Toward a Resolution of the Rorschach Controversy

Howard N. Garb
Pittsburgh VA Health Care System and University of Pittsburgh

M. Teresa Nezworski
University of Texas at Dallas

James M. Wood
University of Texas at El Paso

William M. Grove
University of Minnesota, Twin Cities Campus

William J. Stejskal
University of Virginia Health Sciences Center, Charlottesville Campus

Comments are made about the articles comprising the first round of the Special Series on the Rorschach. G. Stricker and J. R. Gold (1999) and D. J. Viglione (1999) praised the Rorschach, but they consistently failed to cite negative findings. R. M. Dawes (1999) obtained results that provide modest support for the Rorschach, but one of his data sets is flawed. J. B. Hiller, R. Rosenthal, R. F. Bornstein, D. T. R. Berry, and S. Brunell-Neuleib (1999) reported the results of a meta-analysis, but, among other problems, their coders were not blind to the results of all the studies. J. Hunsley and J. M. Bailey (1999) made a strong case for concluding that there is no scientific basis for using the Rorschach. Recommendations are made for resolving the Rorschach controversy.

As noted by Hunsley and Bailey (1999), the Rorschach has the dubious distinction of being, simultaneously, the most cherished and the most reviled of all psychological assessment tools . . . . The Rorschach is held in great esteem by many psychologists for its ability to access intrapsychic material, whereas others point to the Rorschach as a prime example of unscientific psychological assessment. (p. 266)

Indeed, a widely diverse set of opinions about the Rorschach are expressed in this Special Section. Stricker and Gold (1999, p. 244) concluded that “the Rorschach is a valid and reliable method for assessing unconscious experience” and Viglione (1999, p. 251) concluded that “the evidence reveals that Rorschach variables are useful for clinical, forensic, and educational practice,” but Hunsley and Bailey (1999, p. 266) concluded that “there is currently no scientific basis for justifying the use of Rorschach scales in psychological assessments.” In contrast, Dawes (1999) and Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib (1999) were cautious in reaching conclusions, though they both obtained results that provide modest support for the Rorschach.

In our article, we try to explain why scholars have reached contradictory and opposite conclusions. Even more important, we try to clarify whether research supports the use of the Rorschach. We also evaluate whether the new results presented by Dawes (1999) and Hiller et al. (1999) blunt the criticisms of the Rorschach made by Hunsley and Bailey (1999). Finally, suggestions are made for resolving the Rorschach controversy.

Comments on Stricker and Gold (1999)

Stricker and Gold (1999) cited studies that support the validity of the Rorschach, but they failed to cite studies that do not support it. For example, according to Stricker and Gold, “the Rorschach (Exner, 1993) is effective in identifying patients who are depressed, schizophrenic, or acutely suicidal” (p. 4). However, empirical research suggests that the Rorschach is not useful for all of these tasks (Wood, Nezworski, & Stejskal, 1996a, 1996b). Even Viglione (1999) concluded that the Comprehensive System Depression Index (DEPI) “is not routinely sensitive to clinically diagnosed depression in adults or adolescents” (p. 22) and that “more work needs to be done to justify the use of the revised SCZI [Schizophrenia Index] as a diagnostic indicator of schizophrenia” (p. 21). Results on diagnosing schizophrenia, identifying patients who are depressed, and identifying patients who are acutely suicidal are described here.

The SCZI has been found to be related to diagnoses of schizophrenia and other psychotic disorders (Archer & Gordon, 1988; Meyer, 1993). However, as noted by Hunsley and Bailey (1999), the Minnesota Multiphasic Personality Inventory (MMPI) is considerably less expensive and seems to be about equally valid for the diagnosis of schizophrenia and other psychotic disorders as is
the Rorschach. For example, although the Comprehensive System SCZI is related to schizophrenia (Exner, 1986, 1993), no published studies have shown that it can add significant incremental validity for diagnoses of the disorder beyond what can be obtained from an interview and MMPI scores. In short, although the Rorschach can be used to identify schizophrenia, there are usually easier, more effective ways to make this diagnosis.

According to Exner (1993, pp. 260–264, 309–311), the DEPI can be used to differentiate inpatient depressives from nonpatient adults. Using a cutoff score of 5 or higher on the DEPI, Exner found that correct judgments were made for 75% of the inpatient depressives and 97% of the nonpatient adults. However, other investigators have failed to detect a statistically significant relation between the DEPI and diagnoses of depression (Archer & Gordon, 1988; Archer & Krishnamurthy, 1997; Ball, Archer, Gordon, & French, 1991; Caine, Fruch, & Kinder, 1995; Carlson, Kula, & St. Laurent, 1997; Carter & Dacey, 1996; Kadle, 1989; Lipkin, 1989; Lipovsky, Finch, & Belter, 1989; Meyer, 1993; Sells, 1991; Silberg & Armstrong, 1992; Viglione, Brager, & Haller, 1988; but for positive findings in regard to the DEPI, see Jansak, 1997, and Singer & Brabender, 1993). For example, (a) Carter and Dacey (1996) found that DEPI scores were not significantly related to diagnoses of depression in a sample of 118 hospitalized adolescents; (b) Archer and Krishnamurthy (1997) found that there was no significant relationship between DEPI scores and diagnoses of depression in a sample of 152 adolescents; and (c) Carlson et al. (1997) found that only 28% of adults with Major Depressive Disorder were identified by using the DEPI.

According to Stricker and Gold (1999, p. 243), "Exner's (1993) Suicide Index is a remarkable demonstration of the power of this tool [the Rorschach] to predict complex, clinically essential phenomena. The Suicide Constellation (S-Con) was derived by Exner and Wylie (1977), using Rorschach protocols obtained from a sample of patients who had committed suicide and a sample of control subjects. For the derivation sample, the original S-Con was able to correctly identify 75% of the suicide patients and 100% of the nonpatients. In a cross-validation study (Exner, 1986, pp. 411–416; pp. 342–345), the S-Con was able to correctly identify 74% of the suicide patients and 100% of the nonpatients. For both the derivation sample and the cross-validation sample, a cutoff score of 8 was used. It is surprising that hit rates for the validation sample were nearly as high as hit rates for the derivation sample (Wood, Nezworski, & Stejskal, 1996a, 1996b; for a response, see Exner, 1996). Substantial shrinkage of predictive validity should have occurred because the derivation sample was only moderate in size (the sample included 59 patients who had committed suicide) and a large number of variables was considered for inclusion (apparently Exner and Wylie used the derivation sample to look at the predictive validity of over 100 variables). Importantly, in a cross-validation study by independent investigators (J. R. Eyman & Eyman, 1992), only 1 of 50 patients who had committed suicide had an S-Con score of 8 or higher. These researchers used Exner's Comprehensive System to score the protocols, but used the system of Rapaport, Gill, and Schafer (1946) to administer the test. S. K. Eyman and Eyman (1987, p. 10) noted that the average number of responses in their sample was 23 (which is slightly higher than one would expect, given the Exner norms), and the discrepancy between their findings for S-Con and Exner's findings for S-Con was so large that it "could not be accounted for solely by a difference in test administration." Also, as noted by Viglione (1999), the S-Con was not able to discriminate between suicide attempters and nonattempters in either adolescent or adult inpatient samples (Meyer, 1993; Silberg & Armstrong, 1992), though some of the variables making up the S-Con have been related to suicide behavior (Arffa, 1982; Kendra, 1979; Silberg & Armstrong, 1992). Similarly, Long (1995) found that the S-Con did not significantly differentiate between suicidal and nonsuicidal adolescent inpatients. Finally, it is important to note that a new version of the S-Con has been developed (Exner, 1993), but it has not yet been evaluated in a cross-validation study.

