McKinzey, R.K. (April 30, 2002). Prior peer review of "Rorschach interscorer agreement". *WebPsychEmpiricist*. Retrieved (date) from: http://www.wpe.info/papers_table

WPE WebPsychEmpiricist

Prior peer review of "Rorschach interscorer agreement"

R. K. McKinzey, Ph.D.

Oakland, CA

Abstract

The prior peer reviews of previous versions of "The Rorschach, Exner's Comprehensive System, Interscorer Agreement, and Death" are presented verbatim, together with the journals' editor's rejection letters. Alternate explanations for the results are offered; poor protocol administration, demand characteristics, patient characteristics, and unstandardized format. The generalization of the results is questioned, as well as the nature of Interscorer agreement, the CS coding system, the use of single data points, the rhetorical style, and possible bias.

Correspondence regarding this article should be addressed to : R.K. McKinzey, Ph.D., 400 29th St., Ste. 315, Oakland, CA. 94609 510-655-3903, editor@wpe.info or WPEList@yahoogroups.com

Prior peer review of "Rorschach interscorer agreement"

Prior to its publication in *WebPsychEmpiricist*, previous versions of "The Rorschach, Exner's Comprehensive System, Interscorer Agreement, and Death" were submitted to two print journals, which each provided three blind, anonymous reviews. These are presented verbatim. The Authors' Reply is at: www.wpe.info/papers_table.

The first journal was *Journal of Clinical Psychology* (JCP), Larry Beutler, Ph.D., and T. Mark Harwood, Ph.D., editors. The article was submitted 3/2/01. The editors replied 7/18/01.

Reviewer 1 wrote:

> The basic idea for the study is interesting and potentially useful, namely, have an independent set of raters score a high-stakes, forensic Rorschach Test in order to compare codes, indices, and ratios and whether raters would reach the same conclusions as the psychologist administering and interpreting the test.
>
> Twenty-eight raters examined a Rorschach Test administered in a death penalty case. Raters were polled concerning the codes and quality of administration. The protocol and original set of scores by the administering psychologist are included in the manuscript. Based on the extremely poor performance of the raters, involving considerable disagreement, including whether the protocol was scorable at all, the author concludes that the use of the Rorschach in a forensic context is insupportable.
>
> Although the study has a commendable objective, it is fatally flawed in the underlying logic and execution. The problem here is that the raters were not supplied with a clean and unambiguously administered record. Unfortunately, for example, the raters were not informed that Dr. Doe (who himself apparently is apparently in need of a scoring tutorial) scored three separate responses for Card I. Thus, the raters found themselves disagreeing over something as fundamental as the number of responses in the record and whether one response could be distinguished from another. The small group of raters who concluded that the test was inadequately administered were apparently correct. It does not appear that Dr. Doe's administration was so much at fault as the quality of the protocol provided to the raters.
>
> This would be equivalent to assessing whether a set of coders are capable of drawing conclusions from an MMPI-2 profile where the profile is unreadable (maybe retrieved from a mud puddle), or interpreting an MMPI-2 code type where the numbers are indecipherable. Perhaps a better example would be requesting raters to code prosocial behaviors amongst preschoolers from a videotape where the videotape is blurred or incomplete. It is, in fact, an impossible and unfair test and certainly does not provide the basis upon which any sort of definitive conclusions may be drawn.
>
> In a related matter, the author demonstrates a significant ignorance of interrater agreement and reliability. The point is not whether raters agree, but whether they are able to apply a coding system to an observable phenomenon, in this case a sample of behavior. In effect, the Rorschach is an observational methodology. Because the author does not understand or appreciate this, he or she has expected raters to apply a coding system to a poorly prepared sample of behavior, and then blame the coding system! A second serious flaw is the assumption that a single test would be used to make determinations in a forensic context. No self-respecting psychologist that I know would be willing to draw conclusions about a defendant's clinical condition based on a single test in isolation from other sources of information. The authors may wish to refresh their understanding of this by referring to the current Meyer et al article in the American Psychologist.
>
> The author concludes "If the scorers can't agree on the scores, of what use is the test?" A suitable rejoinder in this case would be, "If the author exhibits flawed logic and

faulty conclusions, of what use is the contribution?" For these reasons, I would consider this paper unsuitable for publication in the Journal of Clinical Psychology.

