JOURNAL OF PERSONALITY ASSESSMENT, 1996, 66(2), 308–320

ARTICLES

Standardizing Procedures for Calculating Rorschach Interrater Reliability: Conceptual and Empirical Foundations

Claude McDowell and Marvin W. Acklin
Honolulu, Hawaii

Although the Rorschach test has demonstrated significant refinements in reliability, validity, and statistical power as a result of the procedural standardization and scoring innovations introduced by Exner's Comprehensive System, the issue of Rorschach interrater reliability remains unexplored. This article examines the psychometric foundations of Rorschach interrater reliability and applies notions from applied behavioral analysis to the treatment of Rorschach data. We empirically compare 3 methods of quantifying interrater agreement, their accuracy in estimating interrater agreement, and efficiency in reducing error in Rorschach research. Results indicate that the magnitude of differences between methods of quantifying interrater agreement and the associated reductions of error are significant. We propose a standard method for quantifying interrater agreement in Rorschach research.

The Rorschach test is an anomaly in the field of psychological measurement and assessment when compared to objective and other projective techniques. The undeveloped state of the technique at the time of Rorschach's demise; the Rorschach's serendipitous transplantation from Europe to America; and the fragmenting of the test into different "systems" of administration, scoring, and interpretation would surely be enough to differentiate the test from its counterparts. Even considering the standardization of test administration procedures and scoring system by Exner's Comprehensive System for the Rorschach (Exner, 1991, 1993; Exner & Weiner, 1995), the complex types of data developed by the Rorschach introduce formidable obstacles to the

application of standard procedures and canons of test development. This has been the basis of savage academic criticism of the test. The quality of research and clinical applications made possible by Exner's innovations have done much to satisfy the Rorschach's psychometric critics and extend the test's applications (Meloy, 1991). It has been demonstrated, for example, that the Rorschach's statistical power is not dissimilar to other social science research and that research conducted using the Comprehensive System is more sensitive than non-Comprehensive System research (Acklin, McDowell, & Orndorff, 1992). Further refinements in the test's psychometric properties may be possible with increasingly sophisticated research designs, larger sample sizes, and the use of more powerful parametric statistics (Acklin et al., 1992; Parker, 1983).

One feature of the Rorschach Test that is especially vexing is the heavy reliance on the interpreter of the test, in both scoring and deriving inferences from the psychometrically complex data. This reliance on scoring, or better, coding, procedures and the threat posed by scorer or examiner variance (Anastasi, 1988) elevates the importance of interrater agreement in relation to other approaches to quantifying reliability (viz., internal consistency, split-half, test–retest, or alternate forms approaches). A survey of the literature indicates that approaches to Rorschach interrater reliability have been haphazard and hardly satisfying from a methodological point of view. Standards for reporting interrater agreement were promulgated only as recently as 1991 in the major outlet for American Rorschach research, the *Journal of Personality Assessment* (Weiner, 1991). Exner, as the preeminent Rorschach researcher, has addressed issues of interrater reliability in his published work (Exner, 1991, 1993). He has advocated a procedure for the quantification of interrater agreement based on a percentage agreement method of calculation (Exner, 1991, 1993). Conceptual issues related to interrater agreement and reliability, as far as we know, have not been discussed in the depth they deserve. Specific procedural strategies for undertaking interrater reliability studies have not been described in sufficient detail either (Acklin & McDowell, 1995).

This article examines the conceptual foundations of interrater reliability and examines the historical role of interrater agreement in Rorschach research. To place the Rorschach solidly in the realm of traditional psychometric concerns, we apply concepts from applied behavioral analysis to Rorschach data, viewing the Rorschach as a complex sample of verbal behavior. Of special interest is determining the unit of observation for the purpose of quantifying interrater agreement. We examine the impact on reliability error by empirically comparing three methods of quantifying Rorschach interrater agreement. Based on our findings, we propose a standard method for calculating interrater agreement in future Rorschach research. Specific approaches and strategies for conducting interrater reliability studies are described in a companion article (Acklin & McDowell, 1995).