Stricker and Gold (1999, p. 243) gave another example of how a Rorschach scale can be used to predict "complex, clinically essential phenomena." They argued that the Rorschach Prognostic Rating Scale (RPRS) is "extraordinarily predictive of psychotherapeutic outcome" (p. 243). Although it is true that the RPRS was supported in a recent meta-analysis (Meyer & Handler, 1997), it is also true that the most recent studies on the RPRS are about 20 years old and did not use the Comprehensive System. It is important that the positive results for the RPRS be replicated, using modern administration and scoring systems. Interestingly, such studies do not require the gathering of new data, but they can be done by using data already in clinic archives. Finally, when the RPRS has been compared (in the same studies) to interviewer–diagnostician ratings of psychological health and sickness (e.g., using the Health–Sickness Rating Scale; Luborsky, 1975), interviewer–diagnostician ratings have been more accurate than the RPRS (Fiske, Cartwright, & Kirtner, 1964; Luborsky, Mintz, & Christoph, 1979; see also Luborsky, Crits-Christoph, Mintz, & Auerbach, 1988; Luborsky et al., 1993).1

Though none of the studies that reported negative results were cited by Stricker and Gold (1999), as a matter of fairness it should be pointed out that their article focused on theoretical issues. Unfortunately, their theoretical comments seem designed to provide a rationalization for the sometimes poor performance of the Rorschach. For example, they argued the following:

The profile that emerges from the Rorschach may or may not correspond to the profiles that are obtained from self-report measures, interviews, or behavioral observations. Nor is correspondence or lack

---

1 Two studies have compared interviewer–diagnostician ratings of psychological health and sickness to the RPRS: the Penn Psychotherapy project and the Chicago study. The data for the Chicago study were initially presented by Fiske et al. (1964). These data were later re-analyzed by Luborsky et al. (1979), who also presented the results for the Penn Psychotherapy project. A meta-analysis (using DSTAT; Johnson, 1989) was conducted to compare interviewer–diagnostician ratings to the RPRS. We used the following correlation results: .275 (ratings, Penn study), .155 (RPRS, Penn study), .04 (ratings, Fiske et al., 1964, analysis of Chicago data), –.10 (RPRS, Fiske et al., 1964, analysis of Chicago data), .265 (ratings, Luborsky et al., 1979, analysis of Chicago data), and –.205 (RPRS, Luborsky et al., 1979, analysis of Chicago study). A single effect size for interviewer–diagnostician ratings and a single effect size for RPRS were calculated, using the results from the two analyses of the Fiske et al. data set. Effect sizes were also calculated for the Luborsky et al. (1979) data set. The overall effect size for interviewer–diagnostician ratings is \( d = .419 \) (95% confidence interval = .20/.64), and the overall effect size for the RPRS is \( d = .003 \) (95% confidence interval = 0.30/+.31).
of correspondence seen as more or less meaningful or desirable. (p. 244)

Using this reasoning, one can easily brush aside most negative results that are obtained for the Rorschach. Hunsley and Bailey (1999) have already commented on how the poor performance of the Rorschach is sometimes rationalized:

Although a range of explanations for the lack of convergence has been offered by Rorschach proponents . . . , these explanations are almost entirely post hoc . . . . Claims that the Rorschach’s low-convergent validity reflects its virtues rather than its limitations must be supported by empirical evidence, not simply rhetoric. (p. 270)

That is, rather than provide post hoc explanations, Rorschach advocates need to provide evidence that the Rorschach contributes something that is not already contributed by self-report measures, interviews, or behavioral observations.

All in all, there is a lot of confusion about how Rorschach results should be evaluated. For example, according to Strieker and Gold (1999), the Rorschach is “most likely to be useful, both in research and in clinical practice, when interest is focused on unconscious, longitudinal, and structural dimensions of functioning” (p. 240). Thus, one might think that one can evaluate the validity of the Rorschach by comparing Rorschach results to results from other “projective techniques.” However, Hiller et al. (1999) found in their meta-analysis that validity coefficients for the Rorschach were actually smaller when “projective techniques” were used as criterion measures.

Because there is confusion about how Rorschach results should be evaluated, Rorschach advocates sometimes interpret negative or neutral results as supporting the Rorschach. For example, when Hiller et al. (1999) obtained their negative findings, they concluded that “The Rorschach may have psychometric properties that are different from, and perhaps superior to, those of other ‘projective’ measures that are commonly mentioned in the same breath as the Rorschach” (p. 291).

Comments on Viglione (1999)

Whereas Strieker and Gold focused on theory, Viglione (1999) reviewed the empirical literature on the Rorschach. Unfortunately, he consistently ignored negative findings. We critique his work on the following topics: (a) temporal consistency, (b) intrater reliability, (c) cultural diversity, (d) clinical judgment, (e) DEPI, and (f) biological predisposition for psychosis and schizophrenia.

Temporal Consistency

According to Viglione (1999), “The great majority of Rorschach Comprehensive System (CS) variables and configurations have shown impressive temporal consistency reliability” (p. 252). However, despite Viglione’s claim, test–retest results have been reported for only about 40% of the variables in the Comprehensive System.

Although the Comprehensive System contains at least 125 variables, Exner and his colleagues have reported test–retest coefficients for only 51 variables (Exner, 1974, 1978, 1980, 1986, 1988, 1991, 1993, 1999; Exner, Armbruster, & Viglione, 1978; Exner, Thomas, & Mason, 1985; Exner & Weiner, 1982, 1995; Haller & Exner, 1985). These 51 variables are listed in Appendix A. Test–retest coefficients for these variables usually range from around .30 to .90, with median values in the .80s or mid-to-high .70s (Meyer, 1997a, p. 487). Exner and Weiner (1995, p. 22) have argued that .75 represents a “liberal minimum” for test–retest coefficients over long time periods.

Exner (1978) has stated that the test–retest coefficients for these variables were reported “because they represent the ‘core’ of structural data from which Rorschach interpretation proceeds” (pp. 67–68). However, a close examination of the variables does not support his claim. For example, Appendix B lists 30 Comprehensive System variables whose test–retest reliability has not been reported by Exner. Appendix B indicates that no test–retest figures have been reported for the SCZI, the DEPI, the Coping Deficit Index (CDI), or the Hypervigilance Index (HVI), even though these four variables have all been designated by Exner (1991, pp. 144–145) as being “key variables” in the Comprehensive System.

Investigators other than Exner have also reported test–retest coefficients. Adair and Wagner (1992) and Perry, McDougall, and Viglione (1995) reported results for seven Comprehensive System variables not already reported by Exner and his colleagues. In the Adair and Wagner (1992) study, participants were 50 outpatients with schizophrenia who were tested twice with an average interval of 6.4 years between administrations (range = 1 to 18 years). They received treatment between administrations. Test–retest coefficients were reported for six Comprehensive System variables: DV, DR, INCOM, FABCOM, ALOG, and WSUM6. Coefficients ranged from .16 for ALOG to .50 for DV, with a median value of .38. Similarly, in the Perry, McDougall, and Viglione (1995) study, participants were 17 depressed patients who were tested twice with approximately 5 years between administrations. They had also received treatment between test administrations. Test–retest coefficients were reported for three Comprehensive System variables: Number of responses, FQ—, and WSUM6. Coefficients ranged from .52 to .63, with a median value of .59.

