Psychological Assessment
1997, Vol. 9, No. 4, 480–489

Copyright 1997 by the American Psychological Association, Inc.

1040-3590/97/$3.00

Assessing Reliability: Critical Corrections for a Critical Examination of the Rorschach Comprehensive System

Gregory J. Meyer
University of Alaska Anchorage

Wood, Nezworski, and Stejskal (1996a, 1996b) argued that the Rorschach Comprehensive System (CS) lacked many essential pieces of reliability data and that the available evidence indicated that scoring reliability may be little better than chance. Contrary to their assertions, the author suggests why rater agreement should focus on responses rather than summary scores, how field reliability moves away from testing CS scoring principles, and how no psychometric distinction exists between a percentage correct and a percentage agreement index. Also, after reviewing problematic qualities of kappa, a meta-analysis of published data is presented indicating that the CS has excellent chance-corrected interrater reliability (Estimated κ, M = .86, range = .72–.96). Finally, the author notes that Wood et al. ignored at least 17 CS studies of test–retest reliability that contain many of the important data they said were missing. The author concluded that Wood et al.'s erroneous assertions about the more elementary topic of reliability make suspect their assertions about the more complex topic of validity.

The interchange between Wood, Nezworski, and Stejskal (1996a, 1996b) and Exner (1996) concerning the Rorschach Comprehensive System (CS) may have left some readers wondering where the truth resides between their opposing positions. Alternatively, Wood et al.'s critique may have solidified a suspicion that the Rorschach, even after Exner's extensive efforts, still exemplifies the kind of fuzzy-headed, error-plagued, magical thinking that has set clinical psychology too far adrift from its scientific moorings (e.g., Dawes, 1994).

Wood et al. (1996a, 1996b) addressed both the reliability and the validity of the CS. Although validity is of central importance, they considered reliability to be a fragile beam within the CS, one that ultimately could cause the psychometric infrastructure of the procedure to collapse, if it proved to be as weak as they proposed. Reliability is a less complex and more straightforward issue to consider than validity, and therefore it is the focus of this article.

Wood, Nezworski, and Stejskal's Assertions Regarding Reliability

Wood et al. (1996a, 1996b) made numerous claims about CS reliability. They stated that the typical statistic used to quantify rater agreement, the percentage agreement index, was "unacceptable," "inadequate," "misleading," and "inflated" as a measure of scoring accuracy (p. 4). They also stated that critical components of CS reliability had not been investigated or published. These perceived deficiencies included (a) a lack of evidence indicating that the CS could be reliably used in applied settings (1996a, pp. 4–5); (b) a failure to study errors that may be associated with Rorschach administration, response transcription, or examiner characteristics (1996a, p. 5); and (c) a failure to report "the most clinically relevant" data, which they believe is the reliability of total scores and indices summarized across all responses (1996a, p. 4; 1996b, pp. 14–15). Furthermore, the only actual data that they presented regarding CS reliability were used to support their belief that response scoring may be nothing more than a chance process. They also asserted that a "percentage correct" index was fundamentally different from a "percentage agreement" index, making Exner's (1993) two studies irrelevant for reliability purposes (1996b, p. 14).

Ultimately, Wood et al. (1996a, 1996b) concluded that fundamental issues regarding reliability had not been resolved (1996a, p. 9); that it was false to believe that the CS had high interrater reliability (1996a, p. 9); and that this state of affairs, in conjunction with ethical guidelines and practice standards, warranted a moratorium on the use of the CS (1996a, p. 9; 1996b, pp. 16–17). Although the last recommendation would have been warranted if all their assertions were true, their assertions are incorrect. This article considers each of the points they raise.

Should Interrater Reliability Focus on Each Response or on Final Summary Scores?

Wood et al. (1996a) considered the CS to be in violation of test standards because Exner did not report interrater reliability statistics for the final summary scores in a protocol. Certainly it would be valuable to have more of this information. However, there are no scoring rules applied at this level of analysis. Because the goal of interrater reliability is to demonstrate that the CS provides users with a systematic and consistent way of translating the complex language and imagery from patients into the scores of the system, and because these scores are assigned

I thank Nancy Kaser-Boyd, Glenn Curtiss, Steve Finn, Bill Kinder, John Petraitis, and Bruce Smith for their helpful comments.

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

to each and every response, it is of utmost importance to have rater agreement calculated at the level of individual responses. In fact, if Exner had ever suggested that rater reliability should be based on just the summary scores, he would have been roundly and appropriately criticized.

Summary scores are simple aggregations that are derived by summing across all the responses in a protocol. Because random errors of measurement tend to cancel with aggregation, psychometric theory predicts that summary scores will be more reliable than individual responses. In fact, data clearly demonstrate that this is so (see McDowell & Acklin, 1996). Furthermore, summary scores could appear to be quite reliable despite poor response-by-response reliability. For instance, two raters could agree on the same summary value even though they may have assigned the actual scores to completely different responses. Because the converse is never true, the typical CS procedure for calculating agreement on each and every response is both a more exacting form of reliability and a more specific test of CS scoring clarity.

Interrater Reliability in Field Settings

Wood et al. (1996a) also indicated that field reliability is an essential but missing piece of data regarding the CS. *Field reliability* refers to scoring agreement that is obtained by clinicians "performing under the time constraints and conditions of their work" (p. 4). I echo Wood et al.'s call for more systematic research into this area.[1] However, two points should be kept in mind. First, most Rorschach research is conducted by psychologists who actually work in clinical settings. Thus, most of the reliability data published in conjunction with CS validity research have already been generated by psychologists working under the time constraints and other day-to-day pressures that are part of their work.

Second, it is important to remember that systematic research into field reliability would focus attention on issues that are downstream from the clarity of CS scoring rules. Such studies would not directly assess the consistency with which CS principles can be implemented, because there are many factors that could make field reliability poor even though they have nothing at all to do with the CS scoring rules. For instance, poor field reliability could result from training programs' not committing sufficient time and energy to the proper use of this complex instrument or from clinicians' neglecting scoring rigor because they have a general disregard for the procedure even though they feel compelled to offer it to referral sources. Neither of these factors has anything to do with the intrinsic rules that guide CS scoring, although they would certainly compromise field reliability. In essence, little could be learned about the inherent consistency of CS scoring principles if one were examining results obtained from poorly trained clinicians who had scored the CS in a sloppy manner.

