JOURNAL OF PERSONALITY ASSESSMENT, *78*(2), 219–274

An Examination of Interrater Reliability for Scoring the Rorschach Comprehensive System in Eight Data Sets

Gregory J. Meyer
Department of Psychology
University of Alaska, Anchorage

Mark J. Hilsenroth
Derner Institute of Advanced Psychological Studies
Adelphi University

Dirk Baxter
Department of Psychology
University of Georgia, Athens

John E. Exner, Jr.
Rorschach Workshops
Asheville, North Carolina

J. Christopher Fowler and Craig C. Piers
Austen Riggs Center
Stockbridge, Massachusetts

Justin Resnick
Department of Psychology
Loyola University of Chicago

In this article, we describe interrater reliability for the Comprehensive System (CS; Exner, 1993) in 8 relatively large samples, including (a) students, (b) experienced re-

searchers, (c) clinicians, (d) clinicians and then researchers, (e) a composite clinical sample (i.e., a to d), and 3 samples in which randomly generated erroneous scores were substituted for (f) 10%, (g) 20%, or (h) 30% of the original responses. Across samples, 133 to 143 statistically stable CS scores had excellent reliability, with median intraclass correlations of .85, .96, .97, .95, .93, .95, .89, and .82, respectively. We also demonstrate reliability findings from this study closely match the results derived from a synthesis of prior research, CS summary scores are more reliable than scores assigned to individual responses, small samples are more likely to generate unstable and lower reliability estimates, and Meyer's (1997a) procedures for estimating response segment reliability were accurate. The CS can be scored reliably, but because scoring is the result of coder skills clinicians must conscientiously monitor their accuracy.

Scoring a Rorschach according to the Comprehensive System (CS; Exner, 1993) is a two-step process. First, a sequence of scores is produced for each patient. This grid of data has rows that designate each response given to the 10 inkblots and columns that contain the specific scores that quantify salient response features. Next, a structural summary is generated. As its name implies, the structural summary provides sums that correspond to each of the scores (columns) aggregated across all responses (rows). In addition, the structural summary contains numerous ratios and indexes derived from the combination of these summary scores.

Given the preceding, CS scoring reliability can be evaluated in one of three primary ways. First, one can examine the *patient-level reliability of summary scores.* Here, consistency is evaluated by comparing independently generated summary scores across all patients in a sample. Attention is not given to the scores assigned to each and every response. Second, one can examine the *response-level reliability of specific score options.* Here, each unique score option (e.g., W, D, Dd, S, DQ+, DQo, etc.) is considered separately and consistency is evaluated across all responses in a sample, regardless of which patient provided the responses. Attention is not given to the manner in which scores aggregate to characterize individual patients. A third approach examines the *response-level reliability of multiscore response segments.* Like the second approach, consistency is evaluated across each response in a sample, regardless of which patient produced the response. However, instead of considering each specific score, this approach organizes scores into meaningful segments such that unanimous agreement is evaluated across all location scores, all developmental quality scores, all determinant scores, and so on. This is a more general type of analysis that does not give attention to the distinct score options within a segment.

Meyer (1997a, 1997c) organized interrater reliability information pertaining to the third approach in a meta-analytic review that examined 10 commonly used response segments. Data were obtained from 16 studies published in the *Journal of*

Personality Assessment between 1992 and 1995. For the response segments, reliability coefficients were based on *N*s that ranged from 1,400 to 9,919 responses (*M* = 5,448). Results indicated the CS had excellent chance-corrected reliability, with percentage agreement values that ranged from .87 to .96 (*M* = .92) and estimated kappa values that ranged from .72 to .96 (*M* = .86) across response segments.

Although these results were quite strong, Meyer's (1997a) study had several potential limitations. First, it relied on a formula to estimate kappa from the percentage agreement values available in the published literature. In a comment that accompanied the meta-analysis, Wood, Nezworski, and Stejskal (1997) referred to Meyer's (1997a) statistical procedures as "flawed," "shaky," "dubious," "unsound," and "fatally deficient" (pp. 493–494). Although Meyer (1997c) pointed out how Wood et al. (1997) failed to recognize the mathematics needed to generate correct kappa estimates, it would be useful to directly evaluate his procedures with several independent samples of data.

The meta-analysis also was potentially limited because it focused on multiscore response segments. Although this approach reflects a stringent test of reliability because it requires unanimous agreement for all scores in a segment, it does not explore the reliability for every specific CS score. For instance, the various scores for Color (none, FC, CF, C, Cn), diffuse Shading (none, FY, YF, Y), Human Movement (none, Ma, Mp, Ma–p), and so on, are not considered separately from the broader determinant segment that subsumes all these scores. Meyer (1997a) thus noted how his analysis would not pinpoint potentially problematic scoring rules should they exist for specific scores. Because it is important to evaluate the reliability of all specific CS scores, we report that information in this article.

Meyer's (1997a) meta-analysis also focused on response-level reliability rather than patient-level reliability. Because scores are only assigned to responses, this level of analysis was viewed as appropriate for evaluating the clarity of CS scoring rules. It is also the most stringent means for evaluating coder agreement because it does not permit disagreements to "cancel out" when responses are summarized for patients. For instance, a patient's 20-response Rorschach protocol may have 6 responses in which one of the color determinants (i.e., FC, CF, C, Cn) are assigned. When summed across all responses, Rater A and Rater B may both produce equivalent total scores for the patient, say three FC scores and three CF scores. However, it is possible for the raters to never assign the same score to the same responses. Each time Rater A assigns a CF, Rater B may assign a FC, and vice versa. Reliability coefficients calculated at the level of individual responses would detect this type of disagreement, whereas summary score coefficients calculated at the patient level would not.

Although summary scores present a somewhat more liberal criterion, they are given the greatest emphasis in research and clinical practice because they gener-

ally form the foundation for interpretation and statistical analyses. As a result, summary scores address the applied reliability of CS scoring, and it is essential to understand the precision that accompanies these scores.

As a final issue, Meyer's (1997a) meta-analysis was derived from the published literature. Because researchers may be more conscientious scorers than clinicians working in their day-to-day practice, the meta-analytic results may not characterize CS reliability in an applied setting. Thus, it would be optimal to evaluate reliability using samples that include working clinicians.

In this study, we address each of the preceding issues. First, we report the chance-corrected reliability of CS summary scores in eight diverse samples. In five of the samples, we examined traditional interrater reliability coefficients. In two of these five samples, we examined the scoring of researchers, with one sample comprised of students in training and the other comprised of more experienced researchers. In the third sample, we examined the scoring of a large number of clinicians trained by Rorschach Workshops, the fourth contrasted the scoring completed by clinicians as part of their day-to-day practice to the subsequent scoring of researchers, and the fifth is a composite data set formed by combining the previous four samples. In the remaining three samples, we examined different degrees of experimentally manipulated random scoring error.

Second, we explore the difference between the patient-level reliability of summary scores and the response-level reliability of raw scores. Because aggregated scores should contain less random measurement error (Lord & Novick, 1968; Traub, 1994), we expected summary scores to show consistently larger reliability values.

Third, we explore the impact of sampling error on CS reliability estimates to understand the confounding effects of small samples (Carroll & Faden, 1978). This is accomplished by repeatedly selecting small (n = 20) random subsamples from our larger composite population (N = 219) and contrasting the reliabilities found in the subsamples to the reliability observed in the population.

Fourth, we compare exact kappa values for CS response segments to the statistical estimates generated by Meyer's (1997a) formula to evaluate the generality of his meta-analytic results. Finally, we compare the reliability coefficients derived from our samples to other results obtained from the published literature.

The analyses reported here address numerous issues in a very detailed manner. One could reasonably wonder if such an extensive effort is necessary. We believe it is. Despite the prior meta-analysis, Wood and his colleagues (e.g., Wood & Lilienfeld, 1999; Wood, Nezworski, & Stejskal, 1997; also see Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Lilienfeld, Wood, & Garb, 2000) have continued to claim that CS scoring reliability may be poor or inadequate. Viglione and Hilsenroth (2001) pointed out important problems and inconsistencies in their arguments and reviewed evidence indicating scoring reliability is sound. The detailed findings we present from our data sets should serve to further solidify the evidentiary foundation.

METHOD

Samples

To answer questions about CS reliability with some degree of confidence, we gathered samples that varied along a number of parameters. Our goal was to obtain relatively large samples so there would be a sufficient degree of between-subject variance for each Rorschach score. Ultimately, we collected eight data sets to examine 165 summary scores. The scores we examined are all those that are found on the newest revision of the CS structural summary (Exner, 2001) and those that were on the previous version of the structural summary (Exner, 1995). The eight data sets described here were the only samples we examined for this study and the results were not culled from a larger set of findings. Additionally, within each sample, our analyses employed all usable information from every rater and target protocol. At no time were any data points excluded because of the results they produced.

Sample 1—Student coders. This sample of 66 outpatient protocols was derived from two sources. The first consisted of 23 protocols that had been scored by Gregory Meyer and a bachelor's level research assistant who had no prior assessment experience and was being trained to administer and score the CS. Although Meyer (e.g., 1997b) reported on the reliability for 63 protocols in this data set, most of his archival records only contain summary information about the percentage of agreement observed across response segments. The 23 protocols used here are all those in which two complete sets of independently derived scores were available.

These protocols were obtained when the research assistant was beginning to learn the CS scoring rules. Unlike scoring accuracy, scoring *reliability* is an index of agreement among fallible coders. As such, the least proficient or least experienced scorer in a sample largely determines reliability values.[1] Consequently, even though Gregory Meyer is an experienced rater, because his scoring partner was a student in training it is most appropriate to consider the resulting coefficients as indicating the reliability of a student rater.

The second source of data consisted of 43 outpatient protocols that were recently used as part of another research project (see Meyer, Riethmiller, Brooks, Benoit, & Handler, 2000). Each protocol was independently scored by three graduate students who had completed a semester-long Rorschach course. More than half of these protocols (i.e., 24 of 43) had been used for training purposes to ensure

[1]For instance, even though John Exner, Jr. is recognized as a CS scoring expert, if his scores were compared to those produced by someone with no CS knowledge, reliability coefficients would be low. Superficially, these results would suggest that Exner had poor scoring reliability. However, that would not be the case. The reliability coefficients would be low because of the research design, which paired Exner with an inexperienced criterion rater.

that the raters were scoring properly. Given that these were training protocols, we expected a relatively large frequency of disagreements (particularly because one of the coders had been initially trained in a scoring system other than the CS). One could argue that reliability coefficients should only be derived from coders who are believed to be relatively proficient, and therefore the 24 training protocols should be dropped from this subsample. However, because the first source of student protocols (i.e., the 23 described previously) had been obtained for training purposes, we decided to include all 43 protocols from this second source in our reliability analyses.

Because we were interested in generalizing to student raters across different settings, we wished to combine the second group of 43 protocols with the first group of 23 protocols. To achieve this, we randomly selected two of the three raters for each of the 43 protocols in the second group. Specifically, we placed the 43 protocols in sequential order by code number. For the first 14 protocols, we selected the scoring that was completed by Rater A and by Rater B. For the next 14 protocols, we selected the scoring completed by Rater B and by Rater C. For the final 15 protocols, we selected the coding completed by Rater A and by Rater C.

Sample 2—Experienced coders. This sample consisted of protocols scored by more experienced raters. Four of the authors of this article collected 65 protocols from studies they had conducted. Twenty were scored by Fowler and Piers (see Fowler, Piers, Hilsenroth, Holdwick, & Padawer, 2001), 12 by Fowler and Hilsenroth, 23 by Hilsenroth and an experienced psychometrician with extensive CS administration and scoring experience, and 10 by Meyer and Holaday (see Holaday, 1998). The majority of these protocols were collected from psychiatric inpatients and outpatients, although the last 10 were obtained from children who had experienced severe burns.