Test–retest reliability coefficients have also been reported in other studies. A dissertation by Erstad (1996) examined test–retest reliability for 26 Comprehensive System variables that had also been studied by Exner and his colleagues for test–retest reliability. Participants were 17 normal older adults, ages 61 to 95. The length of time between test and retest was unspecified. Coefficients ranged from .02 to .95, with a median of .68. Coefficients for 18 of the 26 variables fell below the .75 level recommended by Exner.

In this article, the terms “projective” and “projective techniques” are used in the historical sense; they refer to tests that use ambiguous stimuli and elicit open-ended responses.

3 Exner (1980, 1988) and Exner, Armbruster, and Viglione (1978) provided tables in regard to the stability of direction for some additional CS variables that had been dichotomized or dichotomized (e.g., EB was dichotomized into introversive, extroversive, and ambiative). However, these tables do not provide test–retest coefficients, nor do they provide enough detail so that coefficients can be computed for the variables in their continuous, dichotomized, or trichotomous forms. In particular, the tables fail to provide information in regard to the frequency of various scores at test and retest. Without frequency information, the interpretation of the tables is ambiguous. For example, the numbers reported in these tables might represent either “zero” reliability or “near perfect” reliability, depending on the frequency of scores.
and Weiner (1995, p. 22). In a second study (Schwartz, Mebane, & Malony, 1990), a nonstandard form of administration was used. The Rorschach was administered to deaf patients. A written form and a signed form of the test were administered to each patient. These test–retest coefficients also tended to be lower than the coefficients that have been obtained by Exner: Results for 21 of the 25 variables fell below the .75 level recommended by Exner and Weiner (1995, p. 22) as a “liberal minimum” for test–retest coefficients. In this study, test administrations were about 1 week apart.

In conclusion, for the standard administration of the Comprehensive System, test–retest coefficients have not been reported for nearly 60% (74 of 125) of the Comprehensive System scores. Also, studies conducted outside of Exner’s lab have reported low test–retest reliability figures, though test–retest intervals were long (or unspecified) in three of the studies and an unusual administration procedure was used in the fourth study. For these reasons, we cannot accept Viglione’s claim that “the great majority” of Comprehensive System scores have “impressive temporal consistency” (Viglione, 1999, p. 252).

**Interrater Reliability**

Viglione (1999) also defended the interrater reliability of the Rorschach. He made the following anecdotal observation:

> I had the opportunity to review hundreds of interscorer reliability practices and coefficients for all types of scores with all sorts of base rates; with dichotomous, multiple categorical, or rating-scale distributions; completed in Asia, North America, and Europe; and with all sorts of reliability statistics. The fact is that just about everybody reports adequate interscorer agreement (Meyer, 1997a), despite the reported weaknesses in Exner (1993) as characterized by Wood, Nezworski, and Stejskal (1996a, 1996b). (p. 252)

Nevertheless, Viglione did not describe any study on interrater reliability that supports the Rorschach. He did cite an article and meta-analysis by Meyer (1997a) that forcefully argues in favor of the interrater reliability of the Rorschach, but he failed to cite a critique of Meyer’s article by Wood, Nezworski, and Stejskal (1997; see also Meyer, 1997b).

To our knowledge, only one well-done full-scale study on the interrater reliability of clinicians who used the Comprehensive System has been completed. Acklin, McDowell, Verschell, and Chan (2000) computed intraclass correlation coefficients for approximately 95 Comprehensive System scores. Interrater reliability was calculated for a clinical sample (N = 20) and a nonclinical sample (N = 20). Rorschach protocols were scored by two graduate clinical psychology students. Each had advanced training in the use of the Comprehensive System and a minimum of 3 years of experience in Comprehensive System coding procedures. The results for both samples were similar. The median reliability of Comprehensive System scores was in the low 80s, the maximum was 1.0, and the minimum was approximately .20. As Acklin and his coauthors pointed out, interrater reliability was acceptable and at times even excellent for many Comprehensive System scores. However, about 50% of the coefficients fell below the minimum level of .85 set forth by Exner (1978, p. 14) and Groth-Marnat (1997, p. 397). Furthermore, reliability was low for several widely used scores. For example, reliability coefficients for the SCZI were .45 and .56. Similarly, interrater reliability was low for Adjusted D (.53 and .68), X-% (.62 and .66), and FC:CF + C (.54 and .17). These unfavorable values raise serious questions about whether these scores should be used in clinical and forensic settings.

**Cultural Diversity**

According to Viglione (1999), “In the United States, the Rorschach has demonstrated its usefulness with a variety of ethnic and subcultural groups” (p. 252). However, numerous investigators have described shortcomings of the Rorschach for the assessment of American minority groups and for non-Americans (Constantino, Flanagan, & Malgady, 1995; Cuellar, 1998; Dana, 1986, 1993, 1996, 1998; Frank, 1992, 1993; Gray-Little, 1995; Gray-Little & Kaplan, 1998; Howes & DeBlassie, 1989; Okazaki & Sue, 1995; Velasquez, 1995; Velasquez & Callahan, 1992; but also see Butter, Nezami, & Exner, 1998; Ritzler, 1996). An important shortcoming is the lack of appropriate normative data for many ethnic groups, for example, Black adults, Hispanic American adults and children, and Native American adults and children. As observed by Gray-Little (1995), “In the use of inkbolts with ethnic minorities, the assessor must be aware that there are few empirical data to provide a guide” (p. 150).

Rorschach interpreters may overpathologize the Rorschach responses of minority participants who take the test (Dana, 1993, pp. 156), especially if a client’s responses are interpreted within a Eurocentric framework, rather than within the participant’s own cultural context. Problems may also arise because of examiner effects when the examiner and the client are from different cultural groups, although this problem has received little attention from Rorschach researchers. Especially problematic are cases in which the examiner or interpreter comes from a different linguistic group than the participant (Weiner, 1998, pp. 49–51).

In most of the studies cited by Viglione, a Rorschach score was described as being related to a criterion variable. For example, stress was related to an increase in inanimate movement responses in Israel, and risk for schizophrenia was related to thought disturbance indicators in Finland. Though these studies are of value, a more rigorous approach would involve examining slope bias (Cleary, Humphreys, Kendrick, & Wesman, 1975). Using this approach, one can examine the differential validity of Rorschach indices across ethnic groups.

Viglione cited only one study that provided normative data. In this study (Krall et al., 1983), the Comprehensive System was used to score the Rorschach protocols of 272 inner-city Black children. Viglione (1999) claimed that “CS normative data from nonpatient African American children . . . closely resembled the Comprehensive System nonpatient reference sample data available at the time” (p. 252). However, Krall et al. reached a different conclusion: They concluded that “boys and girls in the sample seem to be more similar to themselves than they do to other normative samples” (p. 155). Krall et al. examined 10 variables that overlapped with the Comprehensive System. Compared to the normative data for the Comprehensive System (the normative data in use at that time; Exner & Weiner, 1982), the Black children showed clear and

---

4 We have excluded test–retest results for Rorschach variables that are not listed in Exner’s books as being part of the Comprehensive System.
consistent differences on five (50%) of these variables. Specifically, the protocols of the Black children contained (a) substantially lower $F+\%$ scores compared with the Comprehensive System normative data for children (range = 48% to 69% vs. range = 83% to 89%), (b) over three times as many perseverations, (c) approximately one fewer response ($R$), (d) two to three fewer whole responses, and (e) one to two more detail responses. When compared with more recent normative data for the Comprehensive System (Exner, 1993; Exner & Weiner, 1995), the differences were even more striking: The Black children in the study by Krall et al. differed from the recent normative data on 7 out of 11 variables.

