Accepted November 22, 2004; text = 76832 = 11.75 pp; 11 tables = 7 pp; TOTAL = 19 pages

Running Head: RELIABILITY OF RORSCHACH INTERPRETATION

The Inter-Clinician Reliability of Rorschach Interpretation in Four Data Sets

Gregory J. Meyer and Joni L. Mihura

University of Toledo

Bruce L. Smith

University of California, Berkeley

Author Notes: Correspondence concerning this article can be sent to Gregory J. Meyer, Department of Psychology, University of Toledo, Mail Stop 948, Toledo, OH 43606. Electronic mail may be sent to gmeyer@utnet.utoledo.edu. Portions of this research were presented at the annual meeting of the Society for Personality Assessment, New Orleans, LA, March 26, 1999, and Albuquerque, NM, March 25, 2000.

We are very grateful to the clinicians who provided Rorschach interpretations. Kathie A. Bates, Claude Della Zazzera, Mark-David Janus, Sharon Rae Jenkins, Carol Klopfer, Pam Olsen, Keith Rieder, and John Wallace, each interpreted 65-66 protocols for Data Sets C and D. Bill Alexy, Dan Cady, Phil Caracena, Suzanne Martin, John Newbauer, Piero Porcelli, Chris Sheaffer, Veronica Vivero, and David Zita, each interpreted 10-11 protocols for Data Set C.

Abstract

To examine agreement on Rorschach Comprehensive System (CS) interpretations, 55 patient protocols were interpreted by 3-8 clinicians across four data sets on a representative set of 29 characteristics. Substantial reliability was observed across data sets, though a problematic design produced lower results in one. Unexpectedly, a Q-sort task had slightly lower reliability than a simple rating task. As expected, scales that summarized judgments had higher agreement than judgments to individual interpretive statements and some clinicians produced more generalizable inferences than others. Interpretations for all clinicians were more strongly associated with patients' psychometric true scores (aggregated judgment *M* range = .82-.92) than with the judgments of other clinicians (range = .76-.89). Compared to meta-analyses of interrater reliability in psychology and medicine, the findings indicated these clinicians could reliably interpret Rorschach CS data.

The Inter-Clinician Reliability of Rorschach Interpretation in Four Data Sets

Recent studies have examined interrater reliability for scoring Rorschach protocols (e.g., Acklin, McDowell, & Verschell, 2000; Meyer, 1997; Meyer et al., 2002; Viglione & Taylor, 2003). Logically, it is necessary to have reliable scoring prior to using the Rorschach for other purposes, such as clinical interpretation or empirical research. The available data indicate trained raters can reliably score according to the Comprehensive System (CS; Meyer et al., 2002; Viglione & Hilsenroth, 2001). To extend the research on interrater reliability, we examine how well clinicians agree on the interpretation of CS scores. The key question is whether different clinicians derive similar inferences when interpreting patient protocols.

Many studies have examined clinical judgments in psychology (e.g., Borum, Otto, & Golding, 1993; Garb, 1998; Hammond, 1996; Spengler, Strohmer, Dixon, & Shivy, 1995) and medicine (e.g., Elmore & Feinstein, 1992; Koran, 1975a, 1975b), often investigating factors that may bias judgments (e.g., Chapman & Chapman, 1969) or the accuracy of clinical decisions compared to actuarial equations (e.g., Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954/1996; Westen & Weinberger, 2004). David Funder and his colleagues have advanced the most comprehensive approach to understanding the accuracy of judgments about people (e.g., Blackman & Funder, 1998; Funder, 1995, 1999; Kolar, Funder, & Colvin, 1996).

A more limited body of research has examined the reliability of interpretations derived from psychological tests. In a preliminary review, Mihura, Meyer, and Wright (2002) found 70 samples from 55 studies that reported data on the agreement between two or more judges interpreting tests of personality or cognitive functioning (cognitive tests = 9, MMPI = 17, Rorschach = 23, TAT = 8, figure drawings = 5, sentence completion tests = 4, or a battery of instruments = 4). The first studies appeared in the early 1940s and the majority were published in the 1950s and 1960s. Only four studies were published in the 1980s and none more recently than 1985, indicating that research on this topic had fallen out of favor over the last two decades.

Almost all studies used a design in which judges were asked to rate, classify, or categorize patients on experimenter-specified constructs (e.g., extent of impairment, degree of psychosis, treatment prognosis), with agreement computed across judges on these target constructs. This kind of design was used in 68 of the 70 samples in Mihura et al.'s (2002) review. Just 3 samples used an alternative design (one also examined experimenter-defined constructs), all of which were Rorschach studies conducted between 1942 and 1955. They used a matching task in which clinicians attempted to pair reports or narrative statements to the actual protocols on which they were based.

In general, matching tasks are limited because they do not provide differentiated reliability coefficients for distinct inferences. Rather, matching requires judges to make a single decision per patient (i.e., whether the narrative matches the protocol) and this decision may be based on any of the information contained in the narratives and test protocols. However, there is no way to determine what bits of data actually prompt the judges to match objects in the manner they do. This means that minor but highly 'diagnostic' clues in the data may lead clinicians to agree on how the narratives and test protocols should be matched even though the clinicians might not agree on many (or most) of the remaining characteristics in the narratives or test data. For instance, if one report said the patient was preoccupied with sex and only one Rorschach protocol contained multiple sex responses, this single piece of information would stand out as a highly diagnostic clue. Even though clinicians may not agree on other constructs embedded in the data, they would be given credit for correctly matching all features in this test protocol and report. When this occurs, reliability can be artificially inflated above what would be found if the judges evaluated a broad range of specific, patient-relevant constructs. The reverse can also be true. A minor feature embedded in a complex narrative may prompt one judge to make an atypical or incorrect match, even though this judge may have actually agreed with other judges on a wide range of specific characteristics. Given these issues, it seems more informative to examine interpretive agreement using a range of specific, test-relevant constructs rather than a global matching task.

Studies that examine agreement on test interpretation often gather data with rating scales or Q-sort distributions (see Block, 1961; Ozer, 1993). Rating scale items have dichotomous (e.g., true-false) or dimensional (e.g., 1 to 5 scale) response options that quantify the extent to which a person possesses certain characteristics. Each item is rated independent of the others, so a person may be high, low, or average on all characteristics. There is no requirement that certain characteristics be considered more defining and others less defining of the person. These qualities make rating scales ideal for comparing one person to another person.

In contrast, the Q-sort methodology is explicitly designed to quantify the relative importance of characteristics for a single individual. It does so by having descriptive items sorted into a distribution that ranges from least to most defining of the target person. Q-sorts are thought to provide a more idiographic portrait of personality than ratings because the final distribution indicates the features that most uniquely identify a person. In addition, it takes more thoughtful effort (and time) to complete a Q-sort because each descriptive item has to be compared to all the others before it can be properly placed in the final distribution. Rating scales, in contrast, only require that each item be evaluated on its own merits.

When completing a rating scale, judges make as many decisions as there are items on the scale. For instance, a rating scale with 100 items requires 100 judgments. When completing a Q-sort, however, judges make many more decisions because items are classified via paired-comparisons. For instance, when 100 items are sorted into 9 categories that approximate a normal distribution, a judge must make 4,349 determinations (Ozer, 1993). Because each determination is comparative, the 100 constructs and their placement are reviewed multiple times in the context of creating the final Q distribution. This process should produce fewer errors because faulty placements are likely to be detected and corrected. Accordingly, one could anticipate that Q-sorts may lead to more reliable test interpretations; a finding observed in one prior study (see Ozer, 1993).