CONCEPTUAL FOUNDATIONS FOR INTERRATER RELIABILITY

Replicability of research is a highly valued aspect of the scientific method. Assessing and reporting interrater agreement in research involving scoring and/or rating decisions by judges offers important steps in quality assurance. In a strongly written literature review on the reliability of the Rorschach and other projective techniques, Jensen (1959) declared scoring reliability as minimum evidence that a study is replicable and of scientific and clinical value.

Generalization of findings is another canon of scientific endeavor. In research with rating and scoring systems, it is important that judges are sufficiently consensual in their ratings to approach some external criterion. Tinsley and Weiss (1975) have articulated the factor of judge subjectivity:

> Generality is important in demonstrating that the obtained ratings are not the idiosyncratic results of one rater's subjective judgment. Knowledge of the interrater reliability and interrater agreement is crucial in evaluating the generality of a set of ratings. (p. 359)

Many factors influence scoring agreement between judges: training, similarity in scoring experiences, population from which protocols are drawn, method of accuracy assessment, observer bias and drift, characteristics of the behavior and coding scheme, and statistical methods of quantifying interobserver agreement (Haynes, 1978; Jensen, 1959).

Quantification of interrater agreement is the attempt to account for and ultimately reduce error variance. With regard to the Rorschach, researchers need to be continually aware of more stringent alternatives, given the instrument's complex and controversial psychometrics (Jensen, 1959). Error variance has been implicated as one of three major factors contributing to diminished statistical power in Rorschach research (Acklin et al., 1992) and likely contributed to the "artifactual controversy" surrounding the test (Rossi, 1983). The importance of interrater agreement is central to defining and accounting for the measurement error in psychological research instruments. No psychological measure, of course, is without error. Refinement in psychological description and prediction can hardly be expected without reliable measurement of variables (Bartko & Carpenter, 1976). Quantification of error is imperative in evaluating the fundamental tests of a measure's utility and value, reliability, validity, and power (Cohen, 1988) and in improving the overall quality of research in the behavioral sciences (Rossi, 1990; Sedlmeier & Gigerenzer, 1989).

Error has been defined as any condition that is irrelevant to the purpose of a test (Anastasi, 1988). Accounting for error variance in a test demarcates its reliability. Anastasi (1988) has identified three primary sources of measurement error that may compromise reliability: content sampling, time sam-

pling, and interscorer differences. Sources of error, or error variance, may be statistically quantified and provide information regarding a measure's internal consistency, intersession reliability, and interobserver agreement (Haynes, 1978).

Tests differ in the ways they are affected by error variance. Projective tests, in particular, leave a good deal to the judgment of the scorer, increasing concerns about interscorer agreement and error. Efficient and standardized methods for quantifying interrater agreement in projective techniques is critical in determining their validity, utility, and power.

The first three decades of Rorschach research were characterized by disunity in administration and scoring that fostered diversity regarding the coding of Rorschach variables (Exner, 1968). Rorschach made no mention of interater agreement in his seminal monograph, *Psychodiagnostics* (Rorschach, 1942). Early research efforts rarely undertook, reported, or inconsistently reported evidence of interrater agreement. Cronbach, in his classic discussion of statistical methods in Rorschach research, failed to mention interrater reliability (Cronbach, 1949). Jensen (1959), summarizing the literature up to 1958, reported a growing trend to report reliability data. Reports of studies examining interrater agreement are extremely rare and use statistical procedures that are unsuitable for the purpose (e.g., ϕ coefficients, Pearson's *r*, Spearman–Brown formula). Exner's Comprehensive System (Exner, 1993) has done much to address these issues over the last 20 years by standardizing administration, codifying scoring, and conducting reliability and validity studies. However, assumptions and methods for assessing interobserver agreement in Rorschach scoring have not been well described or examined in the quest to refine the test's value (Jensen, 1959).