Sample 3—Clinician raters. Next, we considered using one of the two reliability samples reported in Exner's (1993) CS text. However, these samples were obtained prior to 1986 and did not incorporate the most up-to-date CS scoring categories. Consequently, we examined data from a study that had been conducted in 1994 with clinicians who had attended training sponsored by Rorschach Workshops. This study was initiated by Rorschach Workshops but never completed because of an error in the data collection procedures. Initially, 300 clinicians were supposed to have been mailed 1 of 10 protocols. Instead, 250 clinicians received 1 of 25 protocols and 50 received 2 of the 25 protocols. Although these data were never analyzed because the design did not proceed as intended, technicians at Rorschach Workshops entered scores for any protocol that had been coded by at least 5 different clinicians. A total of 19 protocols met this criterion: 1 was scored by 7 clinicians, 2 by 6 clinicians, and the remaining 16 were scored by 5 clinicians. For our analysis, in the few instances when 6 or 7 clinicians had scored a protocol, the

scores from 5 were randomly selected. Thus, the final sample consisted of 19 protocols, each of which had been scored independently five times by one of 95 different clinicians. Out of these 95 clinicians, 21 had only attended Rorschach Workshops' basic tutorial, which is designed for clinicians with no prior exposure to the CS. The remaining clinicians had some familiarity with the CS prior to taking a workshop.

The 19 reliability protocols were culled from a sample of outpatients who had been tested at the Payne Whitney Clinic between 1984 and 1988. All patients had been assigned a personality disorder diagnosis, although this was not necessarily the primary diagnosis. Records were selected from the larger pool of patients to ensure none were overly short, long, or constricted. Furthermore, the research assistant who gathered the protocols was instructed to select records only if they were "reasonably complex and difficult to score." In combination, these selection procedures produce more disturbed records than are typically found in an outpatient sample. The selection criteria were designed to serve two goals. First, by excluding short and constricted protocols, the goal was to eliminate those that would be very easy to score and thereby obtain more meaningful reliability data. Second, by excluding very long protocols, it was hoped that raters would not view the task as overly taxing, thereby ensuring a maximal return rate from the volunteer clinicians.

Sample 4—Applied reliability. For this sample, we obtained 69 protocols that were first administered, scored, and used by clinicians as part of their day-to-day work at the Austin Riggs Center, Stockbridge, MA. Later, these protocols were independently scored by Fowler or Piers for research purposes. Because the protocols from this sample were initially used for clinical purposes, this sample allowed us to determine the reliability of CS scoring in a nonresearch, applied setting. The patients in this sample were predominantly young females who were receiving long-term inpatient care for significant psychopathology (most had multiple Axis I and Axis II disorders).

Sample 5—Composite clinical sample. Reliability coefficients are constrained by the degree of variability in the characteristic being rated (see following). For our analysis of summary scores, we examined 165 distinct scores that are part of a CS structural summary. To ensure that each score had roughly the degree of variability that would be encountered by clinicians working in psychiatric settings, the four samples of clinical data (i.e., Samples 1 to 4) were combined to form a single composite sample. For Sample 3, in which five raters scored 19 protocols, we simply selected protocols from two raters chosen at random. The composite sample thus contained 219 protocols that were independently rated by two individuals.

To determine whether these 219 protocols would generalize to a realistic psychiatric setting, they were compared to a sequential series of 440 inpatients and outpatients seen at the University of Chicago Medical Center (UCMC)who had an

electronically stored file of CS scores. Meyer (1997b, 2002) described this sample more fully. Briefly, about half of the patients were inpatients (56%), female (55%), of European American heritage (56%), never married (56%), and diagnosed with a depressive spectrum disorder (52%). The patients had an average age of about 34 and an average of about 13 years of education.

To evaluate whether there were meaningful differences between the UCMC sample and the 219 reliability protocols, we used a fixed standard of effect size magnitude. In a number of investigations with the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943), researchers have suggested that a clinically salient difference is one that exceeds 5 *T*-score points (Greene, 2000; McNulty, Graham, Ben-Porath, & Stein, 1997; Timbrook & Graham, 1994), so we used this as our benchmark. When translated into alternative measures of effect size, this standard produces a Cohen's *d* value of .50 and an *r* value of roughly .25. Thus, our interest was in determining whether any of the 165 CS scores in the composite sample differed from the sequentially collected clinical sample by a *d* value greater than or equal to .50 or an *r* value greater than or equal to .25.

Four analyses were conducted. First, we examined the variance of all 165 scores in the reliability sample relative to the sequential UCMC patients. These analyses were run separately for both raters in the reliability sample. Next, we examined the mean of all 165 scores in the reliability sample relative to the UCMC patient sample. The analyses again were run separately for both sets of reliability scores.

Across 165 variance comparisons with Rater 1, the UCMC sample was less variable than the composite reliability sample on Form Dominated Diffuse Shading (i.e., FY; effect size as $r = .254$, *T*-score equivalent = 5.24). Conversely, the UCMC sample was more variable than the composite reliability sample on the Art content code ($r = .256$, *T* score = 5.29), the Explosion content code ($r = .263$, *T* score = 5.46), and the ALOG Special Score ($r = .254$, *T* score = 5.25). Across 165 variance comparisons with Rater 2, the UCMC sample was more variable than the composite reliability sample on synthesized vague perceptions (i.e., DQv/+; $r = .254$, *T* score = 5.26), the Art content code ($r = .272$, *T* score = 5.64), the Explosion content code ($r = .254$, *T* score = 5.26), and the DV–Level 2 Special Score ($r = .297$, *T* score = 6.22).

Thus, only two of 165 variance differences replicated across both Rater 1 and Rater 2. The Composite reliability sample was too restricted on Art and Explosion content. Theoretically, because variance was restricted in the reliability sample, one could increase the observed reliability values for these scores to more appropriately estimate their true reliability. However, we made no attempt to increase the reliability coefficients for these variables. Importantly, across the 330 analyses, there were no replicated instances when the composite reliability sample had scores with excessively large variance. Consequently, there are no instances in which the composite sample would produce artificially large reliability coefficients.

Across the 165 mean comparisons with Rater 1, the UCMC sample had lower scores than the composite reliability sample on Form Dominated Diffuse Shading(i.e., FY; $r = -.251$, T score = 5.19) and higher scores than the composite on unusual Form Quality (i.e., Xu%; $r = .263$; T score = 5.45). Across 165 mean comparisons with Rater 2, the UCMC sample had a higher mean than the composite sample on just the Art content code ($r = .252$; T score = 5.20). Thus, there were no mean differences that replicated across both raters.

Overall, the results from these analyses clearly suggest that the composite reliability sample approximates the kinds of cases and data one would encounter in applied practice with inpatients and outpatients. Out of 660 tests of variance or mean differences, there were only two replicated differences, both of which suggested the composite sample may artificially produce somewhat lower reliability coefficients. Thus, even though the composite sample combined four diverse primary samples, it accurately represents the type of patients that would be encountered in a medical center and it accurately represents the type of raters that would be encountered in a training environment (i.e., students and more experienced clinicians and researchers).

Samples 6, 7, and 8—Random error samples. These samples used 57 protocols that were drawn from the UCMC patient sample described previously. For each protocol, scoring error was experimentally manipulated by replacing a fixed percent of the original responses with computer-generated erroneous scores. In one sample, 10% of the original responses were replaced with erroneous scores; in another, 20% of the original responses were replaced; and in the final sample, 30% of the original responses were replaced with wrong scores.

We examined these samples for several reasons. First, to our knowledge, no one has experimentally manipulated Rorschach scoring accuracy before. Thus, these analyses provide an initial benchmark for understanding the impact of randomly generated scoring error. Second, raters regularly encounter responses that are difficult to code (e.g., FQu vs. FQo; Form Dominated vs. Form Secondary Color; Shading vs. Achromatic Color). To the extent that certain responses are truly ambiguous, randomly generated scoring decisions mimic the judgment processes that emerge when classifying such responses. Third, a reliability coefficient is designed to indicate the percentage of observed variance in a score that is due to random error (e.g., Allen & Yen, 1979; Nunnally & Bernstein, 1994). However, coding reliability can be lowered both by random error and by systematic error that is not consistent or matched across rater pairs. Random error can emerge from temporary cognitive lapses on the part of one or both raters, transient inattention to the coding task, or from the tentative classifications that are assigned to truly ambiguous responses. Systematic error is quite different. Systematic error can occur when one of the raters misunderstands a coding rule, employs improper benchmarks when applying a coding rule, consistently assigns scores in a liberal or conservative fashion, consistently

neglects certain coding decisions, or all of these. When these types of systematic error affect one coder but not the other, reliability is reduced. Even though both systematic and random error lower CS coding reliability, we believe it is useful to understand the impact of purely random error on the coding process.

To insert randomly generated error into the initial responses, we used the following procedures. First, to determine which responses in a protocol should be altered, a uniform distribution of 10,000 random numbers between 1 and 10 was generated. This distribution was used in a sequential fashion. The first number indicated which Rorschach card should have a response altered. If this card only had one response, that response was changed. If the designated card had more than one response, the list of random numbers was scanned for the next number that fell within the range of responses to that card. This number then indicated which response should be altered. To determine the next response to modify, the table was entered at the point where it had been left and the next number determined the card that should have a response changed. If there was more than one response on this card, the table was consulted again until a number was found that fell within the range of responses to that card. For each protocol, these steps proceeded until 10%, 20%, or 30% of the existing responses had been changed. All altered responses were tracked to ensure they were not modified a second time (i.e., only original responses were changed).

To change each score within a response, a computer program was developed to randomly generate scores in line with their base rates in the full sample.[2] Specifically, a pool of 3,500 score options was generated for each independent score category (i.e., 3,500 options for Location [W, D, Dd], 3,500 options for Space [S, no-S], 3,500 options for Human Movement [Ma, Mp, no-M], etc.). These score options were randomly arranged in a list and used sequentially. Although it would have been easiest to replace an existing response with a complete set of randomly generated scores, we did not replace an original score with an identical random score. Instead, we ensured every score in the altered response had not been assigned in the original and vice versa. To accomplish this we used the following steps. Within a response, the original score option assigned in every category was changed to the next available option in the random pool, as long as the next available score was not the same as the original (e.g., a W would be changed to the next available D or Dd but could not be changed to a W). However, when the original response contained the "no-score" option for a category (e.g., no Color, no Texture, no FABCOM), we did not simply invert the score and assign that which had

[2] *Z* scores were an exception because they are not summarized by their magnitude (i.e., there is no information on the number of times a value of 1.0, 2.0, 2.5, etc. is assigned) and this precludes base rate calculations for each numerical value. Thus, an artificial normal distribution of *Z* scores was created, possessing the same mean and standard deviation as in the full sample. All values in this distribution were rounded to the nearest half point to mimic genuine *Z* scores.

not been present. To do so would have produced very pathological looking responses. Instead, we allowed base rates to determine whether the "score-present" option should be assigned. For instance, the response "Wv CF.YFo Art" has the score-present option for six categories (i.e., Location, Developmental Quality, Color, Diffuse Shading, Form Quality, and Art content) and the no-score option for the remaining 52 categories (i.e., Human Movement, Texture, Pairs, Animal content, Household content, Z Scores, Fabulized Combinations, Aggressive Movement, etc.). To alter the assigned W score, the first non-W location from the random list (i.e., either D or Dd) was inserted. To alter the remaining score-present options, a randomly generated alternative was inserted from the list (i.e., either +, o, or v/+ for Developmental Quality; FC, C, or no-C, for the Color category; FY, Y, or no-Y from the diffuse Shading category; +, u, or – from the Form Quality category; and no-Art from the Art Content category). For the categories with an initial no-score option, simple base rates (i.e., the randomly generated list) determined whether a score-present or no-score option was assigned to the new response. For instance, base rates determined whether Ma, Mp, or no-M would be the appropriate score for the Human Movement category.