Reviewer 2 wrote:

Although this study addresses an interesting topic, the manuscript does not make a sound contribution to the literature in its present form, because the results are not placed in an appropriate context. In effect, the authors have demonstrated that when people agree to score an inappropriately administered Rorschach protocol as part of a research study, they show only modest agreement with each other, and with a computer-driven scoring system. They then imply/conclude that the Rorschach test is "useless" (page 4 and elsewhere).

The problem with the study—and the authors' conclusions—is that similar results would likely be obtained with other psychological tests as well. Suppose, for example, that one asked experienced clinicians to score and interpret a poorly administered MMPI protocol (numerous missing responses, troubling validity scale profiles, etc). Or worse, a poorly done WAIS. Would these clinicians not show similarly modest agreement in their clinical inferences? In the case of a badly done WAIS, wouldn't clinicians confronted with raw responses show modest interrater reliability?

We don't know, because the authors haven't included appropriate control/comparison tests in their study. Until they do, the present findings are essentially uninterpretable.

Reviewer 3 wrote:

In their manuscript, "Interscorer Agreement of Exner's Comprehensive System of the Rorschach: A Forensic Study," the authors presented a Rorschach protocol to a group of psychologists and had them score and interpret the protocol. Interpsychologist agreement was poor both for scoring and interpreting the protocol. The authors conclude (p.11), "Our well-qualified, experienced scorers disagreed on whether this death penalty protocol was scorable, interpretable, schizophrenic, or depressed. They made errors in scoring, and argued with the scoring program."

The manuscript has several attractive features. For example, the protocol is unique and interesting. Also, the topic of Interscorer agreement is an interesting one. There is a controversy over Exner's Comprehensive System, and it would be helpful to have new data that could help resolve it. However, I do not feel this manuscript will help to move the field forward.

One limitation of this study is that only one protocol was given to psychologists. One can wonder if similar results would have been obtained if additional protocols were used. Thus, it is difficult to generalize the results of this study.

Another limitation is that there are problems with the protocol. Many of the psychologists commented on its poor quality, e.g., they were not sure "where one response ended and the other began."

Several minor criticisms can also be made.

The authors overlook the demand characteristics of their study. They note that some psychologists "complained the protocol was of poor quality, then scored it anyway." In clinical practice, this may not have occurred. It may have occurred here because of the demand characteristics of the study: some psychologists may have felt compelled to complete the study.

The tone of the manuscript is not always objective. For example, in their Discussion (p.11), they write, "We came looking for scoring agreement. We found discord." It seems too dramatic and pointed. The same criticism can be made of the last sentence in their Discussion.

I would delete the first full paragraph on p. 10. If these differences are not statistically significant, then we cannot make anything of the direction of the difference.

JCP's editors wrote:

Chief among the reviewers' concerns was:

1. Due to the problematic nature of the protocol and the study, it is unlikely that an honest and unbiased test of the Exner system was possible. That is, the raters found the basic protocol flawed, there were questions regarding whether or not the protocol was appropriate for scoring. Because the Rorschach may have been scored/administered improperly, and because the raters worked with ambiguous and problematic data, any ratings or conclusions drawn from this data (and the study) would logically suffer from flaws as well. Please see comments from all three reviewers.
2. Reviewer #1 raises the author's incorrect conclusions based on interrater agreement and reliability. If the data applied to a coding system is problematic, the veracity of the rating system cannot be tested fairly.

After taking the reviewers' concerns into account, and my own concerns regarding the study, I have decided that I cannot accept this manuscript for publication in the *Journal of Clinical Psychology*. This decision was not an easy one because the topic of this manuscript is one of great interest to the journal. Although some of the issues raised by the reviewers might have been addressed in a revision, others could not have been.

Taking the reviewers' comments into account, the paper was revised and submitted 9/20/01 to Mary Beth Kenkel, Ph.D., editor at *Professional Psychology: Research and Practice* (PPRP). The editors rejected the paper 3/8/02.