Several methods have been proposed for quantifying interrater agreement. Determining the most appropriate method is critically important because different methods significantly affect inferences and estimates of agreement (Haynes, 1978). Appropriate methods for quantifying interrater agreement depend on the type of data or behavior observed (Haynes, 1978). We believe that application of concepts from applied behavioral theory to the Rorschach are appropriate and fruitful in developing conceptual assumptions for Rorschach interrater agreement.

THE RORSCHACH AND BEHAVIORAL ASSESSMENT

Applied behavioral analysis relies heavily on observation, use of behavior coding schemes for the purposes of data collection, and the quantification of behavior as a basis for intervention. Observation is focused on "samples" of behavior, whether across time, event, or situation. Behaviorists prefer observation in structured to naturally occurring environments because of efficiency, particularly in increasing the probability that target behaviors will be emitted (Haynes, 1978). Psychological testing, of course, represents the use

of a structured environment for the observation of behavior. Administration of the Rorschach is no exception. Standard administration includes opportunities for observing and recording *in vivo* behavioral observation and a complex sample of verbal-linguistic behavior. Rorschach "scores" are a coding or rating scheme for various qualities of verbal behavior from which the clinician infers cognitive and perceptual operations and processes. These behavioral codes are used to generate inferences about dispositions, coping skills, and problem-solving behavior. Viewing Rorschach scores as codified elements of behavior, we propose that principles of behavior assessment (Haynes, 1978) provide a useful conceptual framework in standardizing procedures for quantifying Rorschach interrater agreement.

Haynes (1978) reported that the most frequently used method of calculating interrater agreement is based on observer agreement and disagreement within "sampling intervals," that is, the units of observation that serve as the basis for agreement or disagreement concerning a behavior's occurrence. Applied behavioral analysis approaches observable behavior as it is emitted across time, event, or situation. Repp, Deitz, Boles, Deitz, and Repp (1976) and Haynes (1978) have demonstrated significant differences in the calculation of interrater agreement based on the definition of the sampling interval used, whether it is based on whole session, category, or time interval of observation.

One may consider the Rorschach Test as analogous to a behavioral observation. Considering the Rorschach, however, what is the most appropriate unit of study? The possibilities include the protocol considered as a whole, the individual response, response segments (Location, DQ, Determinants, etc.), or individual elements of the response, that is, specific scores or codes (*W*, +, *F*, *M*, *P*, etc.). As early as 1963, Gleser proposed that estimates of reliability might be improved by "grouping stimuli into three or four subgroups on the basis of judgement of similarity along certain major perceptual dimensions such as color, shading, and tendency to elicit whole versus detail responses" (Gleser, 1963, p. 399). We propose, with Exner and Weiner, that "segments" of the individual Rorschach response should be viewed as basic units of observation or, in behavioral terms, the sampling interval. Coded segments of the individual Rorschach response are the units of behavior that, when integrated in the structural summary, provide the basis for interpretive inferences.

There are difficulties associated with interval agreement methods related to the type and nature of the behavior observed. Agreement coefficients are affected by the frequency or occurrence of target behaviors observed (Haynes, 1978). For example, if a behavior has a low or high probability of occurrence (e.g., .2 or .8), coefficients tend to spuriously inflate. A second problem is whether coding should include occurrences of designated behaviors, nonoccurrences, or both. Haynes (1978) and others have reviewed approaches for increasing the power of interval methods. One method in-

volves omitting agreements of nonoccurrence from calculation in low rate behaviors. This method reduces inflation, but according to Haynes (1978), misses a critical assumption:

> Although using only occurrences to calculate interobserver agreement prevents an overestimation of interobserver agreement when low-rate behaviors are being observed, this statistical procedure does not assume that agreement about the nonoccurrence of a behavior is an agreement and is unaffected by errors resulting from the observation of high rate behaviors. (p. 161)

Haynes (1978) has supported the observations of Hawkins and Dotson (1975) and Bijou et al. (1969) that interobserver agreement coefficients should be based on both occurrences and nonoccurrences.