Once a score had been used from the list of 3,500 random scores, it was crossed off. Within a category, the pool of scores was also exhausted. Thus, if a randomly generated W had been skipped during the search for the next available D or Dd, it would be inserted in the next possible instance (i.e., when the next original D or Dd response was changed). This step ensured the newly assigned scores would have roughly the same base rates as the existing scores.

For determinant scores, additional rules were developed to simplify the process. A proportion of responses containing determinants other than Pure Form (F) were automatically changed to F responses. The remaining responses with non-F determinants were altered in the manner described previously. Finally, any F response was changed to a randomly generated determinant or determinant blend, with blends determined randomly by the base rates for every determinant option.

Reliability coefficients for these samples were calculated by comparing the original (i.e., correct) scores to the scores derived from the protocols in which 10%, 20%, or 30% of the responses had been changed. Thus, the original scoring was treated as Rater 1 and the revised scoring with random error was treated as Rater 2.

Data Analysis

Statistics. To assess the reliability of response-level scores, Cohen's (1960) kappa was used. To assess the patient-level reliability of summary scores, we calculated intraclass correlation coefficients (ICC) using a one-way random effects model (Shrout & Fleiss, 1979, Model 1; also see McGraw & Wong, 1996). The ICC is a chance-corrected reliability coefficient suitable for continuous data and equivalent to kappa under appropriate conditions (Fleiss & Cohen, 1973; Shrout, Spitzer,

& Fleiss, 1987). According to Shrout and Fleiss (1979), a one-way random effects model treats each protocol as if it were scored by a different set of raters who are randomly selected from a larger population of raters. Under this model, the effects due to raters, to the interaction of raters and protocols, and to random error cannot be separated. As McGraw and Wong (1996) stated in more simplified terminology, for any given protocol, the designation of who is considered Rater 1 and who is considered Rater 2 (or Rater 3, etc.) is random.

Statistical assumptions: Raters. When an earlier version of this manuscript was reviewed, questions were raised about the Model 1 ICC assumptions in relation to our data sets. We thus present the issues in detail, beginning with assumptions about the assignment of raters. At issue is whether the Model 1 ICC requires each protocol to be scored by a separate and unique pair of raters. If so, then a study examining 25 protocols, each of which are scored twice, would require a total of 50 different people to act as raters to ensure that 2 unique and distinct raters scored each protocol. Although this was the case in Sample 3, some rater pairs scored more than one protocol in Samples 1, 2, 4, and 5. As such, a separate and unique pair of raters was not used to score each protocol. By not having a distinct and unique pair of raters for each protocol in the data set, the residual effects (i.e., effects that go beyond the effects of the protocol under consideration and the population mean across all observations) may not be completely independent because an effect of repeated rater pairs may be embedded within the residuals.

However, statisticians have argued that the Model 1 ICC does not require unique and distinct rater pairs for each protocol. Cicchetti (1991) stated that the Model 1 ICC is "the statistic of choice" (p. 120) for designs like those found in our Samples 1, 2, 4, and 5 (also see Cicchetti & Prusoff, 1983). The statisticians at SPSS consider the Model 1 ICC to be the appropriate model for these samples as well. The Model 1 assumptions are described as follows: "Raters are a random sample from a specified population of raters, and each rater does not rate all subjects/objects. Therefore, each subject/object is rated by *a potentially different set of raters*" (italics added; see Nichols, n.d., para. 3). This principle is spelled out even more specifically and explicitly by Nichols (1998).[3] Andreasen et al. (1981;

[3]Nichols (1998) stated the following, where *N* is the number of objects being rated (e.g., protocols), *j* is the total number of raters who contributed to the data base, and *k* is the number of ratings available for each object:

> Suppose the k ratings for each of the N persons have been produced by a subset of j > k raters, so there is no way to associate each of the k variables with a particular rater. In this situation, the one-way random effects model is used, with each person representing a level of the random person factor. Then, there is no way to disentangle variability due to specific raters, interactions of raters with persons, and measurement error. All of these potential sources of variability are combined within person variability, which is effectively treated as error. (para. 4)

2nd reliability design) relied on the same rationale when confronted with a sample that had a mix of different raters but not a distinct pair of raters for every participant.

Furthermore, because of the documented equivalence between the Model 1 ICC and kappa (Fleiss, Nee, & Landis, 1979), the methodological literature on kappa is relevant as well. Fleiss (1971) explicitly generalized kappa to the situation "where each of a sample of subjects is rated on a nominal scale by the same number of raters, but where the raters rating one subject are not necessarily the same as those rating another" (p. 378). Fleiss et al. reaffirmed the appropriateness of computing kappa under these circumstances and further clarified that kappa can be computed in situations when different participants are rated by "different sets of equal numbers of raters" (p. 974). Given that Fleiss et al.'s article also documented the equivalence of kappa and the Model 1 ICC and was published in the same year as Shrout and Fleiss's (1979) ICC article, it seems clear that Fleis would believe it is appropriate to use the Model 1 ICC in our data sets. In practice, it is also the case that many prominent, large-scale, grant-funded studies have computed kappa from a design like that found in our Samples 1, 2, 4, and 5. These are studies in which two raters (R1 and R2) evaluate some proportion of the cases, two different raters (R3 and R4) evaluate another proportion of the cases, and different pairs of raters (e.g., R5 and R6) or shuffled pairs of raters (e.g., R1 and R4, R2 and R3) evaluate additional portions of the sample. The following citations are for field trials with the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., American Psychiatric Association, 1994) or the International Classification of Diseases (ICD–10; World Health Organization, 1992) that employed this type of statistical design: Buysse et al. (1994), First et al. (1995), Keller et al. (1995), Loranger et al. (1994), Volkmar et al. (1994), and Williams et al. (1992).

If the preceding still leaves doubt about the limiting assumptions of the Model 1 ICC, three other considerations also support our use of this model. First, Shrout and Fleiss (1979) indicated how for a given set of data, the one-way random effects model generally provides more conservative (i.e., lower) reliability estimates than the alternative two-way random effects model or two-way mixed effects model. Second, the alternative ICC models contain assumptions that clearly would have been violated in our Samples 1, 2, 4, and 5. Model 2, the two-way random effects model, assumes that the same fixed coders have rated every protocol. Model 3, a two-way mixed effects model, further assumes the fixed number of coders in the study are the only coders of interest (i.e., there is no desire to generalize the reli-

---

From this it can be seen that the Model 1 ICC is appropriate whenever the total number of raters (*j*) exceeds the total number of ratings (*k*) available for each CS protocol (*N*), as long as it is not possible to associate a given rating designation (i.e., Rater A, Rater B) with a specific rater. As an illustration, our student sample has a total of five raters (*j*) who coded 66 protocols (*N*). Each protocol was scored by two (*k*) of the five raters. Because raters were randomly designated as Rater A and Rater B, there is no way to link specific raters to the designated categories.

ability findings to other potential raters). Neither of these assumptions fit for any of our data sets.

Finally, from a practical perspective, when evaluating this issue empirically with our data it makes no difference if we used the Model 1 or Model 2 ICC. In the composite sample of 219 protocols, across the 164 CS scores that were assigned by at least one rater, the average difference between the Model 1 ICC and the Model 2 ICC was –.0000496510 (with Model 2 results being slightly larger, as predicted by Shrout & Fleiss, 1979) and the maximum difference was –.0050982338. Because we report reliability coefficients to two decimal places, it would not make a difference which ICC model results we reported. However, the results reported in the following are all derived from the Model 1 ICC.[4]

Statistical assumptions: Normality and CIs. Another common assumption of the one-way random effects model (and all alternative models) is that the rated characteristic is normally distributed across participants (McGraw & Wong, 1996; Shrout & Fleiss, 1979). This assumption presents a potential problem for Rorschach scores because many variables have skewed and kurtotic distributions. However, Lord and Novick (1968, pp. 162–166) pointed out how the key assumption for the ICC is that each rater is targeting an equivalent score distribution, regardless of its shape. Thus, the critical assumption is that every replication (e.g., the scores generated by Rater 1 and Rater 2) will have the same population parameters. As such, Lord and Novick indicated that the ICC calculations themselves do not depend on assumptions of normality. However, these assumptions become critical if one wishes to determine the statistical significance of an observed ICC value or if one wishes to generate a confidence interval (CI) about an observed value. Our goal is to use the ICC data descriptively rather than inferentially. As a result, our emphasis is on the magnitude of the reliability coefficients.

Nonetheless, a reviewer requested that we present CIs for our coefficients. We do so for the composite sample. These CIs should be considered tentative estimates because the underlying distribution for many scores is skewed, kurtotic, or both. Subsequently, we also demonstrate the inaccuracy of these CIs.

Contending with the impact of restricted variance. When interpreting an ICC coefficient, it is essential to consider score variance across all the participants in the study. This is because it is impossible to demonstrate reliability if the

[4]As noted in the text, we discussed statistical assumptions in detail because of questions raised during the editorial review process. One argument was that we should discard the results from several of our samples. Given the literature cited in the text, this is a view that we strongly dispute. Another argument was that we produced artificially large reliability coefficients by using the "wrong" ICC model. Clearly, however, because there was an average difference of just –.00005 between the results of the Model 1 and Model 2 ICC, this argument was mistaken.

characteristic being rated does not differ from participant to participant (Finn, 1970; Jones, Johnson, Butler, & Main, 1983; Lahey, Downey, & Saal, 1983; Selvage, 1976; Whitehurst, 1984). Thus, in a sample of 10 patients, if it just so happens that all 10 patients have the same or nearly the same value for some CS score, such as M (Human Movement) or P (Popular), it would be impossible or virtually impossible to demonstrate that scoring was reliable in this sample—even if the coders scored accurately and with considerable agreement. This constraint with the ICC is analogous to the constraint encountered when calculating kappa coefficients on variables with extreme base rates (see Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Meyer, 1997a).

As with kappa, several authors have proposed solutions to correct ICCs for restricted between-subjects variance. One factor that could cause restricted variance is a skewed distribution (Whitehurst, 1984). Many Rorschach variables are inherently skewed because most patients receive values of 0 or 1, whereas a limited number obtain values of 3, 4, 5, or higher. Such distributions have restricted variances relative to distributions in which the scores are normally or randomly distributed across the full range of possible scores (Finn, 1970, 1972; Selvage, 1976; Whitehurst, 1984). If it is reasonable to assume that scores can be randomly distributed across the full range of values (i.e., if *chance* can be defined as what happens when raters blindly assign scores across the range of possible values, without regard to relative frequency), then an alternative statistic, Finn's r (r_F), can be used to generate chance-corrected reliability. Although the merits of r_F have been debated (see Cicchetti, 1985; Whitehurst, 1985), it should be noted that r_F corrects ICC coefficients in the same way that kappa$_n$ (Brennan & Prediger, 1981) corrects Cohen's kappa for problems induced by extreme base rates (Cicchetti, 1985; Meyer, 1997a).