Other investigators have also found that Rorschach scores vary as a function of ethnic group. For example, Moon and Cundick (1983) found important differences between White Americans and Korean Americans on Comprehensive System variables. Glass, Bieber, and Tkachuk (1996) compared Alaskan native and nonnative prisoners and concluded that "there were clear differences between native and nonnative inmates on both the MCMI II and the Rorschach" (p. 583, abstract). Furthermore, Glass et al. (1996) found that the native prisoners differed significantly from Comprehensive System norms for two thirds of all Rorschach variables. Distinctive Rorschach score patterns that used systems other than the Comprehensive System have also been reported for Japanese Americans (Devos, 1989), Japanese citizens (Devos, 1989), Apache Indians (Day, Boyer, & Devos, 1989), and Arab participants (Devos & Miner, 1989). Similarly, the meaning of some scores and ratios is also being challenged for Hispanics in the United States and for populations in several European countries (Andronikof-Sanglade, 1995; Constantino et al., 1995).

In conclusion, given the presence of differences across ethnic groups, the lack of norms for many of these groups, and the small number of studies that have examined the differential validity of Rorschach indices across ethnic groups, we believe that the Comprehensive System should not be used to evaluate members of American minority groups or individuals from outside of the United States. We agree with a conclusion reached by Dana (1993, p. 160): "The Rorschach and the Exner Comprehensive versions are not recommended for routine cross-cultural applications."  

Clinical Judgment

Viglione (1999) also reviewed research on clinical judgment. In these studies, clinicians made judgments after examining Rorschach protocols. According to Viglione (1999, p. 253), "The evidence supports the conclusion that accurate clinical interpretation is achievable by well-trained judges who have clear theoretical concepts in mind." Interestingly, nearly half of the studies cited by Viglione were published in the 1960s and 1970s, with the remainder of the studies published in the 1980s. In none of the studies were Rorschach protocols scored by using the Comprehensive System. Viglione's conclusion is surprising because in a book on clinical judgment (Garb, 1998), the author (one of the coauthors of this article) concluded that "results from judgment studies have been substantially more favorable for objective personality tests than for projective tests" (p. 237).

Unfortunately, Viglione consistently ignored negative findings on clinical judgment. Specifically, he ignored findings on incremental validity, the appropriateness of confidence ratings, and the relation between training in psychology and the validity of judgments. In fact, positive results have never been obtained for the Rorschach in studies on clinical judgment and incremental validity, confidence and validity, and training and validity. For example, the validity of judgments made by clinicians did not increase when (a) the Rorschach was added to demographic data (e.g., Gadol, 1969), (b) a test battery that included the Rorschach and Sentence Completion Test was added to demographic data (e.g., Cochrane, 1972), and (c) the Rorschach was added to other test results or biographical information (e.g., Bilett, Jones, & Whitaker, 1982; Golden, 1964; Perez, 1976). Similarly, when clinicians made judgments, using the Rorschach, the amount of confidence they had in their judgments was unrelated to validity (e.g., Albert, Fox, & Kahn, 1980; Gadol, 1969), and experience, training, and presumed expertise were unrelated to validity (e.g., Gadol, 1969; Turner, 1966; for a review, see Garb, 1989).

It is helpful to look at one of the studies. Turner (1966) gave Rorschach protocols to 25 Fellows of the Society for Projective Techniques, 25 recently graduated doctoral psychologists, 25 graduate students in clinical psychology, and 25 undergraduate psychology majors. Judges indicated whether statements were true of patients. The statements described symptoms, behaviors, motivations, emotions, and conflicts. Criterion ratings were made by hospital personnel. Items were included only if agreement among hospital personnel was good. Validity did not vary as a function of training, experience, or presumed expertise. Thus, Turner (1966) concluded that "general psychology students with no Rorschach experience whatsoever were able to predict as well as the experts with many years of experience" (p. 8).

The studies on clinical judgment are not definitive. Most of the studies are old, and it is unclear whether the Comprehensive System was used by any of the clinicians in the studies. However, a strong argument can be made against Viglione's (1999) conclusion that "the evidence supports the conclusion that accurate clinical interpretation is achievable by well-trained judges who have clear theoretical concepts in mind" (p. 253).

Depression Index

According to Exner (1991), the $DEPI$ will "correlate very highly with a diagnosis that emphasizes serious affective problems" (p. 146). However, as already noted, most studies have found that the original and revised versions of the $DEPI$ are unrelated to diagnoses of depression. Viglione (1999, p. 256) argued that the $DEPI$ can be used to diagnose depression with the help of moderator variables. For example, he argued that "with adolescents, the lack of sensitivity of the $DEPI$ may be associated [to some degree] with low $R$ (low mean $R$ and some $R < 14$ records) and the prevalence of form-dominated, simple responses (high $\Lambda$)." However, he did not cite any studies that demonstrate this moderator effect. Furthermore, in a study that used adults as participants, "differential predictions of $DEPI$ elevations were not found when number ($R$)
Psychosis Proneness and Biological Predisposition for Schizophrenia

According to Viglione (1999), "The available evidence reveals that Rorschach thought disturbance variables (Ego Impairment Index and Thought Disturbance Index [TDI]) are associated with psychosis proneness and biological predisposition for schizophrenia" (p. 254). Each of these claims will be addressed separately.

Viglione (1999) claimed that Rorschach thought disturbance variables are associated with psychosis proneness. This may be true. It is possible to use the Rorschach to measure mild forms of thought disturbance, and it is plausible that individuals with mild thought disturbance are more likely to become psychotic than are individuals with no thought disturbance. However, no longitudinal studies have been conducted to show that Rorschach scores are associated with an outcome of psychosis. That is, in none of the studies cited by Viglione were participants followed over time to learn whether they became psychotic. Thus, it is premature to claim that Rorschach scores are associated with psychosis proneness. Equally important, less time-consuming measures have already been found to be valid for this task. For example, an interview-based assessment system (the Wisconsin Manual for Assessing Psychotic-like Experiences) and five true–false self-report scales (the Physical Anhedonia Scale, the Perceptual Aberration Scale, the Magical Ideation Scale, the Impulsive Nonconformity Scale, and the Revised Social Anhedonia Scale) have been used to measure psychosis proneness (Chapman, Chapman, & Kwapił, 1995). Use of these scales as measures of psychosis proneness and as predictors of schizophrenia-spectrum disorders has been supported (Chapman, Chapman, Kwapił, Eckblad, & Zinser, 1994; Kwapił, 1998; Kwapił, Chapman, & Chapman, 1999; Kwapił, Miller, Zinser, Chapman, & Chapman, 1997), and long-term follow-up data (20–25 years) continue to be collected (Lenzenweger, 1998; Lenzenweger & Korfine, 1995).

Viglione’s (1999) other claim is that Rorschach thought disturbance variables can be used to detect the biological predisposition for schizophrenia. In our view, the results described by Viglione offer little support for this claim. We will first describe studies that looked at family members of individuals with schizophrenia, and we will then look at studies in which individuals with schizophrenia performed laboratory tasks.