Overview of the Current Studies

To address interpretive reliability for the Rorschach CS we conducted two studies, each of which used two separately collected data sets. In Study 1, the three authors interpreted protocols using both rating scale and Q-sort methods. Study 2 was designed as a partial replication and it relied on 8-17 practicing clinicians to make rating scale judgments. None of the clinicians in Study 2 were involved in the design or execution of Study 1. For both studies the same pool of 55 Rorschach structural summaries were interpreted. Table 1 provides an overview of both studies, the four data sets, and the issues addressed in each sample.

In Study 1 the Rorschach interpretations for Data Set A were made on 29 constructs using a 5-point Likert-type scale. Interpretations for Data Set B were derived from a Q-sort in which the same 29 constructs were sorted into a quasi-normal distribution that indicated what was least to most characteristic of each patient. The ratings allowed each interpretive statement to be considered independent of the others. In contrast, the Q-sort required all of the characteristics to be considered simultaneously in order to determine the relative importance of each. We expected that the more judgment-intensive Q-sorts would lead to higher levels of interpretive agreement.

In Study 2, although the clinicians rated only the Likert-type items, we again obtained two independently collected data sets. Data Set C had a complex methodology in which 17 clinicians were randomly assigned to serve as the 1st, 2nd, or 3rd judge for 10 or 11 of the 55 protocols. Statistically, by randomly mixing clinicians across rater positions and protocols, the design assumes that each clinician will use identical cognitive benchmarks when assigning values on the 5-point Likert scale. If judges do not use the same benchmarks, reliability correlations will be lowered. At the outset of data collection, we did not appreciate how this assumption may impact the final results. Consequently, a second data set was collected for clarification. In Data Set D, interpretations were again made by a group of practicing clinicians. However, only eight clinicians were used and each interpreted all 55 Rorschach protocols. This design paralleled that used with Data Set A and it allowed clinicians to use slightly different interpretive benchmarks when completing the Likert-type scale without compromising reliability.

Because Data Sets A and D used the same methodology, analyses were conducted across data sets to address three issues. The first considered the extent to which individual clinicians differed in their overall reliability. The second explored aspects of the classical true score theory of reliability. Multiple clinicians provided ratings on the same target patients so it was possible to average their ratings to reduce random measurement error and obtain approximate psychometric true scores for each patient. These data allowed us to determine if clinicians produced higher agreement with patients' true scores than with each other. Finally, the CS interpretive reliability results were placed in context by considering them relative to a recent summary of 42 meta-analyses examining interrater reliability in psychology and medicine (Meyer, 2004).

General Methodology Across Studies and Data Sets

Target Patients

All data sets relied on CS scored Rorschach protocols from 55 psychiatric inpatients. The patients were selected from a larger group based on factors that may have been useful if subsequent research was conducted with these protocols. Patients were selected if they had a computerized Rorschach record with more than 13 responses, an MMPI-2, a TAT, and inpatient records from their hospitalization. All patients meeting these criteria were used. With the exception of R < 14, no patients were excluded based on their Rorschach scores.

On average the patients were 34.3 years old (*SD* = 12.97), 51% male, 20% married (64% never married), 58% European American (27% African American), and had 14.6 years of education (*SD* = 2.99). Diagnostically, 56% had a disorder on the psychotic spectrum and 87% had a disorder on the depressive spectrum. This background information was gathered when writing this article and no demographic data was given to the clinician raters. A code number identified each protocol and the only clue about the nature of the sample came from using the word "patients" once in our consent form and task instructions.

Patients had the following average Rorschach scores: R = 25.05 (*Mdn* = 22, *SD* = 11.07), Form% = .39 (*SD* = .19), Sum6 = 9.05 (*SD* = 7.42), X+% = .42 (*SD* = .12), and X-% = .26 (*SD* = .11). Twenty-five patients (45.5%) were introversive and 12 (21.8%) were extratensive.

Interpretive Statements

A priori we decided to have clinicians provide judgments on about 30 statements to make the task manageable and increase the prospect that each interpretation would be carefully considered. Ultimately, we used 29 statements because it produced a more normal distribution for the Q-sort task.

The potential item pool began with 166 CS-related statements in the Rorschach Rating Scale (RRS). The RRS was developed for validation research and contains constructs thought to be measured by a variety of Rorschach scoring systems (Meyer, 1996; Meyer, Bates, & Gacono, 1999; see Mihura, Meyer, Bel-Bahar, & Gunderson, 2003, for an updated version of the scale). The 166-item pool was reduced to 77 items by excluding statements that: (a) addressed experimental or nontraditional CS interpretations, (b) were specific to a criterion within a CS constellation index, (c) addressed very rare scores (e.g., CP, FQ+) that were unlikely to apply to most patients, and (d) did not address scores from the lower section of the structural summary that contains the ratios, percentages, and indices that are central to interpretation. Next, we deleted repetitive content, defined as instances when a very similar construct was assessed by more than one statement. For example, we deleted all but one item assessing narcissistic qualities. This winnowing reduced the item pool to 41, but still ensured broad coverege of CS constructs. The final 29 interpretive statements were randomly selected and are listed in Table 2.

Based on psychometric theory, we expected that higher order constructs would be more reliable than interpretations made for single items (Meyer et al., 2002). Although highly redundant content had been deleted from the item pool, higher order constructs were created from the 29 items through both factor analysis and rational aggregation.

To identify factors we used 213 RRS ratings from the clinicians, friends, spouses, family members, or co-workers of target subjects (Meyer et al., 1999). Items were rated on a 5-point Likert scale and were subjected to a principal axis factor analysis and a principal components analysis, with both varimax and oblique rotations. All extraction and rotation solutions were virtually identical, so only the principal axis solution with oblique rotation (oblimin, delta = 0) is described. The number of factors to extract was determined first by parallel analysis (Zwick & Velicer, 1986) and then by Guadagnoli and Velicer's (1988) recommendation to retain factors that have at least four variables with loadings above .60 or 10 to 12 variables with loadings above . 40. Parallel analysis using 25 data sets indicated three factors should be retained. However, in all solutions, the third factor did not meet Guadagnoli and Velicer's retention criteria so only two were extracted. They correlated .28 with each other. Aggregated judgments were created by summing the interpretive ratings for all variables that had a loading above .64 on each factor. The upper portion of Table 3 lists the items, internal consistency, and average inter-item correlation for the scales. Based on item content, Factor 1 was seen as a dimension of Perceptual Distortions and Thought Disorder, while Factor 2 was viewed as a dimension of Negative Emotionality.

Independent of the factor analysis the first author organized the 29 items into meaningful broad constructs, while trying to ensure all items were used. Four aggregate scales were created, one of which duplicated the factor-derived Perceptual Distortions and Thought Disorder scale (see the lower portion of Table 3). Two scales, General Distress/Dysfunction and Poor Coping, had reasonable internal consistency. Defensive Idealization/Intellectualization had just three items, which contributed to its relatively low alpha. Because we wanted to aggregate as many items as possible, this variable was retained. General Distress/Dysfunction shares items with Negative Emotionality, but it is broader and more inclusive. Only two of the 29 RRS items (39 & 167) were not included in a broader construct category.