Lahey et al. (1983) proposed an alternative solution, recommending that ICCs should not be calculated unless scores have statistically significant variance across targets. Although this is reasonable in some respects, statistical significance depends on power, so even a small degree of between-subject variance will be significant if there are many target protocols. Another problem is that scores are less likely to be statistically significant across participants when raters are sloppy, yet more likely to be significant when raters are accurate. This is because significance is determined by the F ratio comparing between-subject variance to within-subject variance. Assuming the same small degree of between-subject variance (MSB), more reliable coders will produce a smaller degree of within-subject variance (MSW) and a larger F value than sloppy coders who will produce a larger MSW and a smaller F value. Thus, the same small degree of between-subject variance is more likely to be statistically significant for good raters than poor raters. If one used significance to determine when an ICC should be calculated, good coders would be held to a more demanding standard than poor coders because they would have to demonstrate reliability on a smaller degree of genuine construct variance (i.e., a smaller MSB). This penalizes good raters who correctly identify samples

with minimal between-subject variance in the same way that kappa penalizes good raters who independently determine that a score occurs infrequently (Meyer, 1997a; Zwick, 1988).

Given the preceding considerations, we believe it is better to avoid a relativistic criterion like the statistical significance approach of Lahey et al. (1983) to determine when ICC calculations are appropriate. We also believe it is better to avoid theoretical assumptions about how scores could be normally or randomly distributed when they are observed to have restricted variance or skewed distributions (e.g., as with Finn's *r*). Instead, to determine when ICC coefficients are likely to be statistically stable and appropriate to calculate, we relied on fixed guidelines akin to those proposed for kappa. Grove et al. (1981) indicated scores with a base rate of less than .05 should be considered statistically unstable and discouraged researchers from calculating chance corrected reliability statistics under these circumstances. In this study, we used a more demanding criterion and set a base rate of .01 as our cutoff. Scores that occurred less than 1 time in 100 responses were considered statistically unstable.

For most of the following analyses, we only present summary data (i.e., mean and median) for statistically unstable variables. Nonetheless, because part of our goal was to present a complete review of the CS, we present all ICC values for the composite sample, regardless of score base rates.

Using a fixed base rate to determine when it is appropriate to compute ICC values protects against most instances of restricted between-subject variance. However, restricted variance can still occur when scores coincidentally happen to be relatively similar across participants (e.g., when, by chance alone, all participants in a sample happen to have similar M scores). Of course, this is more likely in small samples than in large samples. To protect against this problem in our analyses, we tried to obtain a relatively large *N* for each sample. However, Sample 3 contains only 19 patients. As a result, when the 165 structural summary variables are examined across these 19 patients, it is likely some variables will coincidentally have a restricted degree of between-subject variance, even if the score has a base rate greater than .01.

General points. In a reliability study, raters make use a written record of responses. Because the written record generally makes it clear as to what constitutes a response, there should be relatively little ambiguity about two CS variables, R and Afr. Both of these variables are defined by the number of responses in a protocol. Nonetheless, raters can and do disagree about what constitutes a response. Therefore, even though reliability for these two variables may be somewhat high by definition, they were included in our analyses.

For all analyses, *N* refers to the number of target objects rated (i.e., either the number of patient protocols or the number of responses contained in these protocols) and *k* refers to the number of ratings obtained for each object. Traditionally, ICC and kappa (κ) values are interpreted as follows: values greater than .74 are

considered to indicate *excellent* reliability, values from .60 to .74 are considered *good*, values from .40 to .59 are considered *fair*, and values below .40 are considered *poor* (Cicchetti, 1994; Cicchetti & Sparrow, 1981; Fleiss, 1981).[5]

RESULTS

Summary Score Reliability for the Genuine Clinical Samples

Table 1 reports summary ICC information for all eight samples.[6] Sample 1, the student sample, contained 133 statistically stable scores, 27 unstable scores, and 5 scores that were never assigned. The stable scores had a median reliability of .85 (*M* = .82) and ranged from a low of .34 to a high of 1.0. Using ICC interpretive guidelines, the relatively inexperienced coders in this sample displayed poor reliability for 1 score (i.e., C'F), fair reliability for 4 scores (FT, FV, FC, S – %), good reliability for 22 scores (i.e., Wv, MQu, SQu, WDu, CF, C, SumV, YF, Hd, (Hd), An, Ay, Hh, Ls, Idio, DV1, INC1, FAB1, ALOG, An + Xy, Xu%, SCZI), and excellent reliability across the remaining 106 variables (79.70%). The 25 statistically unstable scores had a median ICC of .49 (*M* = .51). Attesting to the fragility of these statistically unstable coefficients, six of the coefficients came from instances in which one rater never assigned the score but the other rater assigned the score one time across the 1,407 responses. In each of these instances, the ICC was zero. For three of the six variables (MQ+, FQx+, WD+), the disagreement occurred because a single response was assigned FQx+ by one rater, whereas the other rater assigned FQxo. If this single FQx+ score is disregarded, the median and mean ICC become .59 and .57, respectively. If one only considers scores assigned at least one time across the 1,407 responses by both raters, the median and mean ICC for the unstable scores in this sample become .69 and .66, respectively.

Sample 2, the experienced coder sample, contained 140 stable scores, 20 unstable scores, and 5 scores that were never assigned. The stable scores had a median ICC of .96 (*M* = .95) and ranged from a low of .64 to a high of 1.0. Thus, the coders in this sample scored virtually all variables with excellent accuracy. The only variable that fell within the good classification was the score for vague Perceptions combined with poor Form Quality (i.e., DQv with FQ–). The remaining 139 scores (99.29%) had excellent reliability. In this sample, the 20 statistically unstable scores had a median ICC of .88 (*M* = .83).

---

[5]Landis and Koch (1977) proposed an alternative convention for interpreting κ and ICC values that is more frequently encountered in the medical literature. They suggested the following guidelines: less than .00 = *poor,* .00 to .20 = *slight,* .21 to .40 = *fair,* .41 to .60 = *moderate,* .61 to .80 = *substantial,* and .81 to 1.00 = *almost perfect.*

[6]A table of specific values for the stable scores in each sample is available from Gregory Meyer.

TABLE 1
The Reliability of 165 Comprehensive System Structural Summary Scores: ICCs for Four Clinical Samples, a Composite Clinical Sample, and Three Experimentally Altered Samples Containing 10%, 20%, and 30% Random Scoring Error

				Statistically Stable Scores (Base Rate ≥ .01)							*Statistically Unstable Scores*			
							% in ICC Classification							
Samples	*N*	*M R*	*k*	*No. of Scores*	*Mdn ICC*	*M ICC*	*Poor < .40*	*Fair .40–.59*	*Good .60–.74*	*Excellent ≥ .75*	*No. of Scores*	*Mdn ICC*	*M ICC*	*No. of Scores Never Assigned*
Clinical samples														
1. Student	66	1,407	2	133	.85	.82	.008	.030	.165	.797	27	.49[a]	.51[a]	5
2. Experienced	65	1,299	2	140	.96	.95	.000	.000	.007	.993	20	.88	.83	5
3. Clinicians	19	388	5	135	.97	.91	.015	.007	.096	.881	15	.59	.66	15
4. Applied	69	1,667	2	139	.95	.94	.000	.000	.022	.978	24	.89	.88	2
5. Composite	219	4,761	2	138	.93	.91	.000	.000	.029	.971	26	.83	.83	1
Forced random error samples														
6. 10% random	57	1,378	2	143	.95	.94	.000	.000	.007	.993	16	1.00	.89	0
7. 20% random	57	1,378	2	142	.89	.88	.000	.000	.042	.958	17	.87	.76	0
8. 30% random	57	1,378	2	141	.82	.80	.007	.050	.206	.738	18	.84	.71	0

Note. ICC = intraclass correlation coefficients; *N* = number of target patients who provided Rorschach protocols; *M* R = mean number of responses contained in the protocols (averaged across raters); *k* = number of raters who coded each protocol.

[a]For six variables, one rater never assigned a score, whereas the other rater assigned it one time across the 1,407 responses. In each of these instances, the ICC was zero.

Sample 3, the clinician sample in which 19 protocols were scored five separate times by a total of 95 different clinicians, had 135 statistically stable scores, 15 unstable scores, and 15 scores that were never assigned. The 15 unstable scores had a median ICC of .59 (M = .66). The stable scores had a median ICC of .97 (M = .91), indicating overall reliability almost identical to that seen in the experienced rater sample. However, in Sample 3, the ICC values for the stable scores had a much larger range, extending from a low of .35 to a high of 1.0. Two values fell in the poor range (i.e., FY and DR2), 1 was fair (i.e., Level 2 Special Scores), 13 were good (i.e., SQu, CF, CF + C, FT, FV, SumY, Idio, DV1, DR1, FAB2, AB, PTI, SCZI), and the remaining 119 scores (88.15%) were in the excellent classification range.

The results from Sample 3 suggest two possibilities. First, it may be the raters were generally as proficient as those in the other samples (reflected in the very similar median and mean ICC values), but stumbled more often when considering some particular scores (reflected in several low ICC values). Indeed, the data for this sample were collected after the scoring rules were altered for Diffuse Shading (Y scores), Achromatic Color (C' scores), and all of the cognitive Special Scores (in which the Level 1 vs. Level 2 distinction was added). Thus, relatively new rules may have been partially responsible for the lower coefficients observed with a number of variables (i.e., FY, SumY, DV1, DR1, DR2, FAB2, Level 2 Special Scores, PTI, SCZI).

The second possibility is that restricted variance in this sample may have made it more difficult to demonstrate reliability for some variables. The scores FY, DR2, SQu, AB, and Id all had less than excellent ICC values. They also had five of the seven smallest MSB values, indicating there was little between-subject variance on these characteristics. In addition, had we followed Grove et al. (1981) and classified all variables with a base rate less than .05 as statistically unstable, then 9 of the 16 scores that did not have excellent reliability values would never have been calculated (i.e., SQu, FY, FT, FV, Id, DV1, DR2, FAB2, AB). More broadly, 15 scores were never assigned by any of the coders in this sample, which is 3 to 15 times the rate of unassigned scores in the remaining samples. This reflects the small number of target protocols in Sample 3 and is another indication that between-subject variability was constrained. Taken together, these factors suggest statistical confounds probably played a role in lowering reliability coefficients for some variables in this sample. We return to this point later.

Sample 4, the applied reliability sample in which 69 protocols were initially scored by clinicians as part of their regular clinical practice and then rescored by researchers, contained 139 stable scores, 24 unstable scores, and 2 scores that were never assigned. The stable scores had a median of .95 (M = .94) and ranged from a low of .64 to a high of 1.0. Three variables had good reliability (INC1, ALOG, COP), whereas the remaining 136 (97.84%) had excellent reliability. Thus, in this sample scores that were initially assigned as part of daily clinical practice corresponded

quite closely to scores that were assigned as part of a formal research investigation. The 24 unstable scores in this sample had a median of .89 (M = .88), suggesting the clinicians and researchers also scored these variables reliably.

The composite sample of 219 protocols had 138 statistically stable scores, 26 unstable scores, and one score (FQf+) that was never assigned. The stable coefficients had a median value of .93 (M = .91) and ranged from a low of .62 to a high of 1.0. Four variables had good reliability (C'F, FV, INC1, ALOG), whereas the remaining 134 stable scores (97.10%) had excellent reliability. The unstable variables had a median of .83 (M = .83). One of these scores (TF) would be classified as having fair reliability, 4 would be classified as good (MQnone, SQ+, C', (Ad)), and the remaining 21 would be classified as having excellent reliability.

Table 2 uses the format of a structural summary to present reliability coefficients for every CS score in the composite sample (N = 219). Coefficients are presented for all variables regardless of base rate. However, the statistically unstable scores are identified and their coefficients should be interpreted cautiously.