Percentage Agreement Versus Percentage Correct

Wood et al. (1996b) accurately noted that Exner's "percentage correct" index is different from a "percentage agreement" index. However, the distinction is not at the psychometric level. Percentage correct is a way to determine scoring discrepancies when a "gold standard" is available or when one rater's scores are designated as a benchmark. Percentage agreement is a more democratic procedure, used when it is impossible or undesirable to say one individual's scoring is better than another's. However, the process of calculating a discrepancy index is the same in each case; a discrepancy exists regardless of whether it is between a rater and a standard or between a rater and another rater. The distinction then resides at the inferential level, where it means something slightly different to diverge from a designated standard than from a potentially erroneous peer.

Interrater Reliability, Percentage Agreement, and Kappa

Regarding interrater reliability, Wood et al. (1996a) accurately noted that Exner (1986, 1993) provided little detail on the method he used to calculate agreement in his two studies. For other studies using multiple raters, this information was provided. However, on some occasions Exner used a stringent procedure reflecting the percentage of responses when all raters were in unanimous agreement; on other occasions, he used a more typical procedure that calculated the percentage of responses for which two raters were in agreement.[2]

Wood et al. (1996a) also noted that the percentage agreement index can be a deficient measure of reliability because it does not correct observed agreement (A_o) for chance levels of agreement (A_c). This can be problematic when raters know that the base rate for a score is very high or very low. Under these conditions, considerable agreement can be obtained when the raters simply make random guesses that parallel the base rate. In the following excerpt, Wood et al. demonstrated the problem by using the inanimate movement variable, *m*, which has a base rate of approximately .05 in a nonpatient population:

> Imagine that two raters independently rate a large number of Rorschach protocols and randomly assign a score of *m* to 5% of responses. Even though the two raters score at random, they will agree that *m* is present in about .0025 (.05 × .05) of responses and absent in about .9025 (.95 × .95). By chance alone, therefore, the total percentage of agreement between the two raters will be .9050 (.0025 + .9025). (1996a, p. 4)

Wood et al. proceeded to point out how Exner's (1993) agreement rates for *m* are only minimally larger than this chance value (.93 in one study and .95 in the second), supporting their belief that CS scoring may be little more than randomness.

The example given by Wood et al. (1996a) derives its assumptions about chance agreement from Cohen's kappa, a measure of interrater reliability designed to indicate how much true

[1] In fact, as one reviewer also noted, it would be helpful to have this information for all the major performance tests in psychology (e.g., the Rorschach, the Wechsler scales, the Halstead-Reitan battery, etc.). Currently, no one knows the extent to which the average clinician working alone administers and scores these measures in an accurate and reliable fashion.

[2] In fact, in my initial effort to make sense of the Wood et al. suggestion that agreement may be no better than chance, I began to suspect that Exner's studies on scoring agreement must have used the former method. However, they actually used the latter (personal communication, J. E. Exner, Jr., May 16, 1996).

agreement occurs beyond chance levels (Cohen, 1960). Although kappa is the most frequently used index of nominal agreement, statisticians continue to disagree about the value of some of its properties. Before evaluating CS reliability with kappa, it is worthwhile to note briefly these points of contention. Two primary issues have been debated. The first concerns kappa's definition of chance, and the second concerns its sensitivity to base rates.

Defining Chance Agreement

Under the assumptions of kappa, *chance* is defined as the level of agreement that would be observed if raters had known a base rate for the phenomenon under study and randomly assigned judgments in line with that base rate. This definition of chance has been referred to as a "fixed marginals" model because the marginal distributions of category assignment are assumed to be known a priori (Brennan & Prediger, 1981). Making this assumption justifies the procedure of determining chance by multiplying the corresponding row and column marginals for each rater and then summing these products (as was numerically demonstrated in the quotation provided earlier).

A problem with this approach is that it does not give raters credit for judgments that are independently agreed on and that actually produce the marginal distributions of scores. For instance, raters could initially expect that an *m* score is as likely to be present as absent on each response in a sample. However, their independent, expert judgments may determine that in fact *m* is very rare, being present in just 5% of the responses. Kappa gives the raters no credit for the parallel reasoning and agreement process that would be inherent in establishing this observed base rate. Instead, kappa "penalizes" the raters by using the extreme base rate that was independently agreed on to now define the chance agreement level the raters must surpass (Brennan & Prediger, 1981; Zwick, 1988). This is done under the presumption that such a chance rate could have been obtained had the raters known the marginal base rates beforehand.

The tenuousness of kappa's assumptions about chance are easier to appreciate when one considers what raters would actually have to do in order to assign CS scores randomly and in a manner that approximated actual data. First, two raters would have to know a relatively similar base rate for each of the CS scores. Thus, they would need to know what kind of sample was being rated and would have to look up, be told, intuit, or rely on their historical experience with similar samples to determine an appropriate value. This is not done with just one or two scores, however. Rather, to code the CS they must retain in working memory a base rate for 85 separate score options (excluding various *Z*-score values). Furthermore, if we assume that reliability will be calculated across 20 Rorschach protocols, for example, the raters will have to code approximately 450 responses. Because each response is scored as a unit before moving on to the next response, the raters must make a decision for each of the 85 options on each response before proceeding. Therefore, to "randomly" assign scores in line with the rules of chance assignment, raters must mentally keep track of how often they are assigning 85 scores, with 85 unique expected frequencies, across 450 responses. When the raters have finished scoring these 450 responses, the rate at which they assigned each of the 85 score options must equal the rate they initially anticipated. Although a computer can easily handle such a task, I doubt that it is something human raters could actually accomplish. Thus, in practice, I believe it would be impossible for raters to "randomly" score the CS according to the dictates of kappa-defined chance.

Over the years, several alternative definitions of chance have been proposed (e.g., Brennan & Prediger, 1981; Cohen, 1960; Zwick, 1988), although each of these contains its own slightly problematic assumptions. The most common alternative is what Brennan and Prediger (1981) have called the "free marginals" model, quantified by the statistic kappa$_n$. Under this model, chance is defined as what raters would do when they have no a priori knowledge about the sample or the characteristics to be rated. Under these conditions, judges would blindly and randomly assign objects to scoring categories. Through mathematical derivation, chance agreement is found to be $1/n$, where n is the number of options in the rating scale (Brennan & Prediger, 1981). In the example given earlier for the *m* score, the options consist of "present" or "absent." Thus, $n = 2$ and chance agreement would be .5, not .905 as indicated by the definition of chance in Cohen's kappa.