APPROACHES TO RORSCHACH INTERRATER AGREEMENT

Bartko and Carpenter (1976) reviewed reliability methods, distinguishing between more or less suitable approaches based on "levels of measurement" for commonly occurring data sets. Data may be defined as nominal: dichotomous (+ or −) with two judges; dichotomous with three or more judges; polychotomous (+, −, *o*, *u*) with two judges; polychotomous with three or more judges; or ordinal—quantitative (*1*, *2*, *3*, *4*) with two or more judges (Bartko & Carpenter, 1976). Tinsley and Weiss (1975) affirmed the idea that different levels of measurement require different approaches for calculating interrater agreement. Thus, levels of measurement in Rorschach scoring must be taken into account. Rorschach coding (e.g., Location, Determinants) offers no quantitative basis of measurement. The observer does not decide how much of something (an ordinal level of measurement), for example, *D* or *Dd*, is being observed, but its presence or absence. Whether a participant gives a *W*, *D*, or *Dd* response (or +, *o*, *u*, −) is more indicative of a style or quality rather than a quantity of behavior. Considering the determinants, the data looks rank-ordered (i.e., ordinal), but from the rater's perspective, categorical decision making is required (Block, 1962). As such, Rorschach coding falls, in Bartko and Carpenter's scheme, within the realm of polychotomous nominal data. Tinsley and Weiss (1975) and others (e.g., Bakeman & Gottman, 1987) have noted that the percentage agreement approach between judges is most common and practical using nominal scales.

The percentage agreement approach to quantifying interrater agreement, though widely used, is problematic because chance agreement between raters is not taken into account. This tends to inflate estimates of agreement (Bartko & Carpenter, 1976; Haynes, 1978; Hubert, 1977; Suen & Ary, 1989; Tinsley & Weiss, 1975). Bartko and Carpenter (1976) reported two situations when percentage agreement may, nevertheless, be appropriate: zero variability between raters (i.e., 100% agreement) and when important variables

occur infrequently. Suen and Lee (1985) argued against percentage agreement as a basis for calculating interrater agreement. Reanalyzing a sample of previously published data using a chance-corrected agreement index instead of percentage agreement, between one fourth and three fourths of the data in prior studies would have been judged as having unacceptably low reliability. Landis and Koch (1977) suggested that a kappa of 0.80 is an indication of good reliability. They offered the following schema for interpreting kappa coefficients: 0 to .2 = none or slight agreement; .21 to .40 = fair agreement; .41 to .60 = moderate agreement; .61 to .80 = substantial agreement; .81 to 1.0 = almost perfect agreement.

The majority of contemporary Rorschach research uses percentage agreement to assess scoring reliability. Editorial standards for the *Journal of Personality Assessment* have included a percentage agreement criterion of 0.80 for published Rorschach research (Weiner, 1991) but a rationale or delineation of procedures for calculating interrater agreement has not been described. The typical article reports the outcomes of interrater reliability studies, typically as percentage agreement, but rarely describes the procedure by which coefficients are calculated. Exner (1991) has advocated the percentage agreement approach for quantifying interrater agreement. Concerning unit of analysis, both Exner and Weiner advocate a percentage agreement method using response segments (e.g., Location, Determinants, Form Quality, etc.) as most practical. More recently, Exner (1993) has conducted percentage agreement studies on individual determinants yielding very high coefficients (ranging from .88 to .99; Exner, 1993). The percentage agreement method is prone to coefficient inflation due to chance agreement and has been considered a "generally undesirable method of reliability assessment" (Bartko & Carpenter, 1976). The Kappa coefficient, as a corrected estimate of percentage agreement, has been recommended as a more suitable alternative to assessing interrater reliability for nominal scale data given that it is a percentage agreement coefficient that corrects for chance agreement between raters (Bartko & Carpenter, 1976; Cohen, 1960; Fleiss, 1971; Hubert, 1977; Haynes, 1978; Tinsley & Weiss, 1975). Interesting studies conducted by DeCato (1983, 1984) examined Rorschach scores with Cohen's Kappa (a stringent method of calculating agreement) using a total frequency or protocol approach, equivalent to the whole session approach in behavioral analysis (a method that inflates coefficients for large samples), yielding very high estimates of interrater agreement.