To generate CIs for variables in the composite sample, we utilized the formulas developed by Shrout and Fleiss (1979) as operationalized in SPSS. The CIs can be perfectly predicted from an observed ICC value using a nonlinear regression equation. For our sample of 219 patients, the following equations have R^2 values of 1.0 for predicting the observed lower bound CI value, the upper bound CI value, and the CI range:

$$\text{Predicted lower 95\% CI value} = -.120247 + (.938253 \times \text{ICC value}) + (.181915 \times (\text{ICC value})^2)$$

$$\text{Predicted upper 95\% CI value} = .138428 + (.957575 \times \text{ICC value}) - (.096036 \times (\text{ICC value})^2)$$

$$\text{Predicted 95\% CI range} = .258688 + (.019308 \times \text{ICC value}) - (.277951 \times (\text{ICC value})^2)$$

Because one can compute specific CIs for each score listed in Table 2 using these equations, we do not report intervals for each score. Instead, the left side of Table 3 reports CIs for some benchmark coefficients.

Summary Score Reliability for the Samples Containing Fixed Proportions of Random Error

Table 1 reports reliability for the three samples containing a fixed proportion of randomly generated error. Sample 6, in which 10% of the responses had been changed to have randomly generated erroneous scores, contained 143 stable scores and 22

TABLE 2

The Reliability of a Comprehensive System Structural Summary: Intraclass Correlation Coefficients for the Composite Clinical Sample Containing 219 Rorschach Protocols

LOCATION FEATURES		DETERMINANTS BLENDS	SINGLE		CONTENTS		SPECIAL SCORES COGNITIVE SPECIAL SCORES	Lv1	Lv2
Zf	= .98		M	= .96	(H)	= .93			
ZSum	= .98	Color Shading = .87	FM	= .95	Hd	= .92			
ZEst	= n/c		m	= .92	(Hd)	= .88	DV =	.84	.80[a]
		All Blends = .93	FC	= .83	Hx	= .96	INC =	.74	.82
W	= .99		CF	= .84	A	= .99	DR =	.82	.85
(Wv	= .92)		C	= .92	(A)	= .94	FAB =	.90	.81
D	= .99		Cn	= .97[a]	Ad	= .95	ALOG =	.69	
Dd	= .98		FC’	= .86	(Ad)	= .72[a]	CON =	.76[a]	
S	= .94		C’F	= .62	An	= .92			
			C'	= .71[a]	Art	= .91	Raw Sum6 =		.91
DQ			FT	= .78	Ay	= .84	Wgtd Sum6 =		.90
		(FQ–)	TF	= .54[a]	Bl	= .98			
+	= .98	(.96)	T	= .97[a]	Bt	= .96	OTHER SPECIAL SCORES		
o	= .99	(.96)	FV	= .72	Cg	= .95			
v/+	= .81[a]	(.87[a])	VF	= .81[a]	Cl	= .87[a]			
v	= .91	(.75[a])	V	= 1.0[a]	Ex	= .86[a]	AB =	.93	
			FY	= .90	Fd	= .94	AG =	.90	
			YF	= .82	Fi	= .93	CFB =	.96[a]	
			Y	= .79[a]	Ge	= .94[a]	COP =	.86	
			Fr	= .98	Hh	= .85	CP =	.79[a]	
			rF	= .85[a]	Ls	= .94	GHR =	.89	
			FD	= .88	Na	= .94	PHR =	.94	
			F	= .97	Sc	= .93	MOR =	.89	
					Sx	= .96	PER =	.94	
					Xy	= .95[a]	PSV =	.84	
			(2)	= .97	Id	= .92			

FORM QUALITY

	FQx	FQf	MQ	SQ	WD
+ =	.91[a]	n/a	.90[a]	.66[a]	.91[a]
o =	.98	.96	.94	.88	.97
u =	.93	.91	.82	.84	.93
– =	.96	.92	.92	.91	.97
None =	.93	—	.66[a]	.82[a]	.93

RATIOS, PERCENTAGES, AND DERIVATIONS

R = 1.0	L = .98	(PureF% = .95)			FC:CF + C	= .83:.89		COP = .86	AG = .90	
EB = .96:.94	EA = .96				Pure C	= .92		GHR:PHR	= .89:.94	
eb = .95:.94	es = .95	EBPer = .92			SmC':WSm C	= .90:.94		a:p	= .96:.92	
	Adj es = .94	D = .89			Afr	= .99[c]		Food	= .94	
		AdjD = .87			S	= .94		Sum T	= .89	
FM = .95	Sum C' = .90	Sum T = .89			Blends/R	= .93		Human Cont	= .98	
m = .92	Sum V = .81	Sum Y = .91			CP	= .79[a]		(HHd):(AAd)	= .95:.92	
			XA%	= .92				H+A:Hd+Ad	= .99:.95	
			WDA%	= .94	Zf	= .98		PER = .94	Isolate/R = .96	
	Sum6	= .91	X – %	= .93	W:D:Dd	= .99:.99:.98		3r + (2)/R	= .96	
	Lv2	= .87	S – %	= .75	W:M	= .99:.96		Fr + rF	= .98	
a:p	= .96:.92	WSum6	= .90	P	= .93	Zd	= .93	SumV	= .81	
Ma:Mp	= .96:.91	M –	= .92	X + %	= .96	PSV	= .84	FD	= .88	
~~2AB+Art+Ay~~	= .93	MNone	= .66[a]	F + %	= .93	DQ+	= .98	An + Xy	= .92	
MOR	= .89			Xu%	= .83	DQv	= .91	MOR	= .89	
								H:(H)Hd(Hd)	= .98:.96	
PTI = .88	SCZI = .87	DEPI = .84	CDI = .87		S–CON = .83		HVI = .91	OBS = .91		

Note. n/a = variable was never assigned in this sample; n/c = ZEst was not calculated because it is not an independent score and it has the same reliability as Zf.

[a]Raters assigned these variables at a base rate of less than .01, making them potentially unstable statistics.

TABLE 3
95% CIs for Selected Benchmark Coefficients in the Composite Sample[a] and in Small Samples[b]

	Composite Sample			*Small Samples*		
Observed ICC Value	*95% CI Range*	*Lower Boundary*	*Upper Boundary*	*95% CI Range*	*Lower Boundary*	*Upper Boundary*
.99	.0054	.9869	.9923	.0213	.9748	.9960
.97	.0159	.9610	.9769	.0608	.9271	.9879
.95	.0262	.9353	.9615	.0992	.8806	.9798
.93	.0362	.9097	.9459	.1362	.8352	.9715
.91	.0461	.8842	.9303	.1721	.7910	.9631
.89	.0557	.8589	.9146	.2067	.7479	.9546
.85	.0743	.8087	.8830	.2724	.6650	.9374
.80	.0962	.7468	.8430	.3482	.5670	.9152
.75	.1168	.6858	.8026	.4171	.4752	.8923
.70	.1360	.6257	.7617	.4796	.3891	.8687
.65	.1538	.5665	.7203	.5359	.3084	.8442
.60	.1702	.5082	.6784	.5863	.2326	.8189
.50	.1989	.3944	.5932	.6710	.0945	.7655
.40	.2219	.2842	.5061	.7362	–.0282	.7080
.30	.2395	.1776	.4171	.7845	–.1384	.6461
.20	.2514	.0747	.3261	.8186	–.2392	.5794
.10	.2578	–.0246	.2332	.8410	–.3336	.5074
.00	.2587	–.1202	.1384	.8544	–.4245	.4299

Note. CI = confidence interval; ICC = intraclass correlation coefficient.
[a]$N = 219$. [b]$N = 20$.

unstable scores. The statistically unstable scores had a median of 1.0 ($M = .90$). The statistically stable scores had a median of .95 ($M = .94$) and ranged from a low of .68 to a high of 1.0. One variable (TF) had reliability values in the good classification range, whereas the remaining 142 (99.30%) fell in the excellent range.

Sample 7, in which 20% of the responses had been changed to random error, contained 142 stable scores and 23 unstable scores. The statistically unstable scores had a median of .84 ($M = .78$). The statistically stable scores had a median of .89 ($M = .88$) and ranged from a low of .63 to a high of 1.0. Six variables had reliability values in the good classification range (TF, Idio, P, AB, Lambda, OBS) and the remaining 136 (95.77%) fell in the excellent range.

Sample 8, in which 30% of the responses were randomly generated error, contained 141 stable scores and 24 unstable scores. The statistically unstable scores had a median of .84 ($M = .71$). The statistically stable scores had a median of .82 ($M = .80$) and ranged from a low of .39 to a high of 1.0. One variable (Lambda) had reliability in the poor classification range, 7 variables had reliability in the fair range (SQo, Fr, Idio, Popular, DV1, Xu%, OBS), 29 variables had reliability in the

good range (DQv/+, TF, Sum T, Fr + rF, Form%, SQu, FQfo, FQfu, H, (A), Ay, Na, Hh, FAB2, PSV, GHR, Mp, EB Pervasive, AdjD, M–, XA%, WDA%, X + %, X – %, S – %, F + %, EGO, PTI, SCZI), and the remaining 104 (76.42%) fell in the excellent range. Overall, Table 1 reveals that the process of inserting randomly generated error worked as expected by psychometric theory (Nunnally & Bernstein, 1994). As the proportion of response-level random error increases, summary score reliability decreases. Simultaneously, across samples, the summary score reliability coefficients remain quite high. These findings may seem surprising. However, the data indicate that CS summary scores are generally quite reliable despite forcing a considerable degree of random error into the responses. Although this is reassuring, it is essential to recognize that random error is very different from systematic error. This is an issue we return to later.

Impact of Sampling Variance on ICC Coefficients

Very often, researchers consider a sample of 20 Rorschach protocols to be sufficiently large for determining CS scoring reliability (cf. Weiner, 1991; also see Acklin, McDowell, Verschell, & Chan, 2000). Samples of this size may be quite appropriate for indexes of absolute coder agreement (e.g., percentage agreement) because these indexes are not dependent on between-subject variance. However, for statistics like κ or the ICC, a sample of 20 protocols may possess some unusual characteristics just by coincidence, and this can adversely affect chance-corrected reliability statistics.

As noted previously, there were several indications that between-subjects variance was constrained for some of the scores in Sample 3, which contained only 19 Rorschach protocols. The upper portion of Table 4 presents the nine scores from this sample with ICC values less than .70. The columns indicate the observed ICC in this sample, the ICC in the full population of 219 protocols, the MSB and MSW for the five raters in this sample, the average variance for each individual rater in this sample, the average variance for each rater in the composite sample, and an *F* ratio comparing the last two variances. Considering the last column of *F* ratios, it can be seen that the variance in Sample 3 was significantly smaller than the corresponding variance in the composite sample for six of the nine variables (and eight of nine using a one-tailed significance test). Treating each of the variances as the meta-analytic estimates they are (with aggregated $df = 90$ and 436), eight of the nine variables have significantly less variance in Sample 3 than in the composite sample. Thus, with the exception of DV1, there was generally restricted between-subjects variance for the coefficients with ICC values < .70.[7]

[7]The DV1 reliability coefficient was low largely because of one rater's lapse on a single case. For this case, four of the raters assigned between three and five DV1 scores, although the remaining rater assigned just a single DV1 response. With this rater excluded, the ICC increased from .64 to .72.

