Psychological Assessment
1997, Vol. 9, No. 4, 495-498

1040-3590/97/$3.00

Thinking Clearly About Reliability: More Critical Corrections Regarding the Rorschach Comprehensive System

Gregory J. Meyer
University of Alaska Anchorage

In this brief comment on J. M. Wood, M. T. Nezworski, and W. J. Stejskal's (1997) response to his article (Meyer, 1997a), the author documents how J. M. Wood et al. continue to make allegations based on a limited subset of the available literature. He also points out specifically how their criticisms regarding kappa, test–retest reliability, true score theory, score aggregation, and his meta-analysis are incorrect. He concludes that these new errors provide additional reasons to be cautious about the conclusions proffered in their other articles on the Rorschach.

Wood, Nezworski, and Stejskal's (1997) response to my article (Meyer, 1997a) on the reliability of the Rorschach Comprehensive System (CS) continues to allege poor rater reliability despite substantial disconfirming evidence and without a single study to support their position. Their response also continues to offer erroneous assertions regarding psychometrics and Rorschach data. In conjunction with my initial article, I hope this reply will help readers to sort through some of the relevant issues.

Did Wood et al. Really Not Intend to Say That CS Scoring Was Little Better Than Random Chance?

It is true that Wood et al. (1997) never made a blanket assertion that CS scoring was random. All they did overtly (Wood et al., 1996a, p. 4) was (a) note that Exner used a percentage agreement (*%A*) index in his two studies (Exner, 1991, 1996), (b) state that *%A* "has long been recognized" as "inadequate," "misleading," and "inflated" because it does not correct observed agreement for chance agreement, and (c) present an example that suggested observed agreement for a CS score was virtually indistinguishable from chance agreement.

Wood et al.'s (1996a) example was the only instance in which statistical information was presented on CS reliability, and nowhere does their text suggest that the example was potentially anomalous. If they did not intend for the example to generalize, it is not clear why they did not temper its implications by citing even a single study from the extensive peer-reviewed literature. As I have indicated, this literature documents that the CS has excellent chance-corrected interrater reliability and excellent test–retest reliability. Also, if they did not intend for the example to reflect a general indictment of the CS, it is not clear why they concluded (Wood et al., 1996a, p. 9; Wood et al., 1996b, p. 17) that the CS did not pass muster as a reliable instrument for assessing personality and that professional standards indicated it should not be used. The message in their article seems clear. Rather than "fiercely dueling with a straw man of (my) own creation" (Wood et al., 1997, p. 490), it seems that the authors may be retreating from the thrust of their earlier report.

My "Fishy" Kappa Example

Wood et al. (1997, p. 490) believe I made an "elementary math error" when discussing the limitations of kappa. However, this is incorrect. Recall the logic of kappa. Kappa indicates the proportion of observed agreement that cannot be explained by chance agreement. If chance agreement explains all of the observed agreement, then kappa must be zero. So, although the formula would yield an "undefined" solution when the denominator is zero, the logic of kappa defines a solution. Fortunately, one does not have to take my word for it. As would be expected, Cohen was well aware of the logic of kappa when he created this statistic. He stated, "When obtained agreement equals chance agreement, $\kappa = 0$" (Cohen, 1960, p. 41). Thus, in my example where observed agreement equals 1.0 and chance agreement equals 1.0, $\kappa = 0$.

Is Kappa Always the Statistic of Choice?

I did not intend to suggest that *%A* should be a preferred index of interrater reliability. However, I also do not universally favor kappa. At times, it is an excellent statistic, but at other times it may not be. Part of my goal was to articulate some of the strengths, limitations, and assumptions that are associated with different reliability statistics. These issues are more complex than many of us generally appreciate, and uninformed choices can have unfortunate consequences. One could certainly circumvent the need to choose judiciously by making arguments from authority or by looking at the issues from a more black-and-white perspective. However, my article was not designed to serve these purposes.

The Relationship Between Scoring Reliability and Test–Retest Reliability

It is true that test–retest reliability does not *necessarily* address coder reliability. For instance, when test–retest reliability

I thank Rhonda Dallas and Irving Weiner for their helpful comments on an earlier draft of this article.