Viglione (1999) overinterpreted the results from two family studies. For example, in describing the results from an article by Arboleda and Holzman (1985), Viglione concluded that "the TDI was associated with biological risk for psychosis" (p. 254). In this article, children born to mothers who had been hospitalized for a psychosis had a level of thought disorder that was three times higher than that of normal children. Viglione interpreted this result as meaning that the Rorschach can be used to detect biological risk for schizophrenia, but the results can also be interpreted another way: Many of these children may have had a thought disorder, the TDI may be able to measure thought disorder, and the thought disorder may have been due to environmental factors. In fact, in discussing the results of their study, Arboleda and Holzman (1985) concluded that the present study does not address the genetic and the environmental influences with respect to the origin of the disordered thought that characterizes the high-risk children. The children in this study had lived with their parents up to the time of the evaluation; their disordered thinking may have been genetically transmitted or may have been an environmentally adaptive response to their parents’ thought disorder. (p. 1012)

In a second family study cited by Viglione (Wahlberg et al., 1997), children of schizophrenic biological mothers and comparison group children with nonschizophrenic mothers did not obtain significantly different scores on a Rorschach measure of disordered thinking. As observed by Wahlberg et al. (1997), "High genetic risk in itself was not associated with greater vulnerability to schizophrenic thought disorder in the adoptees, as indicated by the Index of Primitive Thought" (p. 355, abstract). The Index of Primitive Thought is composed of categories that are also scored when using the TDI (contamination, confabulation, and fabulized combination). Viglione did not describe this result, but instead he noted a positive result: "Adoptees of biological mothers with schizophrenia demonstrated a higher proportion of elevated scores only under conditions when their adoptive parents’ communication patterns were deviant" (p. 254). This does not demonstrate that this Rorschach measure can be used to measure the biological disposition for schizophrenia. Instead, it is likely that adoptees from schizophrenic biological mothers developed a thought disorder only when their adoptive parents’ communication patterns were deviant. One can infer that the Index of Primitive Thought is able to measure thought disorder, but one should not infer that the Index of Primitive Thought is able to measure biological predisposition.

Viglione also described results from laboratory studies. It is true that laboratory studies can be conducted to learn if an assessment
instrument can be used to detect biological predisposition to schizophrenia. Historically, laboratory studies have demonstrated that deficits in attentional functioning and information processing can be used as indicators of biological predisposition to schizophrenia. As observed by Nuechterlein, Edell, Norris, and Dawson (1986),

these studies have found that impairments that are present among many adult schizophrenic patients on specific laboratory performance tasks, such as the continuous performance test (CPT) and the forced-choice span of apprehension, are also found among a subgroup of the offspring of a schizophrenic parent and among nonpsychotic individuals at hypothesized risk due to schizotypal personality characteristics. (p. 408)

However, results from laboratory studies offer only slight (and unreplicated) evidence that the Rorschach can be used to detect biological predisposition to schizophrenia.

In two of the studies cited by Viglione (1999), participants were individuals with schizophrenia (Nuechterlein et al., 1986; Perry & Braff, 1994; see also Perry, Geyer, & Braff, 1999). Schizophrenics who did poorly on laboratory tasks (e.g., attention performance tasks) sometimes did poorly on Rorschach measures of thought disorder. Instead of demonstrating that these Rorschach measures are valid indicators for biological predisposition to schizophrenia, the results indicate that schizophrenics with poor attention and information-processing skills are sometimes more likely to suffer from severe forms of thought disorder.

In another study cited by Viglione (1999), 12 participants with a diagnosis of Schizotypal Personality Disorder received the Rorschach (Cadenhead, Perry, & Braff, 1996). Visual backward masking (a measure of information processing) was significantly related to one Rorschach variable (Ego Impairment Index). Obviously, given the small sample size, these results are in need of replication. Furthermore, they need to be replicated at an independent laboratory.

One more laboratory study will be discussed. The strongest evidence that the Rorschach can be used to detect the biological predisposition for schizophrenia comes from a study conducted by Holzman et al. (1974). Viglione (1999) described the results of this study:

People with schizophrenia and their first-degree relatives produced deviant eye tracking (N = 139; Holzman et al., 1974) and elevations on a modified Rorschach Delta Index. As it turned out, the relationship of the Delta Index and eye tracking was greater than the relationship between clinical diagnosis of schizophrenia and eye tracking. (p. 254)

However, even here problems can be described. For example, because the data were collected in the early 1970s, it is likely that a substantial number of the patients that were diagnosed as having schizophrenia really had bipolar affective disorder. With the publication of the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III; American Psychiatric Association, 1980), the criteria for schizophrenia became more stringent so that fewer diagnoses of the disorder were made. This was not an arbitrary change in the criteria, but instead it was made because some patients were being diagnosed as having schizophrenia even though they had psychotic symptoms only during manic, major depressive, or mixed episodes (J. E. Cooper et al., 1972; Garvey & Tuason, 1980; Joyce, 1984). Certainly the Holzman et al. (1974) study needs to be replicated, using updated methods.

In conclusion, little evidence indicates that the Rorschach can be used to evaluate psychosis proneness or biological predisposition for schizophrenia. Family studies, laboratory studies, and longitudinal studies can be conducted to help us evaluate Rorschach scores, but it seems likely that other tests (e.g., the Revised Social Anhedonia Scale) will prove to be more valuable for these tasks (Lenzenweger, 1998; Lenzenweger & Korfine, 1995).

Comments on Dawes (1999)

To learn about the incremental validity of the Rorschach, Dawes (1999) reanalyzed the data from two studies. He found that when using a linear rule to predict level of psychopathology, the addition of the Ego Impairment Index to two other Rorschach variables (R and XQUAL) led to a small increase in validity from $r^2 = .11$ to $r^2 = .15$ (Meyer & Resnick, 1996) and from $r^2 = .05$ to $r^2 = .12$ (Perry, Moore, & Braff, 1995). He also found that the addition of $R$, $XQUAL$, and the Ego Impairment Index to two MMPI variables led to an increase in validity from $r^2 = .13$ to $r^2 = .24$ (Meyer & Resnick, 1996).

In the Meyer and Resnick (1996) study, Rorschach and MMPI test scores were used to predict level of ego impairment. Ratings of ego impairment were made by the study coauthors based on diagnoses that clients had received by clinicians involved in their treatment. The ratings were made using a 5-point scale. For example, a score of 5 was given to undifferentiated schizophrenia; a score of 4 was given to paranoid schizophrenia; a score of 3 was given to major depression, recurrent, severe nonpsychotic; a score of 2 was given to generalized anxiety disorder; and a score of 1 was given to specific phobia. Meyer and Resnick’s procedure for assessing “ego impairment” assumes that there is a close relation between diagnosis and ego impairment (i.e., that diagnosis $X$ is associated with ego impairment $Y$).

A problem with the Meyer and Resnick ego impairment criteria scores can be described. Level of functioning is variable within diagnostic categories, in part because a mental disorder can be mild, moderate, or severe in intensity. Although Meyer and Resnick’s criterion scale accommodated severity within a diagnostic category, clinicians do not always specify whether a disorder is mild, moderate, or severe in intensity; therefore, it is difficult to know how the criterion judges could have made precise ratings of ego impairment. Thus, some (if not most) of the diagnoses could be used to give only a gross approximation of level of functioning. The ego impairment ratings may have been more valid if they were made by using measures of disability, for example, the Short Form 12 of the Medical Outcomes Study (Ware, Kosinski, & Keller, 1996) or the Brief Disability Questionnaire (Von Korff, Ustun, Ormel, Kaplan, & Simon, 1996).

Overall, results on incremental validity offer negligible support for the use of the Rorschach. As already noted, when clinicians made judgments after being given increasing amounts of assessment information, the addition of the Rorschach did not lead to an.