TABLE 4
Range Restriction As a Factor in the 19–Protocol Clinician Sample

Variable	Sample 3 ICC	Full Population ICC	Sample 3 MSB	Sample 3 MSW	M Variance in Sample 3 per Rater	M Variance in Population per Rater	Sample to Population F(18, 218) Ratio
Variables with ICC < .70 in the 19-protocol sample							
FY	.35	.90	1.44	.39	0.59	4.34	0.14***
DR2	.39	.85	1.56	.37	0.58	0.91	0.64
Level 2 SS	.54	.87	4.56	.65	1.44	3.16	0.46*
DR1	.60	.82	6.13	.72	1.80	3.16	0.57
DV1	.64	.84	4.38	.44	1.23	1.02	1.21
SQu	.65	.84	1.55	.15	0.43	0.96	0.45*
Id	.66	.92	1.53	.14	0.42	2.52	0.17***
PTI	.69	.88	3.06	.26	0.82	2.24	0.36**
AB	.69	.93	1.03	.08	0.28	0.85	0.32**
SQu and FY considered alongside relevant comparison variables							
SQo	.99	.88	4.61	.01	0.93	1.08	0.86
SQu	.65	.84	1.55	.15	0.43	0.96	0.45*
SQ–	.89	.91	5.77	.14	1.27	2.14	0.59
FC	.85	.83	11.23	.38	2.55	2.24	1.14
FC'	.87	.86	10.40	.31	2.32	2.74	0.85
FY	.35	.90	1.44	.39	0.59	4.34	0.14***
FT	.74	.78	3.02	.19	0.75	0.94	0.80
FV	.74	.72	2.21	.15	0.56	0.60	0.94

Note. ICC = intraclass correlation coefficient; MSB = mean square between (between-subjects variance); MSW = mean square within (rating error variance).
*$p < .05$. **$p < .01$. ***$p < .001$.

The lower portion of Table 4 illustrates this further by comparing two target scores with less than optimal reliability (SQu and FY) to relevant comparison variables. Data on SQu are presented along with statistics for two counterpart scores, SQo and SQ–. Data on FY are presented along with information for other Form Dominated Color and Shading determinants. In each case, it can be seen that the between-subjects variance in Sample 3 was constrained for the target score relative to the between-subjects variance for the counterpart scores, even though scoring error (i.e., MSW) was generally no worse for the target score than for the counterpart scores. These analyses demonstrate how sampling error can influence reliability. When 165 scores are examined in a relatively small sample, by chance alone some of the 165 characteristics are likely to be artificially constrained. This constraint will reduce interrater reliability coefficients.

Instability of Small Sample Reliability Estimates

The preceding raises broader questions about the accuracy of reliability estimates derived from small samples. Even though a 95% CI should indicate the range in

which one should find the true population reliability coefficient 95% of the time, as we noted earlier, nonnormal score distributions violate a critical assumption for computing ICC CIs. Thus, it is likely that CIs will be misleading when they are computed on CS scores with skewed and kurtotic distributions, particularly when the reliability sample is small. To more thoroughly investigate this issue, we conducted a series of ancillary analyses. Specifically, we considered the composite sample of 219 protocols to be the full population of interest. Consequently, the ICC values reported in Table 2 are considered population parameters that indicate the true reliability in the population (although with an *N* of 219, the data in Table 2 are themselves sample estimates of underlying population parameters). We then drew 100 random samples of 20 protocols each from this population. For each of the 20-protocol samples, we computed ICCs and the 95% CI around the observed values. This was done for all the scores listed in Table 2. Theoretically, the true population reliability coefficient (i.e., the values in Table 2) should fall within the computed CI 95 times out of 100. In other words, in 95 of the 100 random draws, the value reported in Table 2 should fall within the CI computed from the 20-protocol samples. To the extent that this does not happen, it indicates that problems exist when trying to estimate chance-corrected reliability coefficients from small samples.

Table 5 reports summary data from our analyses. The first column indicates the 164 scores under consideration (because no Pure Form responses were also assigned a FQ+ score, the score FQf+ was dropped from these analyses). The second column reports the mean base rate in the full population rounded to three decimal places (SQ+ and V had base rates of .0003 and .0002, respectively). Base rates are not reported for those variables in which it does not provide a meaningful index (i.e., weighted scores and percentage scores). The third column indicates the reliability coefficient observed in the population. The fourth and fifth columns report the mean and minimum reliability coefficients observed across the 100 randomly selected 20-protocol samples from the population (excluding those samples in which the reliability coefficient could not be computed because of a lack of variance). We do not report maximum values because the vast majority were 1.0 (for 107 scores) and the lowest was .96.

The sixth and seventh columns of Table 5 indicate the difference between the reliability observed in the full population and that observed in the 100 randomly selected 20-protocol samples. Two differences are reported. The sixth column reports the raw difference between the population ICC and the mean ICC across all 100 samples. The seventh column reports the difference after the ICC coefficients have been transformed using Fisher's *Z,* which is designed to correct for the skew inherent in the distribution of correlational values.

The eighth and ninth columns of Table 5 report the percentage of times when the ICC coefficient in the population fell outside the 95% CIs derived from the one hundred 20-protocol samples. Because CIs cannot be determined when a score has no variance or when an observed ICC is unity, the percentages only consider instances when it was possible to compute CIs. Specifically, the eighth column re-

TABLE 5
The Impact of Sampling Error on 164 Observed ICC Values: Interrater Reliability in the Composite Sample of 219 Protocols Relative to100 Random Draws of 20 Protocols Each

	Population		*Sample ICC*		*Population Minus Sample Difference*		*% When Population ICC Is*		*% When Sample ICC Is*		
Score	*M BR*	*ICC*	*M*	*Minimum*	*ICC*	*Z ICC*	*< SLL*	*> SUL*	*< .40*	*< .60*	*< Population – .10*
Zf	0.579	0.98	0.98	0.92	.01	0.00	.12	0.09	.00	.00	.00
ZSum	—	0.98	0.97	0.90	.01	–0.01	.14	0.09	.00	.00	.00
W	0.420	0.99	0.99	0.97	.00	–0.05	.15	0.18	.00	.00	.00
Sum Wv	0.025	0.92	0.83	0.44	**.10**	0.28	.12	0.40	.00	.04	.45
D	0.444	0.99	0.99	0.96	.00	–0.08	.21	0.14	.00	.00	.00
Dd	0.136	0.98	0.97	0.84	.01	–0.10	.27	0.16	.00	.00	.01
S	0.130	0.94	0.93	0.70	.01	–0.07	.18	0.13	.00	.00	.09
DQ+	0.292	0.98	0.97	0.79	.01	–0.06	.25	0.16	.00	.00	.07
DQo	0.658	0.99	0.99	0.94	.00	0.02	.11	0.15	.00	.00	.00
DQv/+	0.009	0.81	0.77	–0.06	.04	0.09	.10	0.16	.07	.15	.29
DQv	0.041	0.91	0.85	0.46	**.06**	0.11	.13	0.27	.00	.05	.33
DQ+ & FQ–	0.075	0.96	0.93	0.36	.03	0.08	.15	0.17	.01	.01	.06
DQo & FQ–	0.173	0.96	0.94	0.80	.01	–0.01	.11	0.13	.00	.00	.04
DQ/ & FQ–	0.003	0.87	0.77	0.00	**.09**	0.47	.00	0.52	.12	.16	.29
DQv & FQ–	0.008	0.75	0.53	–0.12	**.22**	0.24	.14	0.42	.30	.53	.58
FQx+	0.005	0.91	0.68	–0.03	**.23**	0.61	.15	0.60	.23	.28	.45
FQxo	0.486	0.98	0.97	0.91	.01	0.00	.07	0.08	.00	.00	.00
FQxu	0.232	0.93	0.91	0.71	.02	0.01	.10	0.13	.00	.00	.11
FQx–	0.259	0.96	0.95	0.82	.01	–0.01	.13	0.17	.00	.00	.05
FQxNone	0.017	0.93	0.83	0.00	**.10**	0.39	.05	0.51	.03	.11	.39
FQfo	0.180	0.96	0.96	0.87	.00	–0.08	.16	0.04	.00	.00	.00
FQfu	0.088	0.91	0.88	0.47	.03	–0.01	.12	0.12	.00	.02	.15

FQf–	0.096	0.92	0.90	0.49	.03	0.01	.11	0.11	.00	.02	.12
MQ+	0.004	0.90	0.63	–0.06	**.28**	0.68	.23	0.63	.30	.33	.48
MQo	0.103	0.94	0.92	0.74	.02	0.03	.12	0.12	.00	.00	.11
MQu	0.037	0.82	0.80	0.38	.02	–0.02	.10	0.08	.01	.08	.20
MQ–	0.049	0.92	0.90	0.67	.01	–0.02	.09	0.08	.00	.00	.08
MQNone	0.001	0.66	0.45	0.00	**.21**	0.80	.00	1.00	.55	.55	.55
SQ+	0.000	0.66	0.35	0.00	**.32**	0.70	.00	0.87	.62	.62	.62
SQo	0.040	0.88	0.85	0.45	.03	0.00	.13	0.08	.00	.03	.19
SQu	0.030	0.84	0.80	0.33	.04	0.03	.10	0.13	.01	.11	.28
SQ–	0.057	0.91	0.90	0.56	.01	–0.09	.19	0.09	.00	.01	.14
SQnone	0.002	0.82	0.64	–0.03	**.18**	0.78	.00	0.58	.27	.30	.47
WDFQ+	0.005	0.91	0.68	–0.03	**.23**	0.61	.15	0.60	.23	.28	.45
WDFQo	0.459	0.97	0.97	0.88	.01	0.00	.06	0.09	.00	.00	.00
WDFQu	0.185	0.93	0.91	0.77	.02	0.00	.14	0.09	.00	.00	.09
WDFQ–	0.198	0.97	0.96	0.79	.01	0.03	.10	0.16	.00	.00	.03
WDFQNone	0.016	0.93	0.83	0.00	**.10**	0.37	.07	0.49	.03	.12	.40
Blends	0.228	0.93	0.92	0.67	.01	–0.04	.12	0.08	.00	.00	.08
C – Sh Blend	0.042	0.87	0.86	0.37	.02	–0.05	.17	0.08	.01	.02	.16
M	0.192	0.96	0.95	0.82	.01	0.00	.10	0.09	.00	.00	.03
FM	0.149	0.95	0.94	0.83	.01	–0.03	.13	0.03	.00	.00	.02
m	0.079	0.92	0.90	0.66	.02	–0.02	.14	0.08	.00	.00	.10
FC	0.075	0.83	0.79	0.47	.04	0.03	.06	0.11	.00	.08	.25
CF	0.075	0.84	0.80	0.43	.03	0.00	.13	0.10	.00	.06	.26
C	0.027	0.92	0.87	0.16	**.05**	0.13	.14	0.22	.03	.06	.21
Cn	0.001	0.97	0.43	0.00	**.54**	1.92	.00	0.92	.57	.57	.57
FC'	0.080	0.86	0.83	0.52	.03	–0.01	.12	0.10	.00	.03	.24
C'F	0.017	0.62	0.59	–0.03	.02	–0.05	.17	0.10	.23	.47	.34
C'	0.004	0.71	0.56	–0.05	**.15**	0.33	.05	0.41	.32	.40	.40
FT	0.032	0.78	0.75	0.34	.03	–0.04	.16	0.13	.03	.17	.30
TF	0.006	0.54	0.52	–0.12	.02	0.05	.09	0.15	.34	.55	.35
T	0.001	0.97	0.76	0.00	**.21**	1.94	.00	1.00	.21	.21	.29

(continued)

TABLE 5 (Continued)