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 the Internet to afgjm@uaa.alaska.edu.

is poor, one would not know whether the scores fluctuate because of (a) inconsistent scoring, (b) the state-like nature of the construct, or (c) some combination of these factors. However, these are not the conditions that are found with the published CS data. Rather, with the exception of variables thought to assess transient emotional reactions, the temporal stability of CS scores is excellent. Because CS variables must be scored on two separate occasions in order to conduct a test–retest study (i.e., coding accuracy is a nested component of a retest design), it is absolutely impossible to have excellent test–retest reliability without first having excellent score assignment. Thus, contrary to Wood et al.'s assertion (1997, p. 492) that test–retest studies "do not address our criticism of CS interrater reliability," the published data squarely refute their criticism. It is illogical to believe otherwise. It is also scientifically misleading to allege that scoring reliability may be poor when there are no data to support this allegation and a body of published evidence that demonstrates how this allegation must be false.

Response-Level Reliability and Total Score Reliability

Wood et al. (1997) believe that aggregated items (i.e., CS total scores) are not inherently more reliable than single items (i.e., individual Rorschach responses). They appropriately quoted Nunnally's (1978) articulation of the aggregation principle, which is also the basis for the Spearman-Brown formula that can be found in almost any measurement text. However, Wood et al. do not believe this principle applies to Rorschach scores. There are two lines of reasoning that could have been used to support such a position. First, one could argue that Rorschach responses are not equivalent to items on other types of scales because each response is given in a relatively unique contextual field. Second, one could argue that aggregation may not apply to all CS scores because some are rare and have a limited range of values. However, Wood et al. made neither of these arguments. Instead, they argued that aggregation applies to true score theory but *not* to rater agreement, as if somehow rater agreement falls outside the bounds of true score theory. Interested readers should consult an article cited by Wood et al. (Shrout, Spitzer, & Fleiss, 1987, p. 175) that succinctly describes how true score theory applies to interrater reliability.

Wood et al. (1997) also presented an example in which the principal of aggregation fails to hold up. Consider their example to reflect the coding of a 20-item vocabulary test. Clearly, the construct being measured varies little from person to person, which is an essential requirement for norm-referenced reliability statistics such as kappa. Although the example counters my point, it also counters Nunnally's (1978) point and the assumptions of the Spearman-Brown formula.

Wood et al. (1997) said they could not "locate any published demonstration that interobserver reliability is necessarily higher for summary scores than for individual scores" (p. 492). However, they referenced a study that I cited in this regard (McDowell & Acklin, 1996). This study examined the reliability for nine CS response segments, and the authors found the following average coefficients: total protocol = .99, response-level $\%A$ = .87, and response-level kappa = .79. I believe these data are quite clear. However, as further evidence, I also pointed out (Meyer, 1997a, p. 482) that the published literature on chance-corrected reliability indicated CS summary scores were consistently more reliable than response-level scores. For additional data, see Meyer, Exner, Fowler, Hilsenroth, & Piers (1997).

My "Shaky," "Fatally-Deficient," and "Dubious" Meta-Analysis

Given Wood et al.'s (1997) incorrect criticisms regarding kappa, test–retest reliability, true score theory, and score aggregation, it is not surprising that they have mistaken qualms concerning more complicated meta-analytic procedures. Part of their difficulty reflects a rather dramatic misperception of the data I sent them. They received a six-page typed document that contained the 1,343 values I had calculated while generating the chance agreement statistics for the meta-analysis. The document is divided into sections that correspond to the response segments used in the meta-analysis. Below each section are three lines of summary data that are clearly labeled and defined at the top of the document. The first line indicates the number of independent permutations that could be derived from the scores in that segment. The second line includes the phrase "best guess approach" and reports chance agreement rates when two raters always assign the most frequent score in that segment. I did not use these two sets of values in the meta-analysis. Rather, I used the third line of data, which reports chance agreement rates when random scores are assigned in line with base rates. Because this is how kappa defines chance, I used these data in the meta-analysis to derive estimates of kappa. Chance agreement is always lower using the "best guess" approach, so Wood et al. needed to make only a quick comparison with the table of meta-analytic information in order to determine which values were used. Instead, they chose to believe that I had ignored the kappa-relevant data I had calculated in favor of data that would have been incompatible with the purpose of the meta-analysis—a truly perplexing presumption.