Data Integrity

Following Wilkinson and the APA Task Force on Statistical Inference (1999), prior to running analysis the data sets were checked for accuracy and to identify anomalous data points. Ratings were independently entered into a computer program twice and compared to detect entry errors. Q-sorts were entered only once because the fixed distributions allowed for an easy check of entry errors. One rater in Data Set B and another rater in Data Set C produced marked discrepancies when rating item 39. Follow-up investigation revealed that both raters often coded the item in the direction opposite of what they intended so these mistakes were corrected.

Study 1 Methodology

Clinicians and Interpretative Tasks

The three authors served as interpretive judges for Data Sets A and B and the CS Structural Summary page served as the source of Rorschach information. Judges knew the protocols were from patients but did not have other identifying information. The judges worked independently to interpret all 55 protocols (i.e., without discussion and blind to each others ratings). For Data Set A, interpretations were made by rating the 29 items in Table 2 using a 5-point Likert-type scale with the following response options: -2 = *False, Much Less Than Average*; -1 = *False, Somewhat Less Than Average*; 0 = *About Average*; 1 = *True, Somewhat More Than Average*; 2 = *True, Much More Than Average*.

Task instructions explained the study was evaluating the extent to which different clinicians agreed on the information that could be derived from a CS Structural Summary. Specific instructions continued as follows:

> For this study, please begin by carefully reviewing the patient's structural summary. Review it in the same way that you would if you were seeing this patient as part of your clinical practice. Once this is done, answer each question on the rating scale as accurately as you can.
>
> Because we are all prone to make global judgments that may be biased, please take the time to think about each item and consider all the relevant CS information before making your judgment.
>
> Also, when making your ratings, it is important to have a clear benchmark in mind. Please compare this person to what you think is characteristic of an *average person* (not an average psychiatric patient).
>
> If you are not confident about your rating on a particular item, please give your best guess.

Finally, the instructions noted that the 5-point Likert scale was symmetrical and bi-polar, with options extending as far below average as above average. Many Rorschach scores are rare and have skewed distributions in which the modal and median score is zero. Because the typical person obtains a score of zero, it is not possible to classify someone as being less or much less than average on this characteristic. Theoretically then it should not be possible for a clinician to discriminate between the rating scale points of -2, -1, and 0 for such a characteristic. In an effort to contend with this issue without biasing raters by presenting a different Likert scale for certain items (e.g., 0 to +2, rather than -2 to +2), raters were told "You may decide this type of scale does not make sense for all of the items presented below. If so, simply use the portion of the scale that makes sense to you." Ratings were made directly onto a record form containing the 29 items.

For Data Set B, the interpretive task consisted of a Q-sort. Statements in Table 2 were printed on small laminated cards approximately 1" x 3" and for each patient the 29 cards were sorted into seven piles of fixed size that approximated a normal distribution. In addition to the general instructions offset above, clinicians received the following directions:

> To rate each person, you will be using a Q-sort task. To complete the Q-sort for each patient, you must organize the 29 item-cards into a frequency distribution that is approximately normal in shape. The distribution will have 7 columns and these columns will range from "Least Characteristic" on the left to "Most Characteristic" on the right. As shown in the figure below, you will place a set number of item-cards in each column. From left to right, the number of cards

in each column will be: 1, 3, 6, 9, 6, 3, and 1.

The accompanying figure depicted an approximately normal distribution with 29 cells arranged in the format described. For scoring, the Least to Most Characteristic columns were assigned scores from 1 to 7, respectively. Two judges completed Q-sorts prior to the ratings. The other completed the ratings before the Q-sorts. Several months separated each judgment task.

Data Analysis

Our primary agreement statistic was *r*, the traditional Pearson correlation. As a measure of the linear association between raters, it was viewed as the most appropriate index for these studies. Correlations were computed between each pair of clinicians (i.e., A with B, A with C, and B with C) for the 29 individual item-level judgments and the 5 aggregated judgment scales.

We also examined two intraclass correlations (ICC; see McGraw & Wong, 1996), one of which examined consistency and the other absolute agreement. Both ICCs employed a two-way random effects model and determined reliability for a single rater (i.e., one rater with another). The consistency ICC is similar to *r* in that it disregards the level of the scores assigned across raters (e.g., one rater's scores could always be 1/2 point higher than another rater's). As such, raters are allowed to employ somewhat different benchmarks when assigning ratings. The consistency ICC differs from *r* in two ways. First, a single statistic can be computed across multiple raters. Second, consistency ICCs are lowered by unequal variances across judges. Because the clinicians in Data Set A were instructed to use only that portion of the Likert scale that seemed sensible to them, it is likely that each clinician made unique decisions about how much of the 5-point scale should be used, which in turn would produce unequal variances. As such, the consistency ICC could be expected to be slightly lower than *r* in Data Set A. The results from both statistics should be similar in Data Set B because the forced distribution in the Q-sort task would produce more consistent item variances across raters (see Ozer, 1993).

The agreement ICC is a chance-corrected reliability coefficient that is suitable for continuous data and asymptotically equivalent to Cohen's weighted kappa (Fleiss, 1981). The agreement ICC differs from *r* and the consistency ICC in that it counts any discrepancies across raters as error. As such, it requires judges to have an identical understanding of the reference criterion and the anchors on the rating scale. It also requires judges to use the same degree of variability when assigning scores. Although these requirements are very appropriate under many circumstances, such as when scoring a test, it is unlikely these assumptions were met in the current study because a) we did not characterize an "average person" or ensure that judges used this as their benchmark standard rather than an "average patient," b) the Likert and Q-sort scales did not define scale anchors in an explicit or meaningful way (e.g., there were no item-specific benchmarks indicating the criteria that should differentiate *False, Much Less Than Average* from *False, Somewhat Less Than Average*), and c) judges were instructed to disregard portions of the Likert scale if that seemed sensible. As a result, the agreement ICC should produce lower estimates of reliability. Following Fagot (1991; also see McGraw & Wong, 1996), we believe these results will be low for artifactual reasons (related to imprecise definitions in our benchmark criterion and rating scale and by permission to discard portions of the scale) rather than for substantive reasons (related to the judges' inability to consensually discriminate among patient characteristics) and do not believe the agreement ICC should be emphasized or given primary interpretive significance. Nonetheless, we recognize how frequently the agreement ICC is used in the literature and report these results for the sake of completeness.

Statisticians have provided rules of thumb for interpreting the absolute agreement ICC. Values greater than .74 are considered excellent, values from .60-.74 are considered good, values from .40-.59 are considered fair, and values below .40 are seen as poor (Cicchetti, 1994; Fleiss, 1981). Similar rules of thumb have not been proposed for the consistency ICC or Pearson's *r*.

Testing for Spuriously Inflated Reliability

Based on input from a reviewer, a final set of analyses examined whether clinicians may have agreed on interpretations just from knowing the Rorschach protocols came from patients. At issue is whether interpretive agreement may emerge in the absence of any specific Rorschach data simply because all patients share certain levels of pathology. To examine this we created artificial item level ratings that matched the item-level ratings actually obtained from the Study 1 clinicians. At the first step, we obtained the M and SD for Clinician A's ratings on each of the 29 items across all 55 patients. Using the M and SD for an item as seed information, we had SPSS generate a new "item" containing random data for 55 "cases," imposing the restriction that when averaged across cases, the random item had to have the same M and SD as Clinician A's actual item rating. This was done sequentially for each item until the data set contained 29 columns (i.e., "'items") and 55 rows (i.e., "cases") of artificial ratings. These artificial ratings matched the base rate of item endorsements for Clinician A's actual interpretive ratings. Next, we performed the same steps for Clinicians B and C, which resulted in an artificial data set that still contained 55 rows but now had 87 columns (i.e., 29 "items" for 3 clinicians). Next, we computed composite scales from the artificial items, just as we had with the actual items. Finally, the three sets of artificial scores corresponding to Clinicians A, B, and C were correlated to determine if the clinician interpretations could produce artificial levels of agreement when disregarding the actual Rorschach data. These analyses produced unambiguous findings so the same procedures were not applied to any other data sets.

