Comparison of a Chicago counseling center project with a Penn psychotherapy project. *Journal of Consulting and Clinical Psychology, 47,* 469–473.

Lundin, W. H., & Schpoont, S. (1953). The application of the Rorschach Prognostic Rating Scale to one intensively followed case. *Journal of Projective Techniques, 17,* 295–299.

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: Sage.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin, 52,* 194–216.

Meyer, G. J. (1992). The Rorschach's factor structure: A contemporary investigation and historical review. *Journal of Personality Assessment, 59,* 117–136.

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

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

Meyer, G. J. (1997). On the integration of personality assessment methods: The Rorschach and MMPI–2. *Journal of Personality Assessment, 67,* 297–330.

Meyer, G. J., & Resnick, J. (1996, July). *Assessing ego impairment: Do scoring procedures make a difference?* Paper presented at the XV International Congress on the Rorschach and Projective Methods, Boston.

*Mindess, H. (1953). Predicting patient's response to psychotherapy: A preliminary study designed to investigate the validity of the Rorschach Prognostic Rating Scale. *Journal of Projective Techniques, 17,* 327–334.

*Mindess, H. (1957). Psychological indices in the selection of student nurses. *Journal of Projective Techniques, 21,* 37–39.

Mintz, J., Luborsky, L., & Christoph, P. (1979). Measuring the outcomes of psychotherapy: Findings of the Penn Psychotherapy Project. *Journal of Consulting and Clinical Psychology, 47,* 319–334.

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

Munroe, R. L. (1945). Prediction of the adjustment and academic performance of college students by a modification of the Rorschach method. *Applied Psychological Monographs*(Whole No. 7), 3–104.

National Center for Health Statistics. (1987). *Anthropometric reference data and prevalence of overweight: United States, 1976–1980* (DHHS Publication No. PHS 87–1688). Washington, DC: U.S. Government Printing Office.

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

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

*Newmark, C. S., Konanc, J. T., Simpson, M., Boren, R. B., & Prillaman, K. (1979). Predictive validity of the Rorschach Prognostic Rating Scale with schizophrenic patients. *Journal of Nervous and Mental Disease, 167,* 135–143.

*Novick, J. I. (1962). Effectiveness of the Rorschach Prognostic Rating Scale for predicting behavioral change in children following brief psychotherapy. *American Psychologist, 17,* 359–360.

Parker, K. (1983). A meta-analysis of the reliability and validity of the Rorschach. *Journal of Personality Assessment, 47,* 227–231.

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.

Perry, W., McDougall, A., & Viglione, D. J. (1995). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64,* 112–118.

Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56,* 487–501.

Prien, R. F., Kupfer, D. J., Mansky, P. A., Small, J. G., Tuason, V. B., Voss, C. B., & Johnson, W. E. (1984). Drug therapy in the prevention of recurrences in unipolar and bipolar affective disorders: Report of the NIMH Collaborative Study Group comparing lithium carbonate, imipramine, and a lithium carbonate-imipramine combination. *Archives of General Psychiatry, 41,* 1096–1104.

Resnick, J. & Meyer, G. J. (1995, March). *Rorschach assessment of ego functioning: A comparison of the EII and the CESI.* Paper presented at the meeting of the Society for Personality Assessment, Atlanta, GA.

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

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

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

Russ, S. W. (1981). Primary process integration on the Rorschach and achievement in children: A follow-up study. *Journal of Personality Assessment, 45,* 473–477.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

*Schulman, R. E. (1963). Use of the Rorschach Prognostic Rating Scale in predicting movement in counseling. *Journal of Counseling Psychology, 10,* 198–199.

*Seidel, C. (1960). The relationship between Klopfer's Prognostic Rating Scale and Phillips Case History Prognostic Rating Scale. *Journal of Consulting Psychology, 24,* 46–49.

*Sheehan, J. G., Frederick, C. J., Rosevear, W. H., & Spiegelman, M. (1954). A validity study of Rorschach Prognostic Rating Scale. *Journal of Projective Techniques, 18,* 233–239.

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

Shields, R. B. (1978). The usefulness of the Rorschach Prognostic Rating Scale—A rebuttal. *Journal of Personality Assessment, 42,* 579–582.

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

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

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

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

Tuber, S. B. (1983). Children's Rorschach scores as predictors of later adjustment. *Journal of Consulting and Clinical Psychology, 51*(3), 379–385.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third Edition: Manual.* San Antonio, TX: Psychological Corporation.

Weiner, I. B. (1994). The Rorschach inkblot method (RIM) is not a test: Implications for theory and practice. *Journal of Personality Assessment, 62,* 498–504.

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

Whiteley, J. M., & Blaine, G. B., Jr. (1967). Rorschach in relation to outcome in psychotherapy with college students. *Journal of Consulting and Clinical Psychology, 31,* 595–599.

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

Windle, C. (1952). Psychological tests in psychopathological prognosis. *Psychological Bulletin, 49,* 451–482.

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

Zubin, J., & Windle, C. (1954). Psychological prognosis of outcome in the mental disorders. *Journal of Abnormal and Social Psychology, 49,* 272–281.

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage
3211 Providence Drive
Anchorage, AK 99508–8224

Received October 11, 1996
Revised January 17, 1997