Wood et al. (1997) also take issue with other aspects of the procedures I used to estimate kappa values. With respect to multiplying and adding together long chains of numbers, basic probability theory (e.g., Hays, 1981) demands exactly these calculations in order to derive an accurate estimate of chance. As expected, these procedures produced differentiated and theoretically consistent results (Meyer, 1997a, p. 486). With respect to using response segments, this has been the recommended approach for some time (Exner, 1991; Weiner, 1991). With respect to generating estimates by using base rate information from existing samples, Wood et al. did not seem troubled by these procedures when they calculated chance agreement rates from one sample, applied them to Exner's data, and then used the results to suggest that CS reliability may be little better than chance (1996a, p. 4). Unfortunately, however, the authors did not acknowledge that base rates change across samples and did not present any evidence to justify applying nonpatient base rates to Exner's research samples.

Regarding my meta-analysis, I agree it would have been optimal if each study had reported base rate information for each CS variable. However, journals do not publish this level of detail. As is often the case, meta-analyses must then use estimates (see Cooper & Hedges, 1994, chap. 12 and 21; Hunter & Schmidt, 1990, chap. 4). Although Wood et al. (1997) were unaware of

this literature (p. 494), Hunter and Schmidt (1994, p. 331) have referred to approximately 50 meta-analyses in personnel psychology alone that make use of data derived in one sample in order to generate information that is then applied to other samples.

Of course, estimates must be applied appropriately. For my analysis, because some score base rates (and thus, kappa-defined levels of chance) change as a function of psychopathology, the critical issue is whether I may have misclassified some of the samples. Doing so could inflate the final estimates of kappa. Readers should closely inspect my table of classifications and the actual samples in order to allay any doubts in this regard. Also, it is important to know that the meta-analytic results remain stable despite extensive forced error. After artificially increasing the chance agreement rates by 20% in each sample (e.g., from .60 to .72), the average kappa value across segments only dropped from .86 to .83 (range = .61 to .96). Thus, Wood et al. (1997) were in error when they asserted that the meta-analytic results are shaky. Instead, the findings are quite sturdy.

Exner's Percentage Agreement Calculations

Although it was only a passing reference in my article, Wood et al. (1997, p. 491) correctly noted that my understanding of Exner's *%A* procedure differed from his explanation (1996). The mistake is mine, and I apologize for the confusion it may have caused. Exner's procedure is correct as he reported it (personal communication, J. Exner, March 31, 1997). In the language of diagnostic efficacy, his procedure provides indices of rater sensitivity to scoring rules (i.e., it calculates the proportion of correct responses actually recorded by all raters across all responses).

The Dangers of Poor Practice

Wood et al. (1997) and I are in agreement about the danger of clinicians in the field using the CS poorly. I would add, however, that this is not a concern that should be limited to the Rorschach. Rather, any test with complex administration, scoring, and interpretive guidelines can do more harm than good in the hands of poorly trained clinicians. Also, it is important to remember that the same dangers apply in a scientific context. Researchers can create mischief if they try to interpret data about a complex test while being insufficiently informed about the test, psychometric principles, or complex conditions that the test is being used to assess. This danger is compounded if broad conclusions are made from a limited subset of the available literature.

Conclusion

Although I have marshaled evidence that clearly supports the intrinsic reliability of the CS, this does not mean that anyone can accurately score the CS at will. To the contrary, the data only mean that the scoring rules provide sufficiently clear guidance so that the CS can be used reliably when coders understand those rules. It is also the case that many Rorschach validity issues still need to be resolved (see Meyer, 1996a, 1996b, 1997b). However, all of the genuine systematic reviews of Rorschach validity (Atkinson, 1986; Atkinson, Quarrington, Alp, & Cyr, 1986; Meyer & Handler, 1997; Parker, 1983; Parker, Hanson, & Hunsley, 1988) have indicated two things: (a) the Rorschach yields valid data, and (b) the Rorschach is as valid as other personality assessment methods. The challenge then is to develop a more refined understanding of the unique strengths and limitations associated with all personality assessment methods so that we can develop a more scientifically sound and differentiated understanding of personality in its full complexity.

References

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.

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

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

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

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.

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

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York: Russell Sage Foundation.

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

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

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

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

Meyer, G. J. (1997b). On the integration of personality assessment methods: The Rorschach and MMPI–2. *Journal of Personality Assessment, 68,* 297–330.

Meyer, G. J., Exner, J. E., Jr., Fowler, J. C., Hilsenroth, M. J., & Piers, C. C. (1997). *Rorschach psychometrics: I. Response-level and summary score reliability for the comprehensive system in four data sets.* Manuscript in preparation.

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.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

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.

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

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.

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

Received April 10, 1997
Revision received April 17, 1997
Accepted May 6, 1997 ■