Study 1 Results and Discussion

Overall Findings. Tables 4 and 5 provide results for the 5 aggregated judgments and 29 item-level interpretations, respectively, with Likert ratings on the left and Q-sorts on the right. The summary section of each table indicates that these three clinicians had good to excellent reliability when interpreting the 55 Rorschach protocols.

The center data column in Tables 4 and 5 provides results averaged across the three sets of artificial ratings developed for Data Set A. The average reliability for the artificial ratings was -.04 for aggregated judgments and -.02 for item-level interpretations. Because these values are essentially zero while the average correlations for the genuine interpretations were .88 and .79, respectively, it is clear that the observed reliability findings were not artificially inflated as a result of all patients possessing certain common characteristics.[1]

Table 5 shows that item 57 is an outlier. Agreement was poor for this item, regardless of whether it was considered in the rating task (*M r* = .16) or the Q-sort task (*M r* = .34). Thus, each author interpreted this item in a distinct and idiosyncratic fashion. Low reliability may indicate the item was poorly written or addressed a complex construct that was interpreted in multiple ways. In either case, this item's uniquely poor reliability suggests it could have been eliminated from further analyses. Had this been done, the average item reliability would have increased by about .02 (e.g., *M r* from .79 to .81).

As expected from psychometric theory, aggregated judgments (Table 4) were more reliable than the interpretations made for individual items (Table 5). Also as expected, *r* and the consistency ICC produced comparable estimates of reliability, while the absolute agreement ICC produced lower results. The difference was particularly noticeable on the rating task, which did not structure clinician responses into a fixed distribution. Clinicians could readily use different benchmarks when assigning judgments on the rating scale and these differences were considered error by the absolute agreement ICC.

Clinician Differences When Using the Rating Scale. As noted above, we believed the absolute agreement ICC would produce lower results because each clinician's propensity to use different implicit benchmarks and/or portions of the Likert scale would be treated as errors. To determine if clinicians used the 5-point scale differently, means were computed across RRS item-level and aggregate-level judgments for each clinician. Table 6 illustrates these differences using several example item-level constructs and all the aggregated judgments. It can be seen that Clinician C consistently used the lower end of the Likert scale, while Clinician A consistently used the upper end of the scale. As indicated by the final column, some of the differences between these two clinicians approached or exceeded a full standard deviation (i.e., Cohen's $d >= 1.0$), which are large differences. Overall, the average *d* comparing these two raters was .36 for item interpretations and .68 for aggregated judgments.

The key question, however, is whether Clinician C's lower ratings were inherently more correct than Clinician A's higher ratings, or vice versa. We do not believe it is possible to determine the answer. The Likert scale had ambiguous anchors (e.g., 1 = *True, Somewhat More Than Average*; 2 = *True, Much More Than Average*) rather than anchors tailored to fit each statement (e.g., 1 = *Mildly Self-Centered*; 2 = *Significant Narcissism*) and the anchors were not accompanied by clearly articulated benchmarks. Also, clinicians were not provided with a definition of what characteristics should be considered indicative of an average person and it is not clear if they consistently used an average person as their reference standard. Given these factors, we believe it is not appropriate to treat each judge's proclivity for certain regions of the rating scale as errors of unreliability per se.

These issues can be empirically illustrated by using ipsative scores rather than raw scores. Ipsative scores are standard scores computed for each patient on a per rater and per item basis. They retain all information about the extent to which a patient is judged to have a characteristic while also controlling for the rater's style of using the Likert scale. Because ipsative scores are a simple linear transformation of raw scores (i.e., [assigned rating - *M*] / *SD*), they do not affect the Pearson correlations reported in Tables 4 and 5. However, when ICCs are computed on ipsatized scores, results differ from those presented in the tables. For item-level analyses, the consistency ICC results became identical to the *r* values reported in Table 5, while the absolute agreement ICCs became virtually identical, with a maximum difference of .01. When averaged across all item interpretations, *r* and both of the ICCs produced identical results (e.g., *M* = .79 in Data Set A). For the aggregate-level analyses, the absolute agreement ICC and consistency ICC produced results that were identical to all of the *r* values in Table 4 (e.g., in Data Set A all coefficients were

.94 for General Distress-Dysfunction).

Q-Sorts vs Ratings. Tables 4 and 5 show that Q-sort interpretations were not more reliable than those from the rating task. This contradicted our hypothesis. Even though the Q-sorts required refined decisions about how each item should be placed relative to all the other items and took much more cognitive effort to complete than the ratings, the extra effort did not result in improved agreement across clinicians. A partial explanation for this may be the requirement that the 29 items had to form a normal distribution. All patients had to be classified as possessing one Least Characteristic quality, three near-least characteristic qualities, six less characteristic qualities, nine characteristic qualities, six more characteristic qualities, three near-most characteristic qualities, and one Most Characteristic quality. Although some patients may be adequately described by this kind of fixed distribution, others are not. Instead, for some patients an accurate Rorschach interpretation may generate a highly skewed or even bipolar distribution. For instance, a patient may legitimately possess 2 qualities that clearly are least characteristic, 10 qualities of uncertain importance, 10 qualities that are moderately characteristic, and 7 qualities that clearly are most characteristic. This distribution would be decidedly non-normal.

Thus, although the Q-sort forced clinicians to interpret all items in an idiographic manner, the task imposed a nomothetic distributional requirement on the results. By forcing the 29 qualities to be arrayed in a normal curve, the patient descriptions generated for this research task may have differed substantially from the descriptions that would have been naturally generated in applied clinical practice (see Westen & Shedler, 1999, for a discussion of this issue). Each of us experienced this mismatch as we completed the Q-sorts. For many patients, the distribution felt arbitrary and did not conform to the type of descriptions that we wished to make and would have spontaneously generated if we had been considering the case data in a non-research context.

Summary. Overall, the findings from Data Sets A and B indicated that several clinicians could reliably interpret Rorschach protocols. Reliability was present for single interpretive judgments and for aggregated judgments of a broader construct. In addition, judgments were reliable when the clinicians independently considered each construct (i.e., Likert ratings) and when they considered all features of a Rorschach protocol simultaneously (i.e., Q-sorts). Results also supported the psychometric expectations that reliability would be higher for aggregated judgments than for item interpretations and that reliability would be lower when a model containing questionable assumptions was applied to the data (i.e., the absolute agreement ICC).

In order to assess the generalizability of these findings, it is important to determine whether similar results would be found in another sample of clinicians. Study 1 relied on a small set of collaborators and it is possible they produced somewhat atypical results. With this in mind, as Study 1 progressed, Study 2 was initiated to sample a broad range of clinicians who actively used the Rorschach in their clinical practice.