Reviewer A wrote:

The goal of this paper seems to be to prove that the Rorschach suffers some problems in Interscorer reliability. However, the author does not accomplish that goal. The main problem was that the Rorschach record used as the basis for the study was especially badly done. When a record is so badly garbled that qualified scorers disagree as to the number of responses, one cannot believe that this is a fair test of the Comprehensive System. All it goes to show is, as the authors state, "Garbage In and Garbage Out."

While the issue is a hot one—witness all the action from the anti-Rorschach cadre in El Paso—this paper does not add a meaningful voice to the debate. The metaphor is this: if the issue were the reliability of handwriting analysis, the use of a faded, and illegible sample would not measure how well handwriting analysts agreed, but only how hard it is for anyone to make sense of an illegible sample. In this case, we learned that when the signal to noise ratio is low, people make lots of mistakes. This is not news.

If the editors determine that his paper should be published, it needs a great deal of revision. The final sections full of questions should be replaced by the author's best analysis of what the results mean. Putting an opinion on record is desirable under these circumstances and would give those who wish to put forth countervailing arguments a clear target. Second, the practice implications of the research should be explicated. The author appears to suggest that no one should use the Rorschach in forensic settings because of a lack of reliability (or because they would have a bad afternoon on the witness stand). If this is his belief, it should be stated unequivocally.

Reviewer B wrote:

I'd like to make the following comments about your article. The references are to your manuscript's page numbers.

1) pp. 6-8 – Some patients, as a function of their level of disorganization, give very complex protocols that do not, as one of your judges noted, make it clear where one precept ends and the next begins. It would appear that this was such a

protocol and, as another of your judges noted, this would serve to lower reliability artificially. Perhaps the more appropriate inter-judge measure for this protocol would have been judges' ratings of its degree of severity. It may be that its unscorability (three of your participants) or significant difficulty in scoring (several more of your participants) is the most important finding, suggesting significant levels of disorganization.

The fact that so many of your participants noted that this record was not in a format that the Comprehensive System can reliably evaluate raises significant questions as to whether this study is a test of the Interscorer reliability of the CS. It would, for example be inappropriate to give MCMI data to an MMPI scoring program. Any scoring system has expectations about the kind of raw data it is going to receive, and if the raw data deviate markedly from those expectations, the results do not represent a good test of the system's efficiency.

2) pp.10-11 – "However, if the original psychologist…errors."

It's important to note that it was not the psychologist who produced the record, but the patient. The patient's inability to organize his Rorschach responses in the way that most people do (and that the CS can reliably evaluate) is the most important finding here. The inquiry questions that several of your participants suggested would have done some of the organizing for the patient, but the significant issue is that he wasn't able to do it by himself.

3) p. 12 – "Did our…" This paragraph brings up a variety of issues (participant dropout, Dies study, Wood/Meyer hypotheses, renorming approaches) without elaboration.

4) p. 15, footnote 9 – Comprehensive assessment would involve using a variety of techniques (testing, clinical interview, review of history, discussion with collateral individuals) and looking at convergences and contradictions among them as a way of reaching a formulation about psychopathology. Certainly, no single data point such as the SCZI or the DEPI would be used to make a yes-no decision about a diagnosis.

What the Rorschach does allow is a description of some basic components of psychological function. For example, all of your participants agree that the patient has a low Lambda (Table 4, p. 23), suggesting that he has difficulty operating in an affect-free, objective manner; all agree that he has at least one reflection response in his record, suggesting significant self-focus. This is the level at which a test like the Rorschach is most useful.

Reviewer C wrote:

The focus of this manuscript is to exemplify problems with Rorschach intercoder agreement and the resulting adverse consequences for individuals undergoing forensic evaluations that include use of this instrument. The author provides a provocative title ("The Rorschach…and Death"), apparently reflecting the use of a protocol from a Death Row case, but reminiscent of Muriel Lezak's (1998) (sic) "IQ: R.I.P." essay and Garb's (1999) call for a moratorium on the use of the Rorschach.