In this study we examined differences between three common methods of calculating interrater agreement for the Rorschach. We examined differences between the total protocol approach that views the whole protocol as the unit of observation; the percentage agreement approach that views segments of the individual response as units of observation but does not control for chance agreement; and Cohen's Kappa, a corrected percentage agreement coefficient that controls for chance agreement. We hypothesized that

Cohen's Kappa would more accurately estimate interrater agreement and significantly reduce error when calculating interrater agreement for the Rorschach in comparison to the other methods.

METHOD

Participants

The study used 20 protocols that generated a total of 412 responses. Participants were undergraduate psychology student volunteers attending a private urban university in the Midwest. The students took the test for experimental credit in their introductory psychology class. The modal age was 18 years. Seventy percent of the participants were women and 65% were White. The records were administered and scored according to the standard Comprehensive System instructions by clinical psychology students under supervision of Marvin W. Acklin. The records were valid and typical in terms of response productivity and other validity data when compared to Exner's reference data for nonpatient adults.

Calculating Interrater Agreement

The 20 protocols were rescored by Claude McDowell and then independently rescored by an advanced graduate student who had received training from Marvin W. Acklin. Responses were divided into nine segments. The nine segments of the responses yielded 412 data entries ($N = 412$) from the two raters. The ratings were subsequently analyzed and coefficients of agreement calculated.

Total protocol coefficients were calculated by hand. Percentage agreement and Kappa coefficients were calculated using Dynastat's Kappa Program (Dynastat, 1988). The DOS-based program includes several methods

TABLE 1
Example of Interrater Confusion Matrix: Location Segment

	Judge 1						
	W	*WS*	*D*	*DS*	*Dd*	*DdS*	*f (B)*
Judge 2							
W	162	5	8		2		177
WS	1	29	1	1			32
D	3	1	137	3	2		146
DS		1	1	10			12
Dd	1		6		25	1	33
DdS					3	9	12
f (A)	167	36	153	14	32	10	412

Note. Number of agreements / number of agreements + number of disagreements = percentage agreement (372 / 372 + 40 = .90).

for calculating Kappa, including Cohen's original formulation of the statistic. The program provides, additionally, a confusion matrix showing areas of rater coding agreement and discrepancy. Table 1 provides an example of the confusion matrix for Rorschach Location codes. The sum of the matrix diagonal divided by total number of observations yields the percentage agreement coefficient.

RESULTS

Interrater agreement coefficients and their descriptive data for three calculation approaches are presented in Tables 2 and 3. Mean coefficient score for total protocol approach was .99, percentage agreement approach was .87, and Kappa was .79. Total protocol coefficients tend toward 1.00 given the large sample size ($N = 412$). Descriptive data for total protocol coefficients indicate minimal variation and restricted range for response segment agreement. The percentage agreement approach yielded more conservative estimates than the total protocol approach. Descriptive statistics for percentage agreement indicate a wider range of values calculated. Seven of the nine score segments met the .8 editorial criteria (Weiner, 1991). Form quality and special score segments missed criteria by narrow margins.

Kappa coefficients yielded the most conservative estimates of agreement. Kappa coefficients were all smaller than percentage agreement coefficients, but tended to vary in magnitude. For example, the Determinant segment yielded a .03 difference between percentage agreement and kappa coefficients. In the special score section, on the other hand, a substantial .18 difference was found.

TABLE 2
Comparison of Interrater Agreement Coefficients: Total Frequency and Percentage Agreement

Response Segment	*Total Protocol Approach*	*Percentage Agreement*	*z*
Location	1.00	.90	no variance
DQ	1.00	.89	no variance
Determinants	1.00	.83	no variance
FQ	.99	.79	−40.8[a]
(2)	.99	.94	−10.2[a]
Contents	1.00	.91	no variance
Populars	.97	.96	−1.19
Z Score	.98	.81	−24.7[a]
Specific scores	.95	.77	−16.8[a]
M	.99	.87	−24.5[a]
SD	.02	.07	
Range	.05	.19	

Note. $N = 412$.
[a] $p \leq .001$. Coefficients of 1.00 indicate no variability.