Study 2 Methodology

Overview of Research Designs for Data Sets C and D

Study 2 was first conceptualized as a replication that would roughly parallel Study 1, with the same 55 protocols interpreted by three different clinicians per patient using the same instructions as before. However, to ensure conscientious participation by a diverse group of clinicians, we asked them to volunteer a limited amount of time. Because the Q-sort was taxing and time-consuming, only the rating task was used. In addition, to ensure the number of interpretations was manageable, clinicians were randomly assigned to just 10 or 11 of the 55 patients. Each clinician was also randomly assigned to serve as the first, second, or third person interpreting a protocol. Thus, at the outset, Data Set C was to have 55 protocols independently interpreted three times by clinicians who were randomly assigned to patients and to the 1st, 2nd, or 3rd rater positions. Ultimately, 17 clinicians provided judgments for this analysis.

We later realized that by randomly mixing the clinicians who rated each protocol the design of Data Set C paralleled the expectations for the absolute agreement ICC. As such, reliability correlations would be lowered unless each clinician used identical anchors when assigning values on the Likert scale.[2] Because results presented earlier demonstrated that clinicians use different anchors, Data Set D was collected to correct the problematic design. Rather than randomly assigning clinicians to protocols and rater categories, each clinician in Data Set D interpreted all 55 Rorschach protocols. Eight clinicians contributed ratings to this data set.

Clinicians Contributing Judgments For Data Set C

To obtain a broad range of clinicians, an invitation to participate was posted on the Rorschach Discussion List (located at rorschach@maelstrom.stjohns.edu). Out of 18 initial volunteers, one was a student several weeks into her first Rorschach course. She had almost no interpretive experience so was excluded. One of the 17 remaining clinicians did not return the consent form, reducing the initial pool to 16. Each clinician received 10 or 11 unique protocols so that each of the 55 patients would be rated independently by 3 of the 16 clinicians (i.e., 165 total interpretations). However, the initial plan was altered following several complications. The clinician who initially did not return the consent form subsequently wished to participate. Rather than exclude him, the pool of clinicians was raised to 17. However, after six months, only 14 of the clinicians had completed their interpretations, which left numerous gaps in the data set. To rectify this, a second request for participation was posted to the discussion list. Fourteen clinicians replied and the first two respondents were included in the study. After materials were sent to the new clinicians, a judge who appeared to have withdrawn submitted completed materials, which again brought the total number of clinicians up to 17.

In several instances ratings for individual protocols were missing. One clinician would not interpret a protocol that had just 14 responses and a Lambda score of 6.0. The other two clinicians who received this protocol also questioned its interpretability (as did the judges in Study 1). One wrote that the proper judgment for most questions was to state "unknown." The other clinician wrote that any interpretation would be of questionable validity. Given these considerations, this protocol was omitted from subsequent analyses with Data Set C. (Excluding this protocol had a trivial impact on the results; all correlations changed by no more than +/- .02.)

Ultimately, Data Set C consisted of 2 protocols that were rated by two clinicians, 32 rated by three clinicians, 19 rated by four clinicians, and 1 rated by five clinicians (181 total ratings from 17 clinicians). Ratings retained for the final sample were randomly selected. There were 54 protocols interpreted from clinicians designated as the 1st rater, 54 from a 2nd rater, and 52 from a 3rd rater. All interpretations by each judge were completed independent of all other judges.

The judges were an unscreened sample of working clinicians with some interest in the

Rorschach (i.e., sufficient to subscribe to the e-mail list and volunteer for research). The left side of Table 7 shows they were mature, doctoral-level clinicians with an average of about 15 years of clinical experience. They typically worked in private practice and generally interpreted a Rorschach according to the CS about once per week, with a total of about 300 interpretations over the course of their career.

Clinicians Contributing Judgments For Data Set D

After we recognized the design complications with Data Set C, the 17 participating clinicians were asked if they would be willing to interpret data for all 55 patients. Eight agreed to do so. Thus, the Data Set D clinicians were a subset of those who participated in Data Set C. About one year after the Data Set C interpretations had been collected, the Data Set D clinicians interpreted all 55 protocols, including those they had previously seen. As before, all judges worked independently and blind to other interpretations. Table 7 provides their relevant background information. There were no significant differences in the characteristics of the eight clinicians who participated in Data Set D and the nine who did not.

Data Analysis

The ratings in Data Sets C and D were examined primarily by *r*. For each item or aggregated judgment Data Set C produced three pairwise correlations (i.e., 1st rating with 2nd rating, 1st with 3rd, and 2nd with 3rd) and the average of these three correlations is reported. Data Set D produced 28 pairwise correlations (i.e., Clinician D with Clinician E, D with F, D with G, etc.) and the average of these is reported.

If Data Set C was compromised by its design, it was important to document that the design itself caused poorer reliability. To assess this, sets of three clinicians' interpretations were randomly selected for each patient in Data Set D and then shuffled so that each clinician could serve as the first, second, or third rater for a protocol. This allowed us to transform the design of Data Set D into one that mirrored the design of Data Set C. If the design itself was problematic, reliability for Data Set D should drop following this transformation.

Study 2 Results and Discussion

Results from the adequate design (i.e., Data Set D) are presented first in Table 8. The eight clinicians interpreted the 55 protocols with reasonable reliability across the 29 items and five aggregated judgments; the mean *r*s were .68 and .82, respectively.[3] In general, the reliability of judgments observed in this sample of practicing clinicians was lower than though still fairly similar to the reliability observed for the three clinicians in Study 1 (e.g., which had M *r*s of .79 and .88). As with Study 1, item 57 produced the lowest inter-clinician agreement (*M r* = .36).

Table 8 also presents results from Data Set C. Reliability was substantially lower for this sample than for Data Set D. Importantly, however, reliability was also much lower after sets of three Data Set D clinicians were randomly selected for each protocol and assigned to one of the three rater positions. Because the last column of results in Table 8 was obtained after simply rearranging the data that had been gathered for the first column of results, it appears clear that the design itself produces lower reliability estimates. Accordingly, we believe it is only the first column of data in Table 8 (i.e., the initial Data Set D) that provides an adequate estimate of interpretive reliability among these practicing clinicians.

As with Data Set A, ipsative scores were examined in Data Set D by transforming raw scores into relative judgments indicating the extent to which each clinician believed a patient possessed the 29 characteristics relative to his or her own mean rating. The ipsatized scores again produced virtually identical results for *r*, the consistency ICC, and the agreement ICC. All three statistics produced identical mean reliability coefficients for the item-level interpretations and aggregated judgments (i.e., each statistic had means of .68 and .82, respectively).

Exploring Individual Differences and True Score Theory Across Studies and Data Sets

Clinician Differences in Reliability

To assess the performance of individual clinicians, Table 9 presents clinician-by-clinician reliability from Data Sets A and D. The table indicates how each clinician's interpretations corresponded, on average, to the interpretations of every other clinician in their data set. Clinicians were arranged so that pairwise inter-clinician agreement is highest in the upper left quadrant and lowest in the lower right. The bottom rows of the table provide means for each clinician with every other clinician at the level of item interpretations and aggregated judgments.

Examining the last row of Table 9, it can be seen that the clinicians in both data sets agreed with each other to a reasonable degree when considering the summary judgments. In Data Set A the averages were tightly clustered between .86 and .89. In Data Set D the averages were more variable, but fell in a respectable range from .76 to .86. At the same time, the body of the table illustrates an important pattern. Some clinicians generated more consistent and generalizable interpretations than other clinicians. This phenomenon was particularly noticeable in Data Set D, where clinicians in the far right columns tended to produce more unique or idiosyncratic interpretations relative to every other clinician.