A family of kappa-like statistics for determining chance-corrected agreement rates can be defined as having the form $A_o - A_c/(1 - A_c)$, in which A_o is observed agreement and A_c is chance agreement. Given this, one can show that when observed agreement for the CS *m* score is .93 and the base rate for the *m* score is .05, the fixed marginal assumptions of Cohen's kappa would lead to a kappa value of .26 (.93 − .905/[1 − .905]). However, the free marginal assumptions of kappa$_n$ would lead to a kappa value of .86 (.93 − .50/[1 − .50]). Obviously, very different chance-corrected values are obtained depending on how chance is defined.

The Issue of Base Rates

The second quality of kappa that has been debated in the literature is kappa's sensitivity to the base rate of a phenomenon. Specifically, as a base rate moves away from the point of maximum variance (i.e., .50), the same small degree of disagreement among raters will cause kappa values to decline rather dramatically. Several influential statisticians have argued that this is as it should be because kappa is a true reliability statistic (e.g., Bartko, 1991; Shrout, Spitzer, & Fleiss, 1987) and the formulas derived from classical true score theory estimate reliability by testing a group of people in order to determine the ratio of true score variance to observed score variance. As true score variance in the group becomes more restricted, a fixed level of disagreement (i.e., error variance) plays an increasingly large role in observed score variance, so calculated reliability coefficients must decline.

Other statisticians have argued that between-subject variability is not an inherent requirement of classical true-score theory (Traub, 1994). In fact, a basic conceptualization of reliability refers to the consistency of observed scores obtained over repeated independent measurements of a single individual. Because a person's true score is defined as the mean of his or her observed scores across independent measurements, reliability is evident when there is little deviation (variance) in observed scores across repeated measurements. This individualized conceptualization of reliability does not impose any requirements about one person's true score differing from another's. Rather, if repeated independent testing indicates that little error creeps into observed scores, then the test should be considered reliable because it is providing highly consistent information. In particular, this conceptualization of reliability has been advanced by those interested in criterion-referenced testing, as opposed to norm-referenced testing. Within criterion-referenced testing (as with the assignment of Rorschach scores), what is at stake is whether a person exhibits a carefully defined characteristic, regardless of whether someone else also exhibits this characteristic (Traub, 1994). Thus, unlike the norm-referenced approach to testing, the criterion-referenced approach indicates that reliability coefficients should not be influenced by variance in subject characteristics.[3]

An example will illustrate this distinction. Consider the reliability of decision rules for classifying animals from the phylum *chordata* into the class *mammalia.* Two raters are given the rules that each animal must have a notchcord and must nurse its young. Assume that the raters are experts and that they each correctly classify the same sample of 100 species as being mammals. If the clarity and consistency of the scoring rules were indexed by the percentage agreement index, consistency would be 100%. However, if interrater reliability were indexed by the kappa coefficient, observed consistency would be kappa = 0.0 because there is no variance in classification decisions. Thus, from a norm-referenced perspective, all agreement would be attributed to chance. One would be forced to conclude that the supposed experts assigned the *mammalia* designation randomly and that the scoring criteria were highly problematic or unclear.[4] If the reliability study is next repeated on a sample of 100 species drawn from the class *osteichthyes* (bony fish) and this time the raters unanimously agree that each species is an instance of "not-mammalia," percentage agreement again would be 100%, whereas kappa would be 0.0. If kappa were viewed as the more accurate index of agreement, one would have to again conclude that it is impossible for experts to consistently classify animals on the basis of the mammalia criteria.

The reason for these results is the absence of variance in each sample. The base rate is 1.00 in the first study and .00 in the second. In the absence of variance, kappa attributes all agreement to chance agreement because it assumes that no real judgment was involved in establishing these base rates. Instead, kappa assumes that the raters could have somehow known these base rates at the outset of the study and then used this knowledge to assign classifications. Only when both samples are combined and evaluated simultaneously would variance be present. Then, kappa would become 1.00 and reflect the true clarity of the criteria and the expert judgments being used by the raters. However, requiring sample variance for reliability contradicts fundamental assumptions for criterion-referenced testing.

Grove, Andreasen, McDonald-Scott, Keller, and Shapiro (1981) presented more realistic data on kappa's sensitivity to the variance contained within a sample (see also Jones, Johnson, Butler, & Main, 1983). Using hypothetical test scores with fixed validity parameters, they determined how base rates affect the size of kappa coefficients. They began by postulating a test with 95% accuracy identifying a condition as "present" when it is truly present (i.e., sensitivity) and 95% accuracy identifying a condition as "absent" when it is truly absent (i.e., specificity). They determined that the maximum kappa value for this test will be a quite respectable .81 when the base rate for a condition is .50 (maximum variability). When the base rate drops to .25, the maximum kappa value will slide a bit to .76. However, when the base rate drops to .01, the maximum reliability coefficient for this test will be .14. Thus, even though the test will be quite valid for diagnostic purposes, the restricted variance (i.e., the extreme base rate) will make it impossible for kappa to demonstrate anything other than weak reliability. As a consequence, Grove et al. recommended that kappa should probably not be calculated for low base rate phenomena. Similar concerns led Bartko and Carpenter (1976) to recommend that researchers revert to the simple percentage agreement statistic as a way to describe the extent of genuine agreement between raters when they are assessing rare phenomena. Perhaps it is not surprising that this is what Exner did when faced with such circumstances.

Meta-Analysis of Published Interrater Reliability

Despite the limitations associated with kappa, it is important to evaluate CS reliability with this statistic because it is commonly regarded as a gold standard. The published literature is optimal for this analysis for two reasons. First, it contains data from many independent investigators, most of whom collected data under typical "field" conditions. Second, the standard procedure for determining agreement is not subject to the same kind of percentage agreement confounds that were present in Exner's (1986, 1993) two demonstration studies on rater consistency. Unless a researcher has hypotheses that focus on just a few variables, the standard approach is to calculate agreement between two independent raters across segments of scores that can be assigned to each response (see Exner, 1991; Weiner, 1991). Achieving agreement on the joint occurrence of many scores within a segment (e.g., all determinants) is very different from achieving agreement on a single CS scoring option that has a low base rate. Consider the following example.