	Population		*Sample ICC*		*Population Minus Sample Difference*		*% When Population ICC Is*		*% When Sample ICC Is*		
Score	*M BR*	*ICC*	*M*	*Minimum*	*ICC*	*Z ICC*	*< SLL*	*> SUL*	*< .40*	*< .60*	*< Population – .10*
FV	0.016	0.72	0.65	0.10	**.08**	0.05	.13	0.20	.20	.41	.41
VF	0.005	0.81	0.77	–0.03	.04	0.09	.13	0.14	.10	.11	.22
V	0.000	1.00	1.00	1.00	.00	—	—	—	.00	.00	.00
FY	0.072	0.90	0.86	0.51	.04	0.04	.11	0.16	.00	.02	.22
YF	0.018	0.82	0.74	0.12	**.08**	0.07	.14	0.25	.10	.22	.35
Y	0.005	0.79	0.64	–0.06	**.16**	0.44	.06	0.48	.26	.34	.48
Fr	0.019	0.98	0.98	0.79	–.01	0.30	.00	0.22	.00	.00	.05
rF	0.003	0.85	0.54	–0.06	**.31**	0.61	.12	0.68	.39	.40	.52
FD	0.048	0.88	0.86	0.52	.02	–0.05	.17	0.09	.00	.02	.18
F	0.368	0.97	0.97	0.89	.00	–0.10	.22	0.06	.00	.00	.00
Pairs	0.363	0.97	0.97	0.86	.00	–0.15	.31	0.15	.00	.00	.02
H	0.123	0.98	0.97	0.84	.00	–0.08	.20	0.06	.00	.00	.01
(H)	0.075	0.93	0.91	0.66	.02	–0.01	.17	0.14	.00	.00	.14
Hd	0.082	0.92	0.91	0.40	.01	–0.23	.33	0.12	.01	.03	.13
(Hd)	0.022	0.88	0.82	–0.06	**.06**	0.04	.16	0.25	.04	.10	.30
Hx	0.014	0.96	0.81	–0.06	**.15**	0.59	.09	0.49	.10	.12	.38
A	0.363	0.99	0.98	0.82	.00	–0.08	.22	0.15	.00	.00	.01
(A)	0.033	0.94	0.89	0.14	**.05**	0.08	.18	0.27	.01	.04	.23
Ad	0.093	0.95	0.95	0.77	.00	–0.21	.28	0.10	.00	.00	.07
(Ad)	0.005	0.72	0.64	–0.06	**.08**	0.20	.04	0.20	.16	.30	.30
An	0.059	0.92	0.91	0.24	.01	–0.28	.45	0.13	.02	.04	.13
Art	0.031	0.91	0.86	0.44	**.06**	0.10	.18	0.28	.00	.08	.28
Ay	0.025	0.84	0.81	–0.03	.03	0.00	.18	0.17	.03	.12	.23
Bl	0.027	0.98	0.98	0.79	.01	0.53	.00	0.60	.00	.00	.03

Bt	0.050	0.96	0.96	0.87	.00	0.03	.05	0.04	.00	.00	.00
Cg	0.081	0.95	0.94	0.78	.01	–0.03	.13	0.06	.00	.00	.01
Cl	0.011	0.87	0.84	0.00	.03	0.17	.05	0.26	.04	.09	.19
Ex	0.005	0.86	0.84	–0.03	.01	0.35	.00	0.37	.05	.07	.19
Food	0.016	0.94	0.91	0.34	.03	0.22	.03	0.26	.01	.03	.13
Fi	0.024	0.93	0.93	0.51	.00	0.03	.03	0.15	.00	.01	.09
Geog	0.007	0.94	0.92	0.00	.02	0.67	.00	0.68	.02	.03	.20
Hh	0.036	0.85	0.83	0.33	.02	–0.08	.22	0.14	.04	.13	.17
Ls	0.041	0.94	0.91	0.63	.03	0.08	.12	0.15	.00	.00	.13
Na	0.029	0.94	0.92	0.45	.03	–0.01	.18	0.16	.00	.03	.13
Sc	0.035	0.93	0.91	0.68	.02	–0.04	.19	0.15	.00	.00	.14
Sx	0.042	0.96	0.94	0.63	.03	0.10	.10	0.22	.00	.00	.08
Xy	0.006	0.95	0.89	0.00	**.05**	0.82	.00	0.77	.05	.05	.24
Idio	0.053	0.92	0.86	0.45	**.06**	0.12	.09	0.25	.00	.03	.28
DV1	0.022	0.84	0.72	–0.05	**.12**	0.14	.13	0.29	.12	.27	.44
INC1	0.032	0.74	0.72	0.27	.02	–0.09	.19	0.11	.10	.19	.26
DR1	0.064	0.82	0.79	0.26	.03	0.00	.10	0.10	.01	.10	.24
FAB1	0.037	0.90	0.83	0.52	**.08**	0.14	.18	0.36	.00	.02	.41
ALOG	0.017	0.69	0.70	–0.15	–.02	–0.17	.27	0.12	.13	.26	.25
CONTAM	0.004	0.76	0.63	–0.09	**.13**	0.31	.08	0.39	.25	.29	.48
DV2	0.005	0.80	0.59	–0.07	**.22**	0.35	.15	0.50	.33	.41	.48
INC2	0.010	0.82	0.76	0.00	**.05**	0.12	.13	0.25	.11	.20	.27
DR2	0.022	0.85	0.79	–0.03	**.06**	0.01	.16	0.17	.04	.11	.28
FAB2	0.016	0.81	0.75	0.12	**.06**	0.04	.14	0.18	.08	.19	.33
Sum6	0.231	0.91	0.90	0.55	.00	–0.08	.15	0.09	.00	.01	.09
WSum6	—	0.90	0.89	0.63	.01	–0.05	.10	0.07	.00	.00	.11
AB	0.018	0.93	0.87	0.34	**.06**	0.24	.13	0.40	.01	.05	.33
AG	0.039	0.90	0.86	0.41	.03	0.03	.13	0.10	.00	.02	.19
CFB	0.001	0.96	0.81	0.00	**.15**	1.94	.00	1.00	.19	.19	.19
COP	0.039	0.86	0.83	0.28	.03	–0.06	.22	0.16	.02	.08	.25
CP	0.001	0.79	0.60	–0.03	**.20**	0.89	.00	0.76	.36	.36	.45

(continued)

TABLE 5 (Continued)

	Population		*Sample ICC*		*Population Minus Sample Difference*		*% When Population ICC Is*		*% When Sample ICC Is*		
Score	*M BR*	*ICC*	*M*	*Minimum*	*ICC*	*Z ICC*	*< SLL*	*> SUL*	*< .40*	*< .60*	*< Population – .10*
GHR	0.141	0.89	0.88	0.76	.01	0.01	.04	0.00	.00	.00	.05
PHR	0.193	0.94	0.93	0.56	.01	–0.08	.26	0.09	.00	.01	.07
MOR	0.084	0.89	0.87	0.34	.02	–0.01	.08	0.10	.01	.02	.12
PER	0.059	0.94	0.92	0.72	.02	–0.03	.19	0.13	.00	.00	.11
PSV	0.011	0.84	0.80	–0.03	.04	0.08	.08	0.14	.04	.09	.24
R	1.000	1.00	1.00	0.98	.00	0.02	.20	0.26	.00	.00	.00
Lambda	—	0.98	0.94	0.54	.04	0.19	.17	0.50	.00	.01	.15
Form %	—	0.95	0.95	0.73	.00	–0.09	.12	0.05	.00	.00	.02
FM + m	0.229	0.95	0.94	0.77	.01	–0.04	.16	0.06	.00	.00	.03
WSumC	—	0.94	0.93	0.64	.01	–0.12	.21	0.12	.00	.00	.07
Sum Shading	0.257	0.94	0.93	0.76	.00	–0.05	.07	0.05	.00	.00	.05
EA	0.345	0.96	0.95	0.80	.00	–0.06	.14	0.03	.00	.00	.02
es	0.485	0.95	0.95	0.83	.00	–0.05	.06	0.06	.00	.00	.04
Adj es	0.377	0.94	0.93	0.81	.01	–0.02	.06	0.04	.00	.00	.01
EBPer	—	0.92	0.91	0.72	.01	–0.05	.11	0.04	.00	.00	.05
D Score	—	0.89	0.88	0.61	.02	–0.01	.08	0.07	.00	.00	.10
AdjD	—	0.87	0.85	0.64	.02	0.01	.05	0.02	.00	.00	.16
Sum C'	0.101	0.90	0.89	0.73	.01	–0.03	.08	0.05	.00	.00	.09
Sum V	0.021	0.81	0.78	0.27	.03	–0.05	.23	0.15	.03	.17	.30
Sum T	0.039	0.89	0.86	0.55	.03	–0.02	.18	0.18	.00	.02	.27
Sum Y	0.096	0.91	0.87	0.49	.04	0.04	.07	0.15	.00	.03	.20
CF + C + Cn	0.103	0.89	0.87	0.48	.02	–0.03	.13	0.08	.00	.01	.17
Afr	—	0.99	1.00	0.94	.00	0.06	.21	0.31	.00	.00	.00
Blends/R	—	0.93	0.92	0.73	.01	–0.02	.07	0.04	.00	.00	.04

Active	0.276	0.96	0.95	0.84	.01	–0.05	.18	0.09	.00	.00	.02
Passive	0.147	0.92	0.90	0.75	.02	0.03	.09	0.10	.00	.00	.11
All H Cont	0.302	0.98	0.97	0.73	.01	–0.18	.31	0.12	.00	.00	.03
(H) + (Hd)	0.097	0.95	0.94	0.79	.01	–0.01	.15	0.16	.00	.00	.05
(A) + (Ad)	0.038	0.92	0.88	0.16	**.05**	0.05	.19	0.22	.02	.02	.21
H(H)A(A)	0.595	0.99	0.99	0.92	.00	–0.07	.16	0.11	.00	.00	.00
Hd(Hd)Ad(Ad)	0.201	0.95	0.94	0.72	.01	–0.09	.25	0.12	.00	.00	.09
Isol Index	—	0.96	0.95	0.79	.00	–0.16	.23	0.12	.00	.00	.04
Ma	0.126	0.96	0.95	0.80	.01	–0.05	.15	0.15	.00	.00	.01
Mp	0.069	0.91	0.88	0.65	.03	0.03	.10	0.10	.00	.00	.14
Intel Index	—	0.93	0.90	0.55	.03	0.03	.17	0.23	.00	.01	.20
Sum6 Lvl 2	0.054	0.87	0.84	0.46	.04	0.03	.10	0.08	.00	.04	.16
XA%	—	0.92	0.91	0.76	.01	–0.06	.17	0.05	.00	.00	.07
WDA%	—	0.94	0.94	0.77	.01	–0.03	.13	0.09	.00	.00	.07
X – %	—	0.93	0.92	0.77	.01	–0.04	.14	0.06	.00	.00	.06
S – %	—	0.75	0.75	0.18	.00	–0.22	.28	0.13	.08	.24	.26
Popular	0.212	0.93	0.93	0.81	.01	–0.03	.12	0.04	.00	.00	.04
X + %	—	0.96	0.96	0.85	.01	–0.04	.08	0.07	.00	.00	.02
F + %	—	0.93	0.92	0.71	.01	–0.09	.18	0.05	.00	.00	.05
Xu%	—	0.83	0.81	0.46	.02	–0.03	.15	0.09	.00	.09	.21
Zd	—	0.93	0.92	0.77	.01	–0.05	.09	0.07	.00	.00	.07
Egocent Ind	—	0.96	0.95	0.76	.01	–0.05	.15	0.07	.00	.00	.01
Fr + rF	0.023	0.98	0.98	0.86	.00	0.26	.00	0.23	.00	.00	.01
An + Xy	0.065	0.92	0.92	0.24	.00	–0.33	.43	0.15	.01	.04	.14
NonPureH	0.179	0.96	0.94	0.69	.01	–0.10	.27	0.15	.00	.00	.10
PTI	0.074	0.88	0.87	0.64	.01	–0.05	.13	0.06	.00	.00	.12
SCZI	0.115	0.87	0.87	0.61	.00	–0.07	.12	0.05	.00	.00	.13
DEPI	0.187	0.84	0.83	0.29	.01	–0.02	.04	0.07	.01	.01	.17
CDI	0.122	0.87	0.87	0.75	.00	–0.06	.05	0.00	.00	.00	.02
S Con	0.242	0.83	0.82	0.52	.01	–0.04	.10	0.02	.00	.02	.18