Within Data Set D, the three most reliable clinicians (i.e., D, E, and F) produced agreement rates with each other (*M r* = .83 for items and .90 for aggregated judgments) that were slightly higher than those obtained by the three authors in Data Set A (*M r*s = .79 and .88). At the other end of the spectrum, the three least reliable clinicians in Data Set D (i.e., I, J, and K) produced substantially lower rates of agreement among themselves (*M r*s = .55 and .73). These are quite noticeable differences and the findings indicate that individual clinicians systematically differ in their degree of interpretive consistency with other clinicians.

Restricted variance in ratings could produce systematically lower reliability coefficients. To determine if the least reliable clinicians in Data Set D had restricted variance relative to the most reliable clinicians, the pooled variance for Clinicians D and E was compared to the pooled variance for Clinicians J and K. The variances were similar across all 29 items. Clinicians D and E had an average variance of 1.5, while J and K had an average variance of 1.3. The size of the average difference in variances was small (Cohen's *d* = .12, range -.19 to .30), which indicated that restricted variance was unlikely to account for J and K's lower reliability.

The results in Table 9 address consistency with other raters but not the accuracy of interpretations. It is possible that a clinician may have identified subtle qualities or complex markers in the structural summary data that allowed him or her to formulate unique and clinically accurate inferences, even though doing so produced lower rates of agreement with other clinicians. With the data at hand it is not possible to document what may constitute such valid but unique interpretations. However, the prospect of high accuracy with low reliability cannot exist for more than one clinician. For instance, both Clinician J and K cannot be accurate because they had the lowest rates of agreement with each other. If each were accurate, by definition they would have to strongly agree with each other.

Exploring Classical True Score Theory

Many reliability studies contain information from just two judges, which makes it difficult to compute and use true scores for the objects under consideration. As a result, most reliability studies focus on the association between two individual judges rather than the association between each judge's observed score and the target object's psychometric true score. However, Data Sets A and D provide scores from multiple judges, making it possible to compute approximate true scores for these data. Before doing so, a brief review of classical test theory provides context for the analyses (see Allen & Yen, 1979; Nunnally & Bernstein, 1994).

According to true score theory, every observed score (X) is a function of two components, the true score (T) and random error (RE), such that X = T + RE. In the context of the studies reported here, every clinician's judgment (X) is a function of the patient's true score on the characteristic being evaluated (T) as well as random factors that interfere with the clinician's assigned rating (RE).

According to true score theory, when one repeatedly obtains independent measurements of the same object, random errors of measurement, some of which produce negative deviations from the true score and some of which produce positive deviations, should cancel out. As the number of independent observations increases, the average value for random error decreases until it approaches zero and the RE component drops out of the equation given above. As such, with increasing observations, the mean of all the independent observed scores (Xs) approximates the underlying true score. (With an infinitely large sample, the mean observed score defines the true score.) In the current context then, the mean interpretive rating assigned across clinicians provides an estimate of each patient's true score on the rated item because taking the average reduces the random errors that produce unreliability in each clinician's observed score.

To compute approximate true scores for the patients, clinician ratings were averaged. For each of the 29 item-level constructs, the judgments from the three clinicians in Data Set A were averaged. Separately, the judgments from the eight clinicians in Data Set D were averaged.

There is a potential confound when correlating each clinician's interpretive judgments with the approximate true scores. If judgments from a clinician in Data Set A were correlated with the approximate true scores generated from Data Set A, the relationship between the predictor and the criterion would be artificially inflated because that clinician's interpretations also would have contributed to the approximate true score. To avoid this problem, the interpretations for each clinician in Data Set A were correlated with the approximate true scores obtained independently from the Data Set D judgments. Similarly, the clinician ratings in Data Set D were correlated with the approximate true scores independently derived from the Data Set A interpretations.

Before presenting results, it is important to emphasize what psychometric true scores do and do not indicate. In true score theory, the term "true" means consistent or unwavering. It does not mean accurate, correct, or valid (see Streiner, 2003). True score theory partitions every observed score into just two components, the true score and random error. It is not uncommon to misinterpret these components and think of random error as if it referred to any type of error.

However, this is incorrect. Any form of *systematic* bias or *systematic* error affecting both measurements is an element of the true score, not of random error. Because true scores include the systematic error that may be present in data, true scores are not synonymous with accurate scores or valid scores.[4] Whether a true score accurately indicates the construct a test is designed to measure is a question of validity that cannot be addressed by reliability theory (Allen & Yen, 1978; Nunnally & Bernstein, 1994). Despite the potential for misinterpretation, we use the term true score in what follows because it has a precise meaning in the context of reliability analyses.

With the forgoing in mind, Table 10 presents relevant findings. The first two data columns provide interrater reliability results, while the last two columns present correlations with psychometric true scores. The top section of the table provides data for the clinicians in Data Set A, the middle section for clinicians in Data Set D, and the final row presents data for all of the Data Set A interpretations relative to all of the Data Set D interpretations.

The findings in Table 10 support two general conclusions. First, consistent with psychometric theory, every clinician produces strong and substantially higher levels of agreement with approximate true scores than with the ratings of other individual clinicians. For instance, at the level of aggregated judgments, each clinician's average interrater reliability with another clinician ranges from a low of .76 to a high of .89 (*M* = .83, 2nd data column, top two sections). However, the correlation between each clinician's interpretation of the data and the patients' true scores range from a low of .82 to a high of .94 (*M* = .90, 4th data column, top two sections). These differences occur because interrater reliability coefficients are reduced by the random errors made by both of the clinicians being compared, while correlations with approximate true scores are reduced by just the random errors of one clinician. In general, the data reveal that clinicians do a noticeably better job predicting patient true scores than predicting the interpretive ratings of another clinician.

Second, it can be seen that the clinicians who produced the highest levels of interrater reliability also produced the highest levels of agreement with approximate true scores (the correlation of the first and second data columns with the third and fourth data columns is .91). This finding is also in accord with psychometric theory, which stipulates that the square root of the reliability coefficient should equal the correlation between an observed score and its true score. In Table 10 this formal relationship is not exact, in part because psychometric true scores were estimated from a limited number of observed scores, but mostly because the clinician ratings were not all equally correlated with each other (see Nunnally & Bernstein, 1994).

The last row in Table 10 is also informative. The average interrater reliability between each of the clinicians in Data Set A and each of the clinicians in Data Set D (i.e., .74 for item interpretations, .85 for aggregated judgments) is consistent with the interrater results reported higher up in the columns. However, for the approximate true score correlations, the item-level and aggregated judgments in the final row (i.e., .94 and .97, respectively) are now more similar to each other. They are also noticeably larger than any of the other values in these columns. In fact, at the aggregate level, the associations between the approximate true scores in Data Set A and the approximate true scores in Data Set D are almost perfect, having a mean of .97 (the range was .96 to .98 across the 5 constructs). Thus, even though each individual clinician's judgments were affected by random errors, the near perfect association of their averaged judgments suggests that all clinicians were targeting a single core conceptualization of each patient based on his or her CS structural summary. Indirectly, these findings support the notion that "two heads are better than one." Clinicians are likely to refine and correct their clinical inferences when they consult with a colleague who has independently derived his or her own inferences.