[3] That different conclusions can be derived from the individual and group models of classical true score theory is not unique to this set of circumstances. For instance, similar disparities exist in estimating true scores or confidence intervals about true scores. From an individualized, repeated-measurement conceptualization of reliability, a person's observed score is, by definition, an unbiased estimate of his or her true score. However, a group-derived conceptualization of reliability indicates that observed scores are biased estimates of true scores. Specifically, an observed score above the test mean is an overestimate of the true score, whereas an observed score below the mean is an underestimate (see Feldt & Brennan, 1989).

[4] As an aside, it is interesting to consider what could happen if a "field reliability" study were undertaken using raters with less expertise. Had even one or two of the species been drawn from a nonprototypic order within the class mammalia, such as *monotremata* (enchidas and the platypus), *cetacea* (dolphins, porpoises, whales), *pholidota* (pangolins), or *sirenia* (dugongs and manatees), it is possible that both raters would have erroneously assigned a "not-mammal" designation to these animals. If this were the case, a lack of expertise could result in erroneous classifications being made by both raters. However, this would introduce variance into the ratings and lead to a perfect kappa value of 1.0.

Under the fixed-marginals model of chance contained within Cohen's kappa, it was shown that chance agreement would be .9050 for the *m* score when the base rate for this score is .05. This is analogous to the probability that two people will both pick a marble of the same color out of an infinite urn with 95% white marbles and 5% red marbles. However, when the unit of analysis shifts from exact agreement on the *m* score to exact agreement on all determinants, both raters must unanimously agree on the presence and absence of 10 scores in every response: inanimate movement (3 options),[5] human movement (3 options), animal movement (3 options), color (5 options), achromatic color (4 options), diffuse shading (4 options), texture (4 options), vista (4 options), form dimension (2 options), and reflections (3 options).[6] Because each kind of determinant score is assigned independently, the task is analogous to picking from 10 infinite urns, each of which contains marbles of more than one color that occur at a set frequency. Using fixed marginal assumptions, chance is now defined as the probability that two people will draw exactly matching marbles from all 10 urns. Even when each urn has one very common color (analogous to the "not present" option for many determinant scores), it becomes unlikely that two people will coincidentally pull matching marbles from all 10 urns. To determine the probability of chance agreement under these conditions, one must sum the squared base rates for each marble color within each urn and then multiply the sums from each of the 10 urns. The analogous process for the Rorschach would be to sum the squared base rates for each of the options within a score and then multiply the resulting sums across all scores in the target category (i.e., $[p(Ma)^2 + p(Mp)^2 + p(\text{no-}M)^2] \times [p(ma)^2 + p(mp)^2 + p(\text{no-}m)^2] \times [p(FMa)^2 + p(FMp)^2 + p(\text{no-}FM)^2] \times [p(FC)^2 + p(CF)^2$ . . . , where *a* refers to active movement, *p* to passive movement, *M* to human movement, *FM* to animal movement, *FC* to form-dominated color, and *CF* to form-secondary color responses.)[7]

Conceptually, this shift in the unit of agreement has an important implication. When the target for agreement is an aggregated category that requires unanimous agreement across a number of different "lower level" scores, the target category simultaneously evaluates the reliability of the scoring principles for each of the lower level scores. This simultaneous evaluation results because a disagreement for any single scoring option will cause a disagreement in the aggregated target segment. A limitation of this design, however, is that it will not pinpoint problematic scoring rules if they exist.

It should be noted that these procedures for determining agreement count the assignment of true positive and true negative scores as instances of agreement, whereas the assignment of false positive or false negative scores are considered disagreements. In contrast to Wood et al.'s suggestion (1996b, p. 14), all potential scoring errors are considered in this procedure.

To evaluate the interrater reliability of the CS, I reviewed every article in the *Journal of Personality Assessment* from the start of 1992 through the end of 1995. Of 32 studies using CS variables, 26 reported response-level percentage agreement indices, 2 reported response-level kappa coefficients, 2 reported intraclass correlations for summary scores, and 3 did not report specific reliability scores.

Of the 26 studies that reported percentage agreement, 9 reported agreement on single scoring options (weighted mean agreement = 96.6%; calculated on approximately 5,157 responses), so they were excluded from further consideration. Another study (Frueh, Leverett, & Kinder, 1995) used some reliability data that had been counted in an earlier report, so it was also excluded. For the remaining 16 studies, whenever it was not clear from the article, I contacted the authors to verify that agreement rates reflected unanimous agreement across all scoring options within a target category (i.e., response segment). Whenever possible, I also obtained specific agreement values if the study initially reported the range across categories.

Table 1 presents a summary of these 16 studies. The columns indicate the type of sample that was used in the research, the number of responses independently coded by raters, the number of participants contributing responses, and the level of agreement found for each of 10 response segments (the last of which is a combination of the previous 2). The rows in the top section indicate observed values for each study followed by the average percentage agreement coefficient, weighted by sample size, as well as the total number of responses used to generate this statistic. The second section of the table provides indications of chance agreement. Because base rates can change as a function of clinical characteristics (e.g., disrupted thinking is more common in schizophrenic patients than in nonpatients) and because chance agreement for Cohen's kappa is determined by base rates, estimated rates of chance agreement were calculated for five distinct types of samples.[8] The five samples consisted of Exner's (1993) inpatients with schizophrenia (N = 320), my own heterogeneous sample of psychiatric inpatients and outpatients (N = 442), Exner's (1993) inpatients with depression (N = 315), Exner's (1993) sample of outpatients beginning treatment for the first time (N = 440), and Exner's (1993) nonpatient norms (N = 700). The rows for the five samples are followed by chance agreement rates for each score category using the definition of chance provided by the kappa$_n$ statistic.

[5] The three options are "none," "passive," or "active." Each movement score was considered as having these three options in order to eliminate the statistical dependence that would result from considering the active–passive scores separately.

[6] To be conservative, pure form was omitted from this list of determinants and from all calculations in Table 1. Even though it is theoretically possible to assign pure form with any of the other determinant scores, it is such a rare occurrence that including it would have artificially inflated estimates of independent scoring decisions. In turn, this would have led to underestimates of chance agreement and overestimates of kappa.