TABLE 3
Comparison of Interrater Agreement Coefficients: Percentage of Agreement and Kappa

Response Segment	*Percentage Agreement*	*Cohen's Kappa*	*z*
Location	.90	.85	2.84[a]
DQ	.89	.79	4.98[b]
Determinants	.83	.80	1.52
FQ	.79	.68	4.79[b]
(2)	.94	.86	4.68[b]
Contents	.91	.89	1.3
Populars	.96	.90	4.06[b]
Z Score	.81	.76	4.68[b]
Specific scores	.77	.59	7.42[b]
M	.87	.79	3.99[b]
SD	.07	.10	
Range	.19	.31	

Note. $N = 412$.
[a] $p \leq .01$. [b] $p \leq .001$.

To assess the significance of differences between total frequency, percentage agreement, and kappa coefficients for Rorschach score segments, *z* tests for comparison of two independent proportions were used (Ferguson, 1976; Welkowitz, Ewen, & Cohen, 1982). Alpha was set at .05. The difference between the mean total protocol and percentage agreement coefficients was statistically significant, $z = -24.50$, $p < .001$. A power analysis using a one-tailed alpha ($N = 412$) yielded an effect size of 2.33 (large effect) and power of 1.00. The difference between the mean percentage agreement and kappa coefficients was also statistically significant, $z = 3.99$, $p < .001$. A power analysis using a one-tailed alpha of .05 ($N = 412$) yielded an effect size of 3.15 (large effect) and power of 1.00.

DISCUSSION

As predicted, the findings demonstrate significant method variance in quantifying interrater agreement for Rorschach data. They indicate significant reductions in error by calculating Rorschach interrater agreement using the more conservative Cohen's Kappa with individual response segments response as the unit of observation. Controlling for chance agreement between raters reduces error variance in observer agreement. The magnitude of mean differences between methods of calculation, based on the obtained large effect size, suggests that use of Cohen's Kappa provides a significant improvement in the Rorschach's interrater reliability.

A recent study of the Rorschach literature found that research conducted using the Comprehensive System was more powerful, in the statistical sense, than non-Comprehensive System research (Acklin et al., 1992). Procedural

standardization, including administration and a well-described definitional criteria for scoring of the test, were likely the primary factors contributing to this increase in statistical power. Because the standard for calculating interobserver agreement, the percentage agreement coefficient, has been most widely used in the contemporary Rorschach literature, including most of the foundational studies of the Comprehensive System, it is safe to assume that the coefficients reported by Exner and others (e.g., Morgan & Viglione, 1992; Gacono, Meloy, & Berg, 1992) are significantly inflated. The result of our study suggests that further psychometric refinements and increases in Rorschach statistical power, as a result of minimizing error variance in interrater agreement, may be possible through standardizing the method for calculating interrater agreement using Cohen's Kappa. A limitation of our study, based on the fact that estimates of interrater agreement are a function of the participants observed, would require replication using a clinical sample.

The total protocol approach for calculating interrater agreement should be considered inappropriate for Rorschach reliability due to spurious inflation of interrater agreement. It can be argued that the percentage agreement approach offers a practical "middle of the road" estimate of interobserver agreement. Although percentage of agreement provides the basis for Kappa corrections, differences between the two approaches are substantial. We propose that a calculation approach that does not account for chance agreement between raters leads to unacceptable inflation in estimates of reliability. It would appear that the use of Cohen's Kappa provides a more sophisticated and stringent method of quantifying interrater reliability in Rorschach research.

What are the implications of minimizing these issues? Given the nature of the Rorschach Test, interobserver agreement is a far more important component of reliability than for more objective tests (e.g., Minnesota Multiphasic Personality Inventory). Even with the many improvements that have characterized contemporary Rorschach studies, researchers invite methodological criticism and reduce the effectiveness of the Rorschach as a research instrument by failing to use more efficient approaches in accounting for error variance when they are readily available.