Finally, because individual clinicians vary in their level of agreement with other clinicians and with psychometric true scores, it is worthwhile to consider some of the factors that may lead to higher rates of unreliability. Clinicians who were less consistent (e.g., J, K) may have been a) less conscientious when completing the rating scales, b) more rushed to finish a time-consuming and non-remunerative task, c) more confused by the wording of items on the rating scale, d) more inattentive because they completed the scales in the presence of environmental distractions, or e) more idiosyncratic or inconsistent in the way they conceptualized CS variables.

It is not possible to know if the results found in this research task also characterize the reliability and conscientiousness that these clinicians would display in a real clinical context where there are genuine adverse consequences for faulty inferences. Because the contingencies of research participation are not equivalent to the contingencies of clinical practice (cf., Viglione, 1999), one cannot conclude that Clinicians J and K are likely to be inconsistent and idiosyncratic when they interpret the CS protocol of a patient in their office. At the same time, however, the results from Data Set A are from clinicians who likely were motivated to do their best by virtue of their role as authors and designers of the research. Their results may provide a fairly legitimate upper bound for the level of reliability that would characterize actual clinical practice.

Considering the Current Findings Relative to Meta-Analyses of Interrater Reliability in Psychology, Psychiatry, and Medicine

To consider our results in a relevant comparative context we relied on a recent summary of 42 meta-analytic findings examining interrater reliability in psychology, psychiatry, and medicine (Meyer, 2004). That review obtained results by systematically searching PsycINFO and PubMed for existing meta-analyses, literature reviews that allowed meta-analytic results to be computed, and original articles that allowed new meta-analyses to be computed. To obtain the most generalizable findings from the current CS studies, we merged Data Sets A and D to form a single 11-judge data set of Likert-type interpretive ratings. Q-sort results from Data Set A were also used.

Table 11 provides the comparative results roughly ordered by level of agreement. For each construct, the table shows how many independent pairs of observations produced the findings, as well as the average reliability coefficient. Results were computed separately for scales and items. An item was defined as a single unit of observation or a single judgment, while a scale was derived from the aggregation of item-level judgments. Meyer (2004) found this distinction was an important moderator of reliability; in 17 instances when they could be compared directly, reliability was .77 for scales and .62 for items. Meyer also compared types of statistics, contrasting *r* with kappa or the agreement ICC. Across 16 topics that provided both types of statistics, the average kappa/ICC was .70 and the average *r* was .74. Because these differences are not large, findings for those 16 topics were combined in Table 11 (Meyer, 2004, provides the more differentiated results). Our table differs slightly from Meyer's in one other way. As described more fully below, because we explored the consistency of target stimuli as a moderator of agreement, we now present two different findings from Conway, Jako, and Goodman (1995); one for joint interviews and one for separately conducted interviews.

The meta-analytic results on interrater reliability examine a wide range of phenomena and assess target constructs that vary substantially in their complexity. Although having fuzzy boundaries, constructs included in Table 11 range from scoring tasks that code very discrete or circumscribed events (e.g., entries 4, 5, 10, 11, 13, 14) to interpretive tasks that code abstract or higher level inferences (e.g., entries 6, 19, 26, 31, 41, 44). Consistent with general findings from research on judgment complexity (e.g., Jako & Murphy, 1990), circumscribed judgment tasks requiring relatively few bits of information, such as test scoring, object counts, or physical measurements, tend to be more reliable than complex tasks requiring the synthesis of multiple, higher order inferences, such as the quality of medical care or the merits of research articles. Because we asked clinicians to interpret CS protocols using specific constructs thought to be measured by patterns of CS scores rather than global constructs (e.g., pathology, "neurosis") or constructs with an uncertain link to CS data (e.g., intelligence, conscientiousness), our tasks likely fall at a midrange of complexity; more difficult than test scoring but less difficult than global evaluations of a multifaceted construct-like quality.

Another potentially important dimension embedded in Table 11 concerns the static versus changeable nature of the objects being judged. While most studies had clinicians evaluate the same fixed stimulus (e.g., the same test protocols, medical records, MRI or CT scans, or jointly attended interviews; see entries 1-7, 10-15, 20, 22-24, 26-29, 36-38, 40, 41, 44), a smaller number of studies allowed the target stimulus to vary somewhat across judges. This variability occurs, for example, when judges conduct separate and independent interviews with patients (which also nests a test-retest design within the interrater design; see entries 17, 18, 21, 32-34) or when all judges do not share the same observational or inferential parameters, such as when judges are asked to describe what a person is "generally like" or when they make global ratings of performance (see entries 25, 30, 31, 35, 39, 43). In the latter studies, judges may have in mind different time frames, settings, or exemplar behaviors when making their ratings. Like most of the studies in Table 11, our CS interpretive tasks used a static stimulus; clinicians had the same structural summary for each patient. Comparing the results across both types of designs revealed that static stimuli were associated with notably higher levels of agreement (*M* = .69, SD = .21, *n* = 37) than stimuli that were unbounded or somewhat changeable (*M* = .56, SD = .13, *n* = 25; *t*[60] = 2.73, *d* = .72).

Despite these moderators, Table 11 still illustrates how our results fare relative to other findings in the literature. The data indicate that clinicians interpreting the Rorschach can produce a level of interrater reliability that compares quite favorably with the reliability seen for a wide range of other tasks in psychology, psychiatry, and medicine.

General Conclusions

The reliability of inferences derived from psychological assessment instruments is an important area of research, though it has been neglected in recent years. The Rorschach is like many other psychological and medical tests. It is a complex, multifaceted instrument that needs to be interpreted by trained and skilled clinicians. This article presented a series of data sets examining the consistency with which clinicians interpret the Rorschach Comprehensive System. Across data sets, across a representative set of CS-relevant constructs, and across formats for quantifying clinical interpretations, a diverse array of clinicians reliably interpreted the data for 55 psychiatric patients. Thus, when experienced clinicians were presented with the same Rorschach data, they tended to draw similar conclusions about patients.

As expected, agreement was higher for constructs that summarized the clinician's conceptually related inferences than for single judgments made about specific characteristics. In addition, consistent with psychometric theory, each clinician's interpretations were more strongly associated with patients' approximate true scores than with the ratings of another clinician. This illustrates how accuracy in applied clinical practice would be higher than the level suggested by interrater reliability coefficients.

We also showed how individual clinicians differed in their performance. Some clinicians produced more reliable and generalizable inferences than others. These same clinicians also produced higher associations with psychometric true scores. From the available data, it could not be determined if clinician differences in reliability were the result of some clinicians devoting a higher level of conscientious attention to the methodological requirements imposed by this research task, greater sophistication when thinking through various Rorschach findings, or both.

Unexpectedly, Q-sorts produced slightly lower reliability than a rating scale. In part this may due to the requirement that each patient's characteristics had to fit within a quasi-normal distribution. Allowing the shape of the Q-sort distribution to be dictated by each patient's idiographic characteristics may have produced somewhat higher reliability. However, permitting idiographically determined distributions also would introduce new complications at the point of data analysis. It is also possible that the Q-sorts produced lower agreement than the ratings because they required more complex judgments that were dependent on higher level inferences and cross-characteristic comparisons.

In our initial efforts to recruit clinician volunteers, we made the task circumscribed by asking judges to interpret just 10 or 11 protocols. Doing so meant that a different mix of clinicians provided interpretations for each patient. While this initially seemed desirable, it overlooked how clinicians would adopt different benchmarks for completing our Likert-type rating scale and it inadvertently made the reliability for these clinicians appear low. These design problems were corrected for the final data set.