---

7 Viglione (1999, p. 255) incorrectly reported that 14 participants received the Rorschach. In Table 2 of their article, Cadenhead et al. reported that "twelve subjects received the Rorschach" (p. 856).
comments on Hiller et al. (1999)

Hiller et al. (1999) conducted a meta-analysis on the validity of the Rorschach and the validity of the MMPI. Unweighted mean validity coefficients ($r$s) were .30 for the MMPI and .29 for the Rorschach, whereas mean validity coefficients weighted by degrees of freedom were .37 for the MMPI and .26 for the Rorschach. It is important to point out that major textbooks on meta-analysis (e.g., Shadish & Haddock, 1994; Hedges & Olkin, 1985, chap. 6; Hunter & Schmidt, 1990, chap. 11) do not recommend or even mention calculation of unweighted estimates of effect sizes. Although it can be problematic to weight studies by sample size when underlying population parameters are not homogenous across studies, the proper response is not to calculate unweighted effect sizes but instead to identify and describe moderator effects. Of course, the advantage for calculating weighted mean validity coefficients is that effect sizes based on large samples are weighted more heavily than effect sizes based on small samples. Interestingly, the study (Zimmerman & Dillard, 1994) with the largest effect size for the Rorschach ($r = .91$) had the smallest sample size ($N = 8$). If one believes that Hiller et al. appropriately conducted their meta-analysis (we do not), then one could conclude on the basis of the weighted mean validity coefficients that their results favor the MMPI ($r = .37$ and $r = .26$, for the MMPI and Rorschach, respectively).

In addition to conducting their own meta-analysis, Hiller et al. also described problems with previous meta-analyses. We discuss their criticisms of previous meta-analyses, and then we comment on the limitations of their own meta-analysis.

**Previous Meta-Analyses**

The best known meta-analysis on the Rorschach and the MMPI was conducted by Parker, Hanson, and Hunsley (1988). They found that the Rorschach is as valid as the MMPI. Their meta-analysis has been widely cited to defend the validity of the Rorschach. For example, the following conclusion was reached in a recent report of the Board of Professional Affairs, Psychological Assessment Work Group, American Psychological Association: "Validity coefficients produced by the Rorschach are as large as those found for the MMPI." (Meyer et al., 1998, p. 25). This conclusion was based partly on the results of the Parker et al. (1988) article.

Garb, Florio, and Grove (1998) criticized the Parker et al. meta-analysis for several reasons, the main reason being that $F$-test, $t$-test, and chi-square results were not pooled with correlation results. Textbooks on meta-analysis uniformly recommend that results obtained by calculating different types of statistics be pooled. When Garb et al. re-analyzed the Parker et al. data by pooling all of the effect sizes, they found that validity was better for the MMPI than for the Rorschach (MMPI: $r = .48-.55$ in the confirmatory studies; Rorschach: $r = .28-.36$ in confirmatory studies; MMPI: $r = .1$ in exploratory studies; Rorschach: $r = .1-.17$ in exploratory studies).

Hiller et al. (1999) agreed with the criticisms of the Parker et al. meta-analysis, but they have also criticized the Garb et al. (1998) meta-analysis. Specifically, they criticized Garb et al. for using the effect size estimates that were calculated by Parker et al. Hiller contacted Parker and learned that when calculating effect sizes for studies that reported $F$ test results, $df$ for the numerator was frequently greater than 1. We (Garb, Florio, & Grove, 1998) had not realized that Parker et al. (1988) had made this error.

**Hiller et al. (1999) Meta-Analysis**

Unfortunately, the Hiller et al. (1999) meta-analysis is flawed. Five problems are described: (a) poor interrater reliability, (b) incomplete analysis of reliability, (c) coding judges not blind, (d) mistakes in coding the data, and (e) inappropriate assumptions in aggregating effect sizes.

Poor interrater reliability. Interrater reliability for the coding judges was poor (e.g., $\phi = .35$ for the Rorschach judges and .39 for the MMPI judges). Coding judges were given synopses of the method sections for each study, and they were supposed to indicate whether one could reasonably expect the predictors to be signifi-
cantly related to the criteria. Results were included in the meta-analysis only when judges agreed that one could reasonably expect the predictors to be significantly related to the criteria. Importantly, coding books were not constructed to help the coding judges make their ratings.

An example is given to illustrate the problems with interrater reliability. A study by Perry, Sprock, et al. (1995) addressed the effects of amphetamine administration on Rorschach responses. Two of 14 Rorschach scores were statistically significant. Only one of the statistically significant variables and none of the nonsignificant variables were included in the meta-analysis (leading to a high estimate for average effect size). When asked why the other variables were excluded, Hiller (personal communication, May 12, 1999) could not explain the decision. He indicated that raters did not have to give a reason for including or excluding a variable. For this study, the raters agreed to include one variable, agreed to exclude three variables, and disagreed on four variables, yielding a between-raters kappa of only .16, which indicates a very poor level of agreement. Furthermore, the coding judges never even made ratings for six of the variables (this is discussed in the section on mistakes in coding data).

**Incomplete analysis of reliability.** Instead of calculating Cohen's kappa, Hiller et al. (1999) calculated a phi coefficient. Kappa is asymptotically equivalent to an intraclass correlation (Fleiss & Cohen, 1973), and it appropriately penalizes coders for disagreements and also for differences in acceptance rates. On our inquiry, Jordan Hiller calculated a kappa coefficient for his data. He obtained a value of .34. Though this value is only slightly lower than the value obtained for phi, we still recommend that kappa or an intraclass correlation coefficient be calculated when measuring interrater reliability.

**Coding judges not blind.** In addition to problems with reliability, coding judges were not blind to the results of the studies. As observed by Hiller et al. (1999),

> Even though the inclusion judgments were made blindly with respect to authors' original hypotheses and study results, it is likely that judges were nevertheless familiar with at least some of the studies included. In fact, several studies included in the meta-analysis were authored by [the coding] judges. (p. 292)

Thus, the coding judges' ratings may have been influenced by their knowledge of study results.

**Coding ratings may have been biased not because coding judges were intentionally being dishonest, but because of the hindsight bias.** As observed by Fischhoff (1975), research on the hindsight bias indicates that

> finding out that an outcome has occurred increases its perceived likelihood. Judges are, however, unaware of the effect that outcome knowledge has on their perceptions. Thus, judges tend to believe that this relative inevitability was largely apparent in foresight. (p. 297)

In fact, an examination of the studies used in the meta-analysis reveals that the two judges who coded the Rorschach studies did tend to include positive findings and exclude negative findings (e.g., see Perry, Sprock, et al., 1995). Of course, the hindsight bias may have affected the MMPI judges to the same extent as the Rorschach judges.

**Mistakes in coding the data.** Nonsignificant results were not always given to the coding judges. For example, in a study on the effects of amphetamine administration (Perry, Sprock, et al., 1995), only 8 of 14 Rorschach scores were given to the coding judges. We asked Hiller about this, and he responded as follows:

> The first six variables in [Perry, Sprock, et al.'s] Table 1 (R through PHE) were not submitted to the judges for consideration. This appears to have been an oversight on my part—I must have missed them while reading through the article, possibly because they were not explicitly mentioned by name in the text. I certainly believe this is an isolated error. (Hiller, personal communication, May 13, 1999)

We hope that Hiller is right in saying that this is an isolated error: We are not sure if this is the case, because we made inquiries for only a few of the studies.