In recent years, various researchers have expressed conflicting opinions about the adequacy of Rorschach coding as part of the broader debate over the Rorschach's utility and viability. In this context, the goal of this manuscript is relevant, and undertaking an empirical investigation of this issue offers the possibility of obtaining edifying results. Unfortunately, the investigation is seriously flawed at the initial stage of test administration, which renders all subsequent findings and conclusions irrelevant. Specific points are listed below:

1. At the most basic level, the number of responses obtained in the test administration is in question, as is overwhelmingly indicated by the participants who attempted to code this protocol. In a standard administration, the Rorschach

examiner is required to record the response number, which would prevent the type of confusion seen in this case. The response-collection phase of the testing was apparently not regulated by the test examiner in terms of determining when a specific response ended and another one began. It inevitably follows that the accuracy of coding, particularly with reference to determinants and special scores, would be come reduced. Yes, "garbage in, garbage out" is an apt phrase which, in this instance, applies to test administration procedures, but I don't believe this is the point the author was undertaking to demonstrate.

2. The author fuses the issues of coding *agreement* and coding *accuracy* when discussing the coding errors made by participants, which muddles the focus of this manuscript. In my opinion, the coding errors in this case likely relate predominantly to the test administration problem, which makes a true evaluation of intercoder agreement impossible. Furthermore, there is no indication of what analyses were conducted to determine that "the scorers with a Psy.D. made more errors than the scorers with a Ph.D.…) (p. 9) and no rationale is given for undertaking this analysis, therefore giving the impression that this was a "fishing expedition."
3. There is no empirical basis to the author's statement that the protocol obtained by Dr. Doe is representative of those obtained by other Rorschach examiners. In fact, the participant coders' comments make it abundantly clear that Dr. Doe's administration is an *anomaly*. Dr. Doe's credentials aside, the best evidence of Rorschach testing skill comes from the test record itself.
4. As already indicated, the conclusion that "…the practice of using the Rorschach in a forensic context is no longer supportable…" (p. 11) lacks a good foundation in this investigation (and hints at the possibility that the author may have leaned toward this conclusion before the investigation was begun). Rather, the example used in this manuscript underscores the importance of following standardized test procedures. The issue, then, is not about whether the Rorschach should or should not be used in forensic evaluations, but about how to ensure that forensic examiners use *defensible methods* for any and all tests used by them.
5. The rhetorical questions raised in the discussion section, especially those referring to Dies' findings and Meyer's and Wood's arguments, are likely to be unclear for readers who have not been following the series of articles published in assessment journals in the last few years.
6. I have two final points of concern:
    a. The author appears to have misrepresented Acklin et al.'s (2000) results by way of selective emphasis. For instance, the author writes that Acklin et al. found 20% of score combinations in the normal sample, and 9% in the clinical sample, had unacceptable levels of unreliability (p. 4 of manuscript). The fact that these convert to strong overall rates of reliability (80% and 91%, respectively) is not discussed. In actuality, Acklin et al. state, "…most Comprehensive System codes, coding decisions, and summary scores yield acceptable, and in may instances excellent, levels of reliability" (quoted from their article abstract, p. 15) and "we believe that this study provides strong evidence for the reliability of the Rorschach Inkblot Test" (Acklin et al., 2000, p. 43). Moreover, the author's discussion statement that Acklin et al. "…found less than perfect interscorer reliability rates for Exner's Rorschach CS…" (p. 10) is not informative. "Less than perfect" is a phrase that would apply to *every* test currently used, be it psychological, medical, or otherwise.
    b. My impression is that Dr. Doe is someone other than the author. I advise the author to report that permission was obtained from Dr. Doe and the client in question to reproduce the Rorschach protocol presented in this manuscript.

PPRP's editors wrote:

> …Your manuscript deals with a much-debated topic that potentially would be of interest to our readers. However, there are significant concerns with your manuscript. All reviewers agree that little can be learned from your study about the reliability of Exner Comprehensive System because the Rorschach record used in the study was done so poorly. This is a critical flaw so no meaningful conclusions about scoring reliability can be reached.
>
> Because of this concern, we will not further consider your manuscript for publication and are closing the file.

The paper's authors' reply can be found at: www.wpe.info/papers_table.

References

Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach Inkblot Test in clinical and forensic settings. *Assessment, 6*(4), 313-317.

Lezak, M. D. (1988). IQ: R.I.P. *Journal of Clinical & Experimental Neuropsychology, 10*(3), 351-361.

Acklin, M. W., McDowell, C. J., Verschell, M. S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment, 74*(1), 15-47.