[7] In actual practice, the calculations would be slightly different. Rather than squaring a fixed base rate for each score option, the base rates assigned by both of the independent raters would be multiplied. For instance, designating the two raters as *R1* and *R2*, the calculation for the location response segment would be {([R1-$p(W)$ * R2-$p(W)$] + [R1-$p(D)$ * R2-$p(D)$] + [R1-$p(Dd)$ * R2-$p(Dd)$]) * ([R1-$p(S)$ * R2-$p(S)$] + [R1-$p(\text{no-}S)$ * R2-$p(\text{no-}S)$])}, where $p(x)$ refers to the proportion of responses that receive the score option, *W* refers to whole responses, *D* to common details, *Dd* to rare details, and *S* to the incorporation of white space. Note also that *W*, *D*, and *Dd* are mutually exclusive options within a score, whereas *S* and no-*S* are mutually exclusive options within a separate, independently assigned score.

[8] Tables of these calculations are available by writing to me.

Table 1

Interrater Reliability for the Rorschach Comprehensive System: 1992–1995

				Response segments									
Study/summary calculations	Sample type	Rel *N*	Part *N*	Location and space	Developmental quality	Determinants	Form quality	Pair responses	Content[a]	Popular responses	Cognitive special scores	Other special scores	All special scores
					Observed values								
Meloy, 1992	OP	63	1		.98	.83	1.0		.87				.87
Exner & Andronikof-Sanglade, 1992[b,c]	OP	2,100	105	.98	.98	.90	.92	.96	.94	.96			.86
Silberg & Armstrong, 1992[b,c,d]	MP	200	10	.95	.95	.91	.91	.98		.97			.94
Yama, Call, & Entezari, 1993	NP	3,060	102	.99			.94						
Meyer, 1993	MP	312	12	.94	.88	.74	.82	.95	.85	.97	.82	.88	
Singer & Brabender, 1993[b]	MP	320	16	.93	.86	.86	.85				.87	.84	
Kaser-Boyd, 1993	OP	396	22	.92	.95	.83	.90	.97	.90	.98			.83
Berg, Packer, & Nunno, 1993	MP	500	20								.89		
Meloy, Gacono, & Kenney, 1994	MP	540	18	.99	.99	.90	.99		.97				.95
Holaday & Whittenberg, 1994[b,e]	OP	500	25	.98	.94		.92						
Frueh & Kinder, 1994[b,c,d]	OP	360	20	.82	.82	.88	.87	.94	.88	.95			.85
Abraham, Lepisto, Lewis, Schultz, & Finkelberg, 1994[b]	MP	400	20	.98	.94	.96	.94				.83	.88	
Ganellen, 1994[d]	MP	368	16	.96	.88	.78	.88	1.00	.92	.98	.96	.92	
Perry, McDougall, & Viglione, 1995[b,d]	OP	340	17	.96		.86	.94				.88		
Sloan, Arsinault, Hilsenroth, Harvill, & Handler, 1995[f]	OP	460	20	.88	.88	.84	.88		.88				.93
Adrian & Kaser-Boyd, 1995[b,d]	MP	500	25	.92	.94	.84	.86	.94	.92	.96			.82
Weighted mean percentage agreement				.96	.94	.87	.92	.96	.92	.96	.88	.88	.87
Total number of responses				9,856	6,519	6,359	9,919	4,236	5,099	4,236	2,240	1,400	4,619
					Expected chance values and final calculations								
κ-defined chance agreement													
Inpatient schizophrenic				.28	.49	.19	.33	.53	.18	.68	.41	.64	.26
Mixed psychiatric				.29	.46	.15	.34	.55	.17	.66	.51	.58	.30
Inpatient depressive				.30	.46	.16	.37	.58	.16	.64	.63	.64	.40
Outpatient				.32	.47	.25	.42	.59	.20	.59	.69	.62	.43
Nonpatient				.40	.46	.15	.59	.53	.20	.57	.86	.64	.55
κ_n-defined chance agreement				.167	.25	4.5×10^{-6}	.20	.50	7.0×10^{-9}	.50	.0062	.0039	2.4×10^{-5}
Estimated weighted mean Cohen's κ				.96	.92	.85	.86	.93	.91	.91	.76	.72	.80
Estimated weighted mean κ_n				.96	.92	.87	.90	.92	.92	.93	.88	.88	.87
Combined weighted estimate of z				118.9	71.9	118.6	81.3	46.1	123.4	44.2	32.4	22.9	60.0
File drawer analysis				77,102	23,246	76,483	45,680	5,736	49,502	4,635	2,270	762	11,804

Note. Rel *N* = responses included in the reliability analysis; Part *N* = participants included in the analysis; OP = outpatient; MP = mixed psychiatric; NP = nonpatient.

[a] Authors treated content scores in different ways. Some studies calculated agreement on primary and secondary content, some included tertiary content, and some considered all content. To contend with this, I conservatively assumed that only primary contents were scored in the reliability analyses. Chance agreement rates were then calculated as if each response had been given only one mutually exclusive and exhaustive content score. Thus, for content scores, observed agreement rates are underestimated, chance agreement rates are overestimated, and kappa values are underestimated.

[b] Study did not report the average number of responses, so each protocol was conservatively estimated to contain 20 responses.

[c] These studies calculated interrater agreement across all Location and all Developmental Quality scores. Rather than treating this as a separate category, the reported agreement was conservatively assigned to both categories.

[d] Author initially reported a range of reliability values and was contacted for the specific findings in each category.

[e] Study calculated percentage agreement for only the scores actually assigned by an "anchor" rater. As such, the only data suitable for inclusion consisted of those categories for which a mutually exclusive and exhaustive set of scores must be assigned to each response.

[f] Study did not report exact agreement for each category but instead reported the lowest and highest values and the mean. Therefore the two known values were inserted and the mean value was used for the remaining unknown categories.