**Inappropriate assumptions in aggregating effect sizes.** The Rosenthal and Rubin (1986) procedure aggregates effect sizes within studies by weighting them for importance and allowing for intercorrelations between dependent variables. Hiller et al. did not say what weights they gave to multiple variables from a single study. We assume that if they included them, they weighted them all equally. Hiller et al. also did not state what value for intercorrelations (denoted rho by Rosenthal and Rubin) they assumed for dependent variables (primary study authors do not usually report these correlations). Hiller (personal communication, May 12, 1999) stated the following:

> When no such reference material [test manuals giving a value for rho] was available, I generated an estimate according to how conceptually similar the measures seemed. I tried to err on the side of using values of rho that were too large (I generally chose rho's over .30, at least), because larger intercorrelations lead to somewhat smaller composite effect sizes.

This is both a subjective and sometimes incorrect procedure. First, it is clearly subjective because there is no explicit, replicable method for estimating rho. Second, Hiller (personal communication, May 12, 1999) implicitly acknowledged that his procedure was sometimes incorrect ("I tried to err on the side of... ").

**Methodological Recommendations**

Suggestions are made for resolving the Rorschach debate. Recommendations are made for improving the methodological rigor of individual studies, conducting replication studies, making data accessible, studying incremental validity, and conducting more meaningful meta-analyses.

**Methodological Rigor of Individual Studies**

To resolve the Rorschach debate, the methodological rigor of individual studies must be improved. Four common methodological errors are described. They are as follows: (a) lack of comparison groups, (b) criterion contamination, (c) selective reporting of results, and (d) failure to adjust level of alpha when conducting multiple statistical tests. To support his argument that the Rorschach is valid, Viglione (1999) cited studies that committed all four of these types of errors.

**Lack of control groups.** According to Viglione (1999), inanimate movement (m) and diffuse shading (y) are related to stress and anxiety. He cited five field studies of traumatized patients to support his claim. However, m and y were not even analyzed in
one of the studies (Sloan, Arsenault, Hilsenroth, Handler, & Harvill, 1996). In the other four studies (Hartman et al., 1990; Kasser-Boyd, 1993; Sloan, Arsenault, Hilsenroth, Harvill, & Handler, 1995; Swanson, Blount, & Bruno, 1990), results for traumatized patients were not compared to results from a comparison group of participants. Instead, results for patients compared either implicitly or explicitly to Comprehensive System normative data.

As Exner, Kinder, and Curtiss (1995, p. 151) have pointed out, "one of the most common errors noted in Rorschach research occurs when investigators attempt to use normative data as some control or reference sample." There are two main problems with using the Comprehensive System normative data as a control group. First, group members (e.g., members of an anxiety group) may differ from the Comprehensive System normative group on a number of variables but not the variable of interest (e.g., they may differ in age, occupation, socioeconomic status, and educational level, but not anxiety level). When this occurs, investigators may erroneously conclude that differences in Rorschach scores are due to differences in anxiety rather than to differences in demographic characteristics. Second, differences between an experimental group (e.g., an anxiety group) and a Comprehensive System normative group may due to the different scoring styles (and biases) of different groups of raters. Preferably, a single set of raters would score the Rorschach protocols for both groups.

Criterion contamination. According to Viglione (1999, p. 257), "individuals with borderline personality disorder produce more malevolent, inaccurate, and elaborated human representations." He cited three studies to support his conclusion; one of the studies will be discussed here, and the other two will be discussed in the following section.

Criterion contamination occurs when the same information is used to make assessment ratings and criterion ratings. In a study by Murray (1985), one third of the criterion diagnoses were made by a hospital clinician who attended to psychological test results, including Rorschach results (pp. 457, 463). As Murray acknowledged, "this practice resulted in a contamination of independent and dependent variables" (p. 463).

Selective reporting of results and failure to control alpha level. Viglione (1999, p. 257) also cited two other studies that describe the response patterns on the Rorschach for individuals diagnosed as having borderline personality disorder. These studies are also flawed.

In an article published in Psychiatry, Lerner and St. Peter (1984a) performed 64 main analyses and 246 post hoc comparisons. In an article published in Journal of Personality Assessment, using the exact same data that had been analyzed in their Psychiatry article, Lerner and St. Peter (1984b) reported the results from only 14 of the main analyses. This is the article that was cited by Viglione (1999). There was no warning in the Lerner and St. Peter (1984b) article that 50 parallel analyses had been performed and discarded, nor was any attempt made to control the alpha level for the many statistical tests (310 or more) that had been carried out by the researchers.

In the other study (Stuart et al., 1990), results for 3 of 30 analyses of variance (ANOVAs) were statistically significant, and post hoc analyses were performed. To control for alpha inflation due to conducting multiple statistical tests, Stuart et al. performed a multivariate analysis of variance (MANOVA). However, the use of MANOVA in this way does not necessarily protect against Type I error in subsequent univariate tests (Bray & Maxwell, 1985, pp. 40-41; Miller, 1966). Thus, most of the positive findings found by Stuart et al. may be due to chance.

Replication Studies

Positive results for the Rorschach have frequently not been replicated. As observed by Wood, Nezworski, Stejskal, Garven, and West (1999), some individual CS scores (e.g., the SCZf) have a well-established relationship to schizophrenia, and several others (e.g., R) appear related to intellectual disabilities. Beyond that, however, we cannot identify any CS score with a well-demonstrated relationship to a psychological symptom or diagnosis, according to the three minimal criteria in our "central question" (consistent validity, methodological quality, independent replications). So far, Rorschach proponents have not challenged this negative appraisal by publishing a list of individual CS scores that meet these minimal criteria, along with citations to the relevant scientific literature. However, if such a list is ever published, the Rorschach debate is likely to achieve greater clarity and focus. (p. 117)

Unfortunately, the importance of replicating results has been neglected. For example, Viglione (1999) described a very impressive longitudinal, predictive field study (Rose & Bitter, 1980). The Palo Alto Destructive Content Scale, a scale encompassing general aggression but emphasizing the notion of victim and attack, predicted with a large effect size (Cohen's $d = 1.7$, $p < .002$) reoffense after release from prison in a group of rapists and murderers. (p. 258)

Though Viglione (1999) praised the Rose and Bitter (1980) study, it is important to note that the findings have never been replicated even though the study was published nearly 2 decades ago. Replication is especially important for this study because the Rorschach was used to make only a small number of predictions: Predictions were made for only 23 participants (11 rapists who reoffended and 12 rapists and murderers who did not reoffend). In comparison, non-Rorschach studies of criminal recidivism often include hundreds of prisoners (Quinsey, Harris, Rice, & Cormier, 1998, pp. 27-42).

Incremental Validity

In an effort to resolve the debate over the Rorschach, it will be important to establish a body of research on incremental validity. For example, non-Rorschach predictors of criminal recidivism have well-established validity (Quinsey et al., 1998). Before using a Rorschach measure to predict recidivism, one will want to know if it adds anything to already established predictors, and one will want to make sure that the addition of the Rorschach measure does not lead to a decrease in validity.

Meta-Analyses

In an effort to resolve the Rorschach debate, it will also be important to conduct well-designed meta-analyses. Two recommendations can be made. First, unpublished studies should be included. Meta-analyses on the Rorschach have never included a random, let alone comprehensive, sample of unpublished studies.
This could be done by including dissertations. Hiller et al. (1999) argued that results are unlikely to change if unpublished studies are included, but when unpublished and published studies have been compared in other areas, published studies have been more likely to report significant findings (e.g., H. Cooper, DeNeve, & Charlton, 1997).