ACKNOWLEDGMENT

We thank Stephen Haynes for his comments on an earlier draft of this article.

The authors equally share authorial credit for this study.

REFERENCES

Acklin, M. W., & McDowell, C. (1995). *Standardizing procedures for calculating Rorschach interrater reliability: Procedural strategies.* Manuscript in preparation.

Acklin, M. W., McDowell, C., & Orndorff, S. (1992). Statistical power and the Rorschach Test: 1975–1991. *Journal of Personality Assessment, 59*, 366–379.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bakeman, R., & Gottman, J. M. (1987). *Observing interaction: An introduction to sequential analysis*. Cambridge, England: Cambridge University Press.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease, 161*, 307–317.

Bijou, S. W., Peterson, R. S., Harris, S. R., Allen, K. E., & Johnston, M. S. (1969). Methodology for experimental studies of young children in natural settings. *Psychological Record, 19*, 17–210.

Block, W. E. (1962). Psychometric aspects of the Rorschach Technique. *Journal of Projective Techniques and Personality Assessment, 26*, 162–222.

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

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

Cronbach, L. J. (1949). Statistical methods applied to Rorschach scores. A review. *Psychological Bulletin, 46*, 393–429.

DeCato, C. M. (1983). Rorschach reliability: Cross-validation. *Perceptual and Motor Skills, 56*, 11–14.

DeCato, C. M. (1984). Rorschach reliability: Toward a training model for interscorer agreement. *Journal of Personality Assessment, 48*, 58–64.

Dynastat. (1988). *Dynastat's kappa program*. Eugene, OR: Author.

Exner, J. E. (1968). *The Rorschach systems*. New York: Wiley.

Exner, J. E. (1991). *The Rorschach: A comprehensive system. Volume 2: Interpretations*. New York: Wiley.

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

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

Ferguson, G. A. (1976). *Statistical analysis in psychology and education* (4th ed.). New York: McGraw-Hill.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382.

Gacono, C., Meloy, J. R., & Berg, J. (1992). Object relations, defensive operations, and affective states in narcissistic, borderline, and antisocial personalities. *Journal of Personality Assessment, 59*, 32–49.

Hawkins, R. P., & Dotson, V. A. (1975). Reliability scores that elude: In Alice in Wonderland's trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research applications* (pp. 359–376). Englewood Cliffs, NJ: Prentice-Hall.

Haynes, S. N. (1978). *Principles of behavioral assessment*. New York: Gardner.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84*, 289–297.

Jensen, A. R. (1959). The reliability of projective techniques: Review of the literature. *Acta Psychologica, 16*, 108–136.

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Meloy, J. R. (1991, Fall–Winter). Rorschach testimony. *The Journal of Psychiatry and Law*, 221–234.

Morgan, L., & Viglione, D. J. (1992). Sexual disturbances, Rorschach sexual responses, and mediating factors. *Psychological Assessment, 4*(4), 530–536.

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

Repp, A. C., Deitz, D. E., Boles, S. M., Deitz, S. M., & Repp, C. S. (1976). Differences among

common methods for calculating interobserver agreement. *Journal of Applied Behavioral Analysis, 9*, 109–113.

Rorschach, H. (1942). *Psychodiagnostics*. New York: Grune & Stratton.

Rossi, J. (1983, April). *Inadequate statistical power: A source of artifactual controversy*. Paper presented at the meeting of the Eastern Psychological Association, Baltimore.

Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–655.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Journal of Consulting and Clinical Psychology, 105*, 309–316.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative observation data*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Suen, H. K., & Lee, P. S. C. (1985). Effects of the use of percentage agreement on behavioral observation: A reassessment. *Journal of Psychopathology and Behavioral Assessment, 7*, 221–234.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 22*, 358–376.

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

Welkowitz, J., Ewen, R. B., & Cohen, J. (1982). *Introductory Statistics for the Behavioral Sciences* (3rd ed.). New York: Academic.

Marvin W. Acklin
850 West Hind Drive, Suite 209
Honolulu, HI 96821

Received June 5, 1995
Revised July 17, 1995