As one would expect if the CS is a valid measure of psychopathology, for several of the target categories there is a strong linear association between the kappa-defined chance agreement rates and the severity of disturbance within a sample. As psychiatric impairment increases across samples, there is a parallel change in several kinds of scores. Increasingly disturbed samples show an increase in unusual blot locations (Location and Space), perceptual inaccuracies (Form Quality), and cognitive disorganization (Cognitive Special Scores and All Special Scores), as well as a decrease in Popular responses. The change in base rates across samples has a direct bearing on the probability that two raters will agree by chance alone. Specifically, as the base rate for an initially rare score increases, the kappa-defined probability of chance agreement decreases. For instance, cognitive disorganization is rare in nonpatients and more common in patients with schizophrenia. As a result, the Table indicates that chance agreement for Cognitive Special Scores is high in the nonpatient sample (.86) but lower in the schizophrenic sample (.41). This relationship does not mean that scoring becomes easier or more accurate across samples. Rather, it simply demonstrates that chance agreement is dependent on base rates when chance is defined in a fixed-marginals model.

Cohen's kappa was estimated for each score category within a sample by inserting the observed agreement rate and the chance agreement rate for the appropriate sample type into the formula presented earlier ($A_o - A_c/[1 - A_c]$). Following Rosenthal (1991), each kappa value was then transformed to Fisher's Z, weighted by sample size, averaged, and then transformed back to the original coefficient in order to obtain the average weighted coefficient. Using the appropriate estimates of chance agreement, Kappa$_n$ coefficients were calculated and summarized in a parallel fashion, although they were not transformed to Fisher's Z because their equivalence to a correlation has not been demonstrated.

The estimated values for Cohen's kappa range from .72 to .96, having a mean of .86. A value of .86 indicates that agreement was found in 86% of the instances when chance could not account for the agreement. Kappa values of .75 or higher are generally taken to indicate excellent agreement beyond chance, values between .60 and .74 are considered good agreement, those between .40 and .59 are considered fair, and those below .40 are considered poor (Fleiss, 1981). Using these guidelines, one scoring category would be considered to have a beyond-chance level of agreement falling at the upper end of the "good" classification, whereas the remaining nine categories have "excellent" agreement beyond chance. Using the alternative definition of chance found with kappa$_n$, the chance-corrected reliability of the CS would be considered "excellent" across the full range of scores.

Of interest is that the simple percentage agreement index does not appear to be extremely misleading relative to the reliability indices that are corrected for chance. Across categories, the mean for kappa is .86, the mean for kappa$_n$ is .91, and the mean for percentage agreement is .92. Thus, when reliability is assessed in clinical samples using these broad scoring categories and when observed agreement rates are substantial (i.e., >.80), there appears to be little cause for concern about the percentage agreement index. Although this may seem surprising to some, the reason behind this is simple. The greatest degree of aggregation in response segments occurs for those scores that have the most extreme base rates (i.e., Determinants, Content, and Special Scores). Thus, the typical procedure of calculating percentage agreement within response segments actually protects against inflated chance agreement rates.

To further evaluate Wood et al.'s (1996a, 1996b) suggestion that CS scoring may be a random process, I tested the null hypothesis across studies that true agreement was actually equal to random chance agreement. For this and the following analysis, chance was defined according to Cohen's kappa. Using the method described by Rosenthal (1991; Equation 5.5), a standard normal deviate was generated for each study and weighted by sample size. The initial z values for each study were calculated using Cohen's formula for statistical significance (1960; Equation 10), even though this produces a slight underestimate of z (see Fleiss, Cohen, & Everett, 1969). The average combined z value in Table 1 is 71.97, with a range from 22.9 to 123.4. Keeping in mind that a z value of 1.645 occurs 1 time in 20 by chance alone (i.e., $p = .05$, one tailed) and a z value of approximately 8.00 occurs by chance 1 time in a quadrillion (i.e., $p = .0000000000000001$), it is clear that chance factors do not account for the demonstrated reliability of CS scoring in the published literature.

Another way to demonstrate this point is to conduct a "file drawer analysis" (Rosenthal, 1991). Such an analysis is instructive because it also helps to evaluate the impact of any potential publication bias. A file drawer analysis determines the number of studies with null results that would be required to bring the observed findings down to a marginal level of statistical significance (i.e., $p = .05$). The number derived from a file drawer analysis can refer to studies that have been conducted but not published as well as to studies that have yet to be conducted. For our purposes, the obtained value would indicate the number of unsampled Rorschach studies that would have to find that interrater reliability is no better than chance. Taking the average across all scoring categories presented in the final row of Table 1, the data indicate that approximately 29,722 studies with random levels of agreement must be discovered lurking about before the literature summarized here could be brought to a trivial level of significance or attributed to publication bias.[9] Clearly, CS scoring rules do not generate random variables.

The same point is demonstrated with the other chance-corrected reliability coefficients uncovered in this literature review. Perry, McDougall, and Viglione (1995) evaluated the Location, Determinant, Form Quality, and Special Score categories and obtained kappa coefficients that ranged from .63 to .89, or from "good" to "excellent." Another study (Perry, Sprock, et al., 1995) used the same scoring categories and three independent

[9] There is a considerable range in the number of studies that would be needed to negate the present findings, from a low of 762 to a high of 77,102. This variability is a direct function of the amount of data available for the meta-analysis. The exponential relationship between the number of studies contributing data and the number of studies needed to negate the findings is $r = .89$. This suggests that all categories would have had file drawer numbers in the tens of thousands if only a few more primary studies had been available for the Pair, Popular, and Special Score categories.

raters. The authors obtained weighted kappa values that ranged from .71 to .82. In a third study (Greco & Cornell, 1992), although the authors did not report specific values, they indicated that the intraclass correlations for three summary scores were all above .85. The final study (Netter & Viglione, 1994) reported intraclass correlations in excess of .95 for seven summary scores and a value of .90 for one other score. Note how the latter findings are consistent with the psychometric principle that aggregation enhances reliability and that CS summary scores should be more reliable than individual responses.