A second recommendation is that all of the predictors in a study be coded and included in the meta-analysis. Hiller et al. (1999) did not use this approach. Instead, they had expert clinicians decide which predictors should be included. Of course, in the Hiller et al. study, interrater reliability was poor, and judges were not really blind to the results of the studies (being familiar with the research literature, they knew which predictors have been found to be significant for a particular task). If all of the predictors in a study are coded, as we recommend, then one can try to learn what predictors have been valid for particular tasks. When conducting a meta-analysis, one should be able to determine if positive findings have been replicated for particular predictors.

Unwritten Articles and the Accessibility of Data

There has been a heated argument over the accessibility of studies that have been cited to support the Rorschach. The Rorschach controversy will not be resolved until both sides agree on how information should be made available for independent inspection.

Unpublished studies sponsored by the Rorschach Workshops are frequently cited as supporting the Comprehensive System (e.g., Exner, 1991, 1993). Unfortunately, attempts to obtain copies of the studies have often been unsuccessful (Nezworski & Wood, 1995; Wood et al., 1996a, 1996b; but also see Exner, 1995, 1996). Attempts to obtain copies of the studies were described in an earlier article (Wood et al., 1996a):

Many readers of TRACS [The Rorschach: A Comprehensive System] are probably under the impression that the Workshops Studies are actual documents that can be examined by other scholars. However, this impression is often mistaken. In preparation for writing this article, we requested 23 of the Workshops Studies cited in TRACS. Letters from the Rorschach Workshops informed us that some of the Workshops Studies were not in their files. The methods and results of the remaining studies either had not been formally written or could not be released. We were informed that the Rorschach Workshops could provide raw data related to specific questions, but that we might have to pay for computer costs. (p. 8)

Because Method and Results sections of papers could not be obtained, there seemed to be little point in paying for the data. Exner (1995) has presented his view of the controversy:

They requested copies of more than 20 studies and were informed that, while most were not written in a form that could be released, they could obtain statements concerning the designs if those in various volumes were not clear, plus raw data matrices, plus any tables relating to the data analyses still available, or the analyses could be regenerated provided they would assume the cost of downloading the data from a mainframe computer and/or the costs of any technician time. (p. 205)

Several comments can be made about Exner's argument. First, Rorschach Workshops did say that data (and presumably data matrices) could be obtained, but only at the expense of the person requesting the data. Second, they did not offer tables relating to the data analyses, and in fact they did not send tables when Wood and his colleagues asked for information about 23 of the unpublished studies. Also, in the letters that were written by Rorschach Workshops to Wood and his colleagues, no offer was made to provide statements describing the designs of studies though the letter mentioned that "brief statements" existed. According to Patricia Greene, Administrative Assistant of the Rorschach Workshops (in a letter to James Wood dated March 11, 1994):

Dr. Exner has asked that I respond to your letter of February 5 requesting copies of some of our unpublished works. During the period from 1968 to 1990 more than 1000 studies were undertaken at Rorschach Workshops to address various issues. The majority of these are not written in a publishable form. Instead, they usually include a brief statement concerning the methodology of the study. (p. 1)

Even if a brief statement, concerning the methodology of the study, was made available, this would be inadequate. To evaluate a study, one needs more than a brief statement about the Method section.

In conclusion, it is important that research papers and data be made accessible. Certainly, if unpublished studies demonstrated that the Rorschach is invalid, Rorschach advocates would be unhappy if they could not obtain copies of the Method and Results sections for the studies. The accessibility of research papers and data continues to be an important issue: A recent request for a copy of the Comprehensive System normative data was rejected by J. E. Exner (personal communication, December 8, 2000), even though J. M. Wood offered to pay for any expenses that this would entail (J. M. Wood, personal communication, August 5, 2000; see footnote 9).

Discussion

One of the coauthors of this article recently called for a moratorium on the use of the Rorschach Inkblot Test (Garb, 1999). The moratorium is to extend to clinical and forensic settings, but not to research settings. The call for a moratorium was issued not because all Rorschach scores are invalid, but because we do not know which Rorschach scores are valid and which ones are invalid. Until this changes, clinicians will continue to interpret invalid scores along with valid scores. In general, a test should not be used if its use leads to the interpretation of invalid scores along with valid scores.10

When deciding whether an assessment instrument should be used, a number of issues need to be addressed. First, the assessment instrument should be supported by methodologically sound studies. Methodological problems described in this article include the following: (a) the use of inappropriate control groups (e.g., using Comprehensive System normative data as a control group), (b) criterion contamination, (c) selective reporting of results, and (d) failure to control alpha level. Methodological problems with meta-analyses were also described. Also, when unpublished stud-

---

9 On request, we will provide copies of all correspondence.

10 Criteria for evaluating validity are described in the Standards for Educational and Psychological Testing (American Psychological Association, 1999).
ies are cited to support an assessment instrument, written manuscripts should be available so that the Method and Results sections of the studies can be evaluated. Second, validity needs to be established. Validity can be broadly defined to include not only studies on the relation between test scores and construct measures, but also studies on inter scorer reliability, test-retest reliability, bias (e.g., whether use of the instrument leads to race bias), and the validity of norms. Positive findings need to be replicated by independent investigators. The issue of incremental validity is also important; for example, a test should not be used if its use leads to a decrease in the validity of inferences and judgments.

To end on a positive note, we describe Rorschach scores that have been supported by empirical research. First, deviant verbalizations on the Rorschach are related to thought disorder and schizophrenia, and they are probably also related to bipolar disorder as well as to schizotypal and borderline personality disorders (Wood, Lilienfeld, Garb, & Nezworski, 2000). However, it is not known whether Rorschach thought disorder measures can be used to improve the accuracy of psychiatric diagnoses in clinical work, beyond what can be learned from a standard interview and a self-report instrument such as the MMPI. Second, empirical research generally supports the validity of the Rorschach Prognostic Rating Scale (Meyer & Handler, 1997). However, most studies on this topic have serious methodological limitations and are old (Hunsley & Bailey, 1999). Furthermore, the Rorschach Prognostic Rating Scale depends on complicated scoring rules and lacks current norms. Future research on the Rorschach Prognostic Rating Scale will need to show that it is more valid than other measures that are less time consuming to use and that have also been supported by validity research (e.g., Leon, Kopta, Howard, & Lutz, 1999; Luborsky et al., 1979). Finally, research on the Rorschach Oral Dependency Scale is encouraging (Bornstein, 1999). However, though positive results have been found for the Rorschach Oral Dependency Scale, their clinical relevance is unclear. Bornstein (1999) found that, “Projective test scores were positively correlated with behavioral ratings obtained in laboratory, field, and classroom settings but not in clinical settings” (italics added)” (p. 52). In sum, positive and encouraging results have been found for some Rorschach scales, but to a surprising degree little is known about the validity of many Rorschach scores that are used in clinical and forensic practice.

References


Joyce, P. R. (1984). Age of onset in bipolar affective disorder and misdiagnosis as schizophrenia. Psychological Medicine, 14, 145–149.


Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict


### Appendix A

**Comprehensive System Variables Whose Test–Retest Coefficients Are Listed by Exner**


### Appendix B

**Comprehensive System Variables Whose Test–Retest Coefficients Are Not Listed by Exner**

- SCZ, DEPI, CDI, HVI, S-Con, OBS, Pure H, H: (H) + Hd + (Hd), H + (Hd); (A) + (Ad), FC: CF + C, S−%, F+, F−%, a, p, M 5, M 6, EB, EB

Per, W, Dd, Blends; R, W:M, An + Xy, CONTAM, PER, PSV, CP, Hx, An + Xy, Fd, Most contents

Received May 24, 1999

Revision received May 30, 2001

Accepted June 25, 2001