Overall, the present studies demonstrate that clinicians can reliably interpret Rorschach CS data. This is the first time interpretive reliability has been explored with the CS so these studies add a new dimension to the Rorschach reliability literature (e.g., Acklin et al., 2000; Meyer et al., 2002). Future investigations could expand these findings in several directions. First, scoring reliability was held constant by providing clinicians with already coded protocols. To more closely mirror applied practice, future research could determine reliability after having clinicians both score and interpret the target protocols, or even after independent testing. Second, it would be informative to examine the reliability of Rorschach interpretations relative to the interpretation of other personality tests, such as the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), the Personality Assessment Inventory (PAI; Morey, 1991), and the Millon Clinical Multiaxial

Inventory (MCMI; Millon, 1994). Currently, there are no published studies on the interpretive reliability for these contemporary tests. However, on average, our Rorschach CS interpretive reliability findings seem comparable to the results of similar studies conducted with the original MMPI (e.g., Cooke, 1967; Little & Shneidman, 1959; Poythress & Blaney, 1978; Sines & Silver, 1963).

Finally, in clinical practice, personality assessments are highly idiographic and inferences are generally built upon a wide range of test and extra-test information (Meyer et al., 2001). Although complicated, it would be useful for researchers to embark on efforts to quantify the extent to which independent clinicians derive similar inferences from an open-ended assessment. Not only would this type of research increase our knowledge about interpretive reliability, but it also could go a long way toward the development of models for more ecologically valid studies of the assessment process. We hope the data reported here help to further these goals.

References

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

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/ Cole.

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology, 34*, 164-181.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C. Thomas.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry & Law, 21*, 35-76.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI-2. An administrative and interpretive guide*. Minneapolis: University of Minnesota Press.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271-280.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Cooke, J. K. (1967). Clinician=s decisions as a basis for deriving actuarial formulae. *Journal of*

Clinical Psychology, 23, 232-233.

Elmore, J. G., & Feinstein, A. R. (1992). A bibliography of publications on observer variability (final installment). *Journal of Clincial Epidemiology, 45*, 567-580.

Fagot, R. F. (1991). Reliability of ratings for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement, 15*, 1-12.

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

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception*. San Diego, CA: Academic Press.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311-337.

Koran, L. M. (1975a). The reliability of clinical methods, data and judgments (first of two parts).

New England Journal of Medicine, 293, 642-646.

Koran, L. M. (1975b). The reliability of clinical methods, data and judgments (second of two parts). *New England Journal of Medicine, 293*, 695- 701.

Little, K. B., & Shneidman, E. S. (1959). Congruencies among interpretations of psychological test and anamnestic data. *Psychological Monographs, 73*(6, Whole No. 476), 42.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

Meehl, P. E. (1954/1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Northvale, NJ: Jason Aronson, Inc (reprinted 1996).

Meyer, G. J. (1996). 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). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment, 9*, 480-489.

Meyer, G. J. (1999). Simple procedures to estimate chance agreement and kappa for the interrater reliability of response segments using the Rorschach Comprehensive System. *Journal of Personality Assessment, 72*, 230-255.

Meyer, G. J. (2004). The reliability and validity of the Rorschach and TAT compared to other psychological and medical procedures: An analysis of systematically gathered evidence. In M. Hilsenroth & D. Segal (Eds.), Personality assessment. Volume 2 in M. Hersen (Ed.-in-Chief), *Comprehensive handbook of psychological assessment* (pp. 315-342).

Hoboken, NJ: John Wiley & Sons.

Meyer, G. J., Bates, M., & Gacono, C. (1999). The Rorschach Rating Scale: Item adequacy, scale development, and relations with the Big Five Model of personality. *Journal of Personality Assessment, 73*, 199-244.

Meyer, G. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Hilsenroth, M. J., Piers, C. C., & Resnick, J. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78*, 219-274.

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Mihura, J. L., Meyer, G. J., Bel-Bahar, T., Gunderson, J. (2003). Correspondence among observer ratings of Rorschach, Big Five Model, and DSMIV personality disorder constructs. *Journal of Personality Assessment, 81*, 20-39.

Mihura, J. L., Meyer, G. J., & Wright, A. (2002, March). *Review and meta-analysis of the interpretive reliability of psychological tests*. Paper presented at the annual meeting of the Society for Personality Assessment, San Antonio, TX.

Millon, T. (1994). *MCMI-III manual*. Minneapolis: National Computer Systems.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd edition). New York: McGraw-Hill.

Ozer, D. J. (1993). The Q-sort method and the study of personality development. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development*. Washington, DC: American Psychological Association.

Poythress, N. G., & Blaney, P. H. (1978). The validity of MMPI interpretations based on the Minimult and the FAM. *Journal of Personality Assessment, 42*, 143-147.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901-912.

Sines, L. K., & Silver, R. J. (1963). An index of psychopathology (Ip) derived from clinicians' judgments of MMPI profiles. *Journal of Clinical Psychology, 19*, 324-326.

Spengler, P. M., Strohmer, D. C., Dixon, D. N., & Shivy, V. A. (1995). A scientist-practitioner model of psychological assessment: Implications for training, practice and research. *Counseling Psychologist, 23*, 506-534.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99-103.

Viglione, D. J., Jr. (1999). A Review of Recent Research Addressing the Utility of the Rorschach. *Psychological Assessment, 11*, 251-265.

Viglione, D. J., Jr., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future.

Psychological Assessment, 13, 452-471.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of Rorschach comprehensive system coding. *Journal of Clinical Psychology, 59*, 111-121.

Westen, D., & Shedler, J. (1999). Revising and assessing Axis II, Part I: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry, 156*, 258-272.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59*, 595-613.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Footnotes

[1]. The reviewer who prompted these analyses also hypothesized that reliability would be higher when clinicians were the same gender and when the clinician and patient were matched on gender. Because clinicians did not know the gender of the patient being rated, the latter hypothesis was not tested. However, we examined the first hypothesis with interpretations from Data Set A in Study 1 and Data Set D in Study 2. At the item level, average reliability in Study 1 was .78 for clinicians of the same gender and .79 for clinicians of different genders. At the level of aggregated judgments, the average same-gender coefficient was .86, while the cross-gender coefficient was .88. Although not significantly different, these results were in the direction opposite of the reviewer's hypothesis. In Study 2, the average reliabilities were identical for item interpretations (.68) and aggregated judgments (.82) when the judges were matched or mismatched on gender.

[2]. In fact, the average Pearson correlation, the consistency ICC, and the absolute agreement ICC produced virtually identical results for all the item-level and aggregated judgments in Data Set C.

[3]. The item-level and aggregated judgment means for ICC(C2,1) were .64 and .79, respectively; corresponding values for ICC(A2,1) were .56 and .67.

[4]. While the text presents the traditional psychometric definition of true scores in relation to random and systematic error, it should be noted that these concepts operate somewhat uniquely in the context of interrater reliability research. Within interrater reliability, systematic error is only present when the same errors or biases occur across raters. If just one rater in a pair produces systematic error (e.g., has a consistent, mistaken notion about how to score or interpret a variable), this produces a discrepancy across raters that then reduces the reliability coefficient. In other words, when only one rater in a pair has systematically biased or erroneous ratings, the error is treated as random error for purposes of reliability. This exception to the general rule has sparked some recent debate (Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000).