Test–Retest Reliability

A tenet of the Wood et al. (1996a, 1996b) critique was that the empirical foundation for the CS resided in Exner's books, making the data presented there of pivotal importance. Particularly if this reasoning were accurate, it would have been essential to review all of the reliability data presented by Exner. However, Wood et al. focused on only a bit of data drawn from Exner's two interrater reliability studies. They never mentioned the more extensive test–retest data that were also documented in Exner's texts. Detailed methodology and findings from eight adult studies were reported in Exner's 1978 book (pp. 63–80 and 86–93), and his 1993 book discussed three additional adult studies (pp. 45 and 51) as well as six studies conducted with children of various ages (pp. 46–48). Disregarding this more extensive reliability research is not a trivial omission. Not only are their data quite compelling, but these studies also address many of Wood et al.'s (1996a, 1996b) other criticisms. For instance, Exner's retest studies used *only* the summary scores that Wood et al. felt were lacking in reliability data. These studies also quantified unreliability that could be attributed to examiner influence, administration errors, and the inaccurate recording of responses—issues that Wood et al. asserted had not been subject to empirical scrutiny.

Exner's core reliability study consisted of 100 adult nonpatients retested after a 3-year interval (see Exner, Armbruster, & Viglione, 1978, for the peer-reviewed publication of this study). Twenty-six examiners collected data at the initial testing and 22 collected the retest protocols. Importantly, *all* Rorschachs were scored at the conclusion of the study by eight raters who were unaware of the purpose of the investigation and of the fact that every participant had contributed two protocols. Over the course of 3 years, Exner found that the median retest coefficient for 23 variables believed to reflect trait-like dispositions was .81, whereas the values for two variables believed to measure transient state-like reactions were .39 and .23 (Exner, 1993, p. 46).

In another study (Exner, 1980, 1993), 60 eight-year-old children were randomly assigned to one of two retest conditions that were given 3 or 4 days after the initial test. The control group was retested under standard conditions, whereas the experimental group was instructed to give answers that were different from what they had given on the first test. Ten examiners were randomly assigned to test the children, with no examiner testing the same child twice. After testing, the 60 pairs of Rorschach protocols were randomly ordered, and 20 were given to each of three judges who were instructed to determine whether the responses from the second protocol were the same or similar to the responses given on the initial test. The analysis revealed that 86% of the responses from the control participants were similar, whereas only 14% of the responses from the experimental participants were similar. Despite the fact that the experimental participants gave predominantly *different* responses, the retest reliabilities were largely the same for both groups. For 19 summary scores (Exner, 1980, p. 572), the median retest coefficient for the experimental group was .83 (range = .27–.94), whereas the median retest coefficient for the control group was .87 (range = .48–.94). In both cases, the lowest retest coefficient was for a variable of state-sensitive transient stress.

A similar experimental design was used with 50 adults who were newly admitted inpatients suffering from depression (see Haller & Exner, 1985, for the peer-reviewed publication). The baseline test was completed within 5 days of admission by 1 of 10 examiners, whereas the retest was completed 3 to 4 days later by a different examiner. The 25 participants randomly assigned to the experimental condition were instructed to give different answers on the retest, whereas the control participants were informed that the retest was a standard part of the research protocol. For the experimental group, 34% of the responses were judged to be similar on both tests, whereas the figure was 66% for the control participants. Despite the fact that all of the patients were in the midst of psychological restabilization, across 28 summary scores the median retest coefficient for the control group was .74 (range = .28–.87), whereas for the experimental group it was .71 (range = .33–.88).

If Wood et al. (1996a, 1996b) had attended to the strong and differentiated retest coefficients found in these studies, it would have been virtually impossible to suggest that the CS suffers from random scoring or potentially salient examiner and administration confounds.[10]

Conclusion

Wood et al. believed CS reliability was "presently unknown" (1996b, p. 14), not nearly as high as many people may have assumed (1996a, p. 9), and potentially little more than a random

[10] Of course, there are still factors that will alter the reliability of Rorschach-derived data. Exner (1993, pp. 338–342; or Exner, 1988, for the peer-reviewed version) has presented data indicating that brief records will yield unreliable and therefore interpretively invalid protocols, particularly if these records are accompanied by minimal determinant articulation. Schwartz, Mebane, and Malony (1990) have identified additional constraints, although most examiners are unlikely to encounter these circumstances in general practice. These authors had 24 prelingually deaf participants take the Rorschach in a counterbalanced design under two administration conditions, neither of which were typical. In one, examiners used prewritten cards for the administration and inquiry, and then had participants provide written responses. In the second condition, the examiner and participants communicated in sign language for the administration and inquiry. The retest interval was 1 week. The median retest correlation for 25 CS variables was .62 (range = −.07–.83). To place these findings in context, they can be contrasted with the results of a similar MMPI study. Brauer (1992) had 35 deaf participants respond to MMPI critical items in a counterbalanced design consisting of two different videotaped individuals communicating the items in sign language. The retest interval was 30 *minutes*. The median retest correlation across 34 items was .475 (range = −.06–1.0), whereas the retest coefficient for the sum of all items was approximately .53.

process (1996a, p. 4). In part because of their grave reservations about reliability, they implied that ethical canons and practice standards justified a moratorium on the use of the CS until more sturdy data were available.

However, the facts that were already available indicate that these criticisms of CS reliability are unwarranted and quite misleading. Not only did Wood et al. (1996a, 1996b) mistakenly elevate some relatively minor concerns to matters of critical importance, but they simply did not consider a wealth of data that contradicted the points that they were attempting to make. Wood et al. (1996a) claimed that summary scores should be the focus of research on interrater reliability. However, the current focus on individual responses is both a more appropriate and more exacting test of interrater agreement. Wood et al. (1996a) also suggested that CS interrater agreement was not much better than random chance. This was done after defining chance in a fixed-marginals model (i.e., Cohen's kappa) and presenting hypothetical data from a single variable with a low base rate. Unfortunately, the authors made no mention of alternative ways to define chance, nor did they mention kappa's sensitivity to extreme base rates. In addition, the authors made no effort to evaluate the many independent, peer-reviewed studies of CS reliability available in the published literature. As summarized earlier, this literature indicates CS scoring has excellent estimates of chance-corrected reliability.

Finally, Wood et al. (1996a, 1996b) never seemed to recognize that the form of reliability they were so concerned about, interrater reliability, provides a less stringent evaluation of the CS than the many test–retest reliability studies they ignored. Studies of interrater reliability do not evaluate random errors that could be due to examiner skill, the interpersonal style of the examiner, test administration styles, oddities in the response verbalization or inquiry process, omission and commission mistakes that affect how responses are recorded, peculiarities of the testing occasion, developmental processes, or fluctuating emotional states. Instead, studies of interrater reliability use a static stimulus (the written transcript of responses) and only assess potential scoring error. In contrast, *all* of the preceding sources of error are free to vary in a retest design. Thus, even though Wood et al. (1996a) asserted that these kinds of potential error had not been studied, they dismissed numerous investigations that provided relevant data.

Wood et al. (1996a, 1996b) also raised many concerns about CS validity. Of course, validity is a more central concern, and a number of the challenges that they raised deserve careful consideration. However, validity is a more complicated matter to address than reliability. Because validity is more complicated and because Wood et al. made erroneous assertions about the more elementary topic of CS reliability, it would be wise for those who have a genuine scientific interest in the Rorschach to consider the complexity of the issues and the full array of data (see Atkinson, 1986; Atkinson, Quarrington, Alp, & Cyr, 1986; Bornstein, 1996; Hilsenroth, Fowler, Padawer, & Handler, 1997; Meyer, 1996a, 1996b, 1997; Meyer & Handler, 1997; Parker, 1983; Parker, Hanson, & Hunsley, 1988) before accepting Wood et al.'s (1996a, 1996b) assertions about Rorschach validity.

References

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

*Abraham, P. P., Lepisto, B. L., Lewis, M. G., Schultz, L., & Finkelberg, S. (1994). An outcome study: Changes in Rorschach variables of adolescents in residential treatment. *Journal of Personality Assessment, 62,* 505–514.

*Adrian, C., & Kaser-Boyd, N. (1995). The Rorschach Ego Impairment Index in heterogeneous psychiatric patients. *Journal of Personality Assessment, 65,* 408–414.

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.

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin, 17,* 483–489.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease, 163,* 307–317.

*Berg, J. L., Packer, A., & Nunno, V. J. (1993). A Rorschach analysis: Parallel disturbance in thought and in self/object representation. *Journal of Personality Assessment, 61,* 311–323.

Bornstein, R. F. (1996). Construct validity of the Rorschach Oral Dependency Scale: 1967–1995. *Psychological Assessment, 8,* 200–205.

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

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687–699.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

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

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

Exner, J. E., Jr. (1980). But it's only an inkblot. *Journal of Personality Assessment, 44,* 563–577.

Exner, J. E., Jr. (1986). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (2nd ed.). New York: Wiley.

Exner, J. E., Jr. (1988). Problems with brief Rorschach protocols. *Journal of Personality Assessment, 52,* 640–647.

Exner, J. E., Jr. (1991). *The Rorschach: A comprehensive system: Vol. 2. Interpretation* (2nd ed.). New York: Wiley.

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

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

*Exner, J. E., Jr., & Andronikof-Sanglade, A. (1992). Rorschach changes following brief and short-term therapy. *Journal of Personality Assessment, 59,* 59–71.

Exner, J. E., Jr., Armbruster, G. L., & Viglione, D. (1978). The temporal stability of some Rorschach features. *Journal of Personality Assessment, 42,* 474–482.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L., Cohen, J., & Everett, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323–327.

*Frueh, B. C., & Kinder, B. N. (1994). The susceptibility of the Rorschach Inkblot Test to malingering of combat-related PTSD. *Journal of Personality Assessment, 62,* 280–298.

Frueh, B. C., Leverett, J. P., & Kinder, B. N. (1995). Interrelationship between MMPI-2 and Rorschach variables in a sample of Vietnam veterans with PTSD. *Journal of Personality Assessment, 64,* 312–318.

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

Greco, C. M., & Cornell, D. G. (1992). Rorschach object relations of adolescents who committed homicide. *Journal of Personality Assessment, 59,* 574–583.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry, 38,* 408–413.

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

Hilsenroth, M. J., Fowler, J. C., Padawer, J. R., & Handler, L. (1997). Narcissism in the Rorschach revisited: Some reflections on empirical data. *Psychological Assessment, 9,* 113–121.

*Holaday, M., & Whittenberg, T. (1994). Rorschach responding in children and adolescents who have been severely burned. *Journal of Personality Assessment, 62,* 269–279.

Jones, A. P., Johnson, L. A., Butler, M. C., & Main, D. S. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. *Academy of Management Journal, 26,* 507–519.

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

McDowell, C., & Acklin, M. W. (1996). Standardizing procedures for calculating Rorschach interrater reliability: Conceptual and empirical foundations. *Journal of Personality Assessment, 66,* 308–320.

*Meloy, J. R. (1992). Revisiting the Rorschach of Sirhan Sirhan. *Journal of Personality Assessment, 58,* 548–570.

*Meloy, J. R., Gacono, C. B., & Kenney, L. (1994). A Rorschach investigation of sexual homicide. *Journal of Personality Assessment, 62,* 58–67.

*Meyer, G. J. (1993). The impact of response frequency on the Rorschach constellation indices and on their validity with diagnostic and MMPI-2 criteria. *Journal of Personality Assessment, 60,* 153–180.

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

Meyer, G. J. (1996b). 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. (1997). On the integration of personality assessment methods: The Rorschach and MMPI. *Journal of Personality Assessment, 68,* 297–330.

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

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

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., Jr. (1995). A five-year follow-up study on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64,* 112–118.

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

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

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

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44,* 172–177.

*Silberg, J. L., & Armstrong, J. G. (1992). The Rorschach test for predicting suicide among depressed adolescent inpatients. *Journal of Personality Assessment, 59,* 290–303.

*Singer, H. K., & Brabender, V. (1993). The use of the Rorschach to differentiate unipolar and bipolar disorders. *Journal of Personality Assessment, 60,* 333–345.

*Sloan, P., Arsenault, L., Hilsenroth, M., Harvill, L., & Handler, L. (1995). Rorschach measures of posttraumatic stress in Persian Gulf war veterans. *Journal of Personality Assessment, 64,* 397–414.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and application.* Thousand Oaks, CA: Sage.

Weiner, I. B. (1991). Editor's note: Interscorer agreement in Rorschach research. *Journal of Personality Assessment, 56,* 1.

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

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

*Yama, M. F., Call, S. E., Entezari, P. (1993). A new test of an old hypothesis: A quantification of sequence in the Rorschach. *Journal of Personality Assessment, 60,* 60–73.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374–378.

Received October 9, 1996
Revision received April 17, 1997
Accepted May 6, 1997 ■