(continued)

TABLE 5 (Continued)

	Population		*Sample ICC*		*Population Minus Sample Difference*		*% When Population ICC Is*		*% When Sample ICC Is*		
Score	*M BR*	*ICC*	*M*	*Minimum*	*ICC*	*Z ICC*	*< SLL*	*> SUL*	*< .40*	*< .60*	*< Population – .10*
HVI	0.133	0.91	0.90	0.74	.01	–0.05	.11	0.05	.00	.00	.08
OBS	0.053	0.91	0.89	0.33	.02	0.00	.13	0.13	.01	.01	.15
M	.11	.90	0.86	0.48	.04	.10	.13	.21	.05	.08	.17
Minimum	.00	.54	0.35	–0.15	–.02	–.33	.00	.00	.00	.00	.00
Maximum	1.00	1.00	1.00	1.00	.54	1.94	.45	1.00	.62	.62	.62
SD	.16	.08	.12	.34	.07	.33	.08	.21	.11	.14	.15
Kurtosis	8.65	3.39	2.92	–1.22	15.01	15.68	2.08	4.20	9.20	4.19	.10
Skewness	2.57	–1.64	–1.66	–.40	3.39	3.53	.83	2.10	2.94	2.16	.91

Note. Raw differences of .05 or greater between the population ICC and the mean sample ICCs in Column 6 are in bold. In the last column, an underlined value indicates instances in which the small samples underestimated the population coefficient by a magnitude of .10 or more at least 10% of the time. ICC = intraclass correlation coefficient; BR = mean base rate in the full population; SLL = sample lower limit; SUL = sample upper limit.

ports the percentage of times when the population ICC fell below the lower limit of the 95% CI from the samples, whereas the ninth column reports the percentage of times it fell above the upper limit. In theory, if the CIs are accurate, the population ICC should fall below the lower limit 2.5% of the time and above the upper limit 2.5% of the time, so the values in each column should be .025.

Finally, the last three columns in Table 5 indicate the percentage of times in which the one hundred 20-protocol samples deviated from fixed benchmarks. Specifically, the tenth column indicates the percent of times that the 20-protocol samples indicated reliability was poor (i.e., ICC < .40). The eleventh column indicates the percent of times when the randomly drawn samples indicated reliability was not good (i.e., ICC < .60). The final column indicates the proportion of samples that found ICC values that were lower than the population ICC value by more than .10 (e.g., when a sample indicated reliability was < .75 when in fact the population ICC was .85).

To illustrate, consider the fourth row examining the sum of Wv responses. This score had a base rate in the population of .025, indicating that it occurred 2.5 times out of every 100 responses. It had an ICC of .92 in the full population. Across the 100 random samples of 20 protocols, the mean ICC was .83 and the minimum was .44 (the unreported maximum was .99). The average small sample ICC was lower than the population parameter by a raw value of .10 (i.e., .921 – .8255 = .955) and by a magnitude of .28 after transforming the ICC values by Fisher's *Z*. With respect to the 95% CIs from the 100 small samples, the population ICC value (i.e., .92) fell below the sample's lower limit 12% of the time but fell above the sample's upper limit 40% of the time. None of the 100 subsamples produced an ICC coefficient less than .40. However, 4% of the time the small samples produced an ICC less than .60 and 45% of the time the small samples produced an ICC that was lower than the population value by a magnitude of at least .10. Overall, the data indicate that small samples consistently underestimated the true reliability for this score.

Perusing Table 5, it can be seen that many variables had relatively constant ICC values (e.g., Zf, D, DQo, F, R), whereas others changed considerably across the randomly selected subsamples. For instance, DQv/+ reliability ranged from –0.06 to 1.00 even though the population parameter was .81, FQxNone ranged from .00 to 1.00 even though the population parameter was .93, (Hd) ranged from –0.06 to 1.00 even though the true reliability was .88, FAB2 ranged from .12 to 1.0 despite a population parameter of .81, COP ranged from 0.28 to 1.0 around a parameter of .86, and Sum Y ranged from 0.49 to 0.99 even though the population parameter was .91.

A careful examination of Table 5 reveals many instances in which there are marked disparities between the mean ICC derived from the small samples and the true ICC in the population (e.g., FQx+, MQNone, Cn, T, rF, DV2, CP). Furthermore, for these kinds of scores, the small samples give very imprecise and misleading estimates of reliability. Considering just the 38 scores with a population versus sample raw difference of .05 or more (in bold), on average the true reliability in the popula-

tion is higher than the upper limit of the sample's CI about half the time (i.e., 48%). The true reliability is lower than the lower limit of the sample's CI an average of just 10% of the time. Thus, small samples often produce misleading reliability results. When this occurs, the small sample results underestimate the true reliability about five times more often than they overestimate it.

More important, the scores that are most likely to be underestimated by small samples can be predicted with a substantial degree of accuracy. In particular, as scores becomes less frequent in the population, small samples are more likely to provide underestimates of their true reliability. Figures 1 and 2 plot this relation. Figure 1 shows the association between score base rates in the population (horizontal axis; using a logarithmic scale) and the raw difference between the population ICC and the mean ICC observed across the 100 random samples (vertical axis). In other words, the figure plots the data from columns 2 and 6 in Table 5. It can be seen that there is a negative curvilinear relationship ($R = .80$) such that the true reliability of a score is more drastically underestimated for increasingly rare scores. Small samples provide adequate estimates of scores that occur at least 5 times in 100 responses (i.e., base rate > .05). For scores that occur less often, and particularly those that occur less than 1 time in 100 responses (which corresponds to about 15% of the scores in our analyses), small samples are increasingly likely to underestimate reliability and the underestimates are increasingly severe.

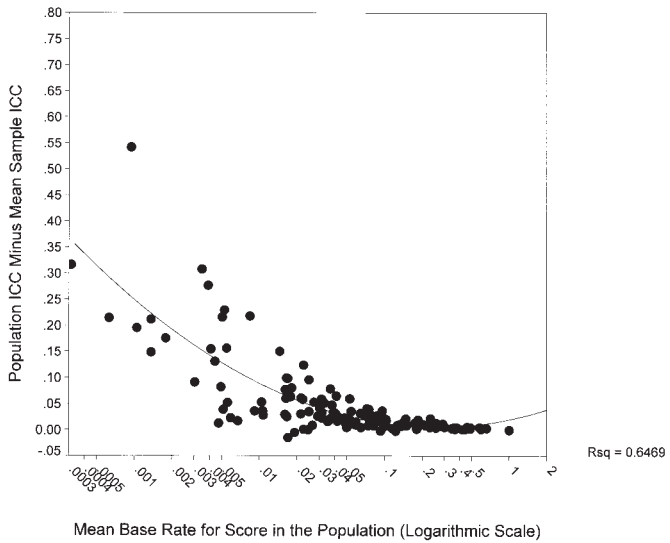

FIGURE 1 The association between score base rates in the population and the raw difference between the population intraclass correlation coefficient (ICC) and the mean ICC observed across the 100 random samples of 20 protocols.

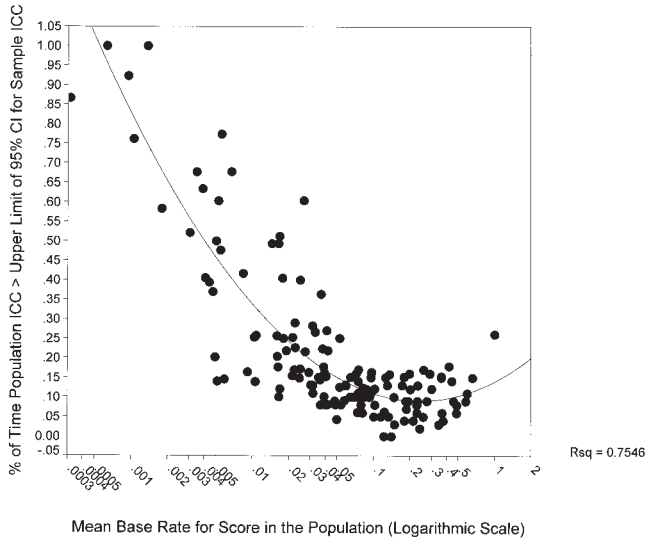

FIGURE 2 The association between score base rates in the population and the extent to which the population intraclass correlation coefficient (ICC) exceeds the upper limit of the 95% confidence interval (CI) in the 100 random samples of 20 protocols.

Figure 2 plots a similar relation. The vertical axis now quantifies the extent to which the population ICC exceeds the upper limit of the 95% CI in the smaller samples (i.e., column 9 in Table 5). There is again a strong negative nonlinear relationship with base rates in the population ($R = .87$). Once again, the small samples provide more reasonable CIs for scores that occur at least 5 times in 100 responses (although even at this frequency, the true ICC values fall above the confidence range much more often than the expected rate of .025). In general, for scores that occur less than 1 time in 100 responses, true reliability falls above the upper limit of the sample's 95% CI between 30% and 100% of the time.

Returning to Table 5, note that there are two instances in which the mean ICC across samples is slightly higher than the population ICC (see column 6; Fr = –.0052; ALOG = –.0151). For both scores, the disparity occurred because the rater pair for a single protocol in the full population produced outlier results. Consequently, when the random samples did not include this rater pair, the sample reliabilities tended to be a bit higher. More specifically, one rater in the student sample assigned two Fr scores to a patient, whereas the other rater did not assign any. Excluding this protocol, the ICC difference became +.007 rather than –.0052. In the applied reliability sample, one rater assigned five ALOG scores to a patient, whereas the other rater did not assign any. Excluding this protocol, the ICC difference became +.0600 rather than –.0151.

A careful review of Table 5 also reveals numerous instances in which the raw difference between the population ICC and the average sample ICC (i.e., column 6) has a positive value, whereas the Fisher's *Z* transformed difference (column 7) has a negative value. In these instances, the Fisher's *Z* coefficients are artifactual. The positive raw difference truly indicates that the population ICC is larger than the mean ICC obtained across the 100 small samples. The transformed Fisher's *Z* coefficients suggest just the opposite (i.e., that the average small sample ICC is larger than the population ICC) because the Fisher's *Z* transformation is imprecise. Particularly when a raw coefficient is large, the transformation creates an upward bias that artificially inflates the *Z* transformed coefficient (see Hunter & Schmidt, 1990). Across the 100 small samples, there are many opportunities for this bias to emerge, which in turn creates the seemingly contradictory data in those two columns in Table 5.

Overall, the data in Table 5 indicate that chance factors related to sampling error affect observed ICC values in relatively small samples (for additional compelling data, see Carroll & Faden, 1978). Sampling error can occur through at least two relatively independent processes. First, chance affects the specific protocol–coder pairs selected for analysis. To the extent protocols are relatively easy or hard to score, or to the extent coder skills are aligned across pairings, the observed disagreement among coders (i.e., MSW) will either be too large or too small relative to the population parameter. If the MSW is biased by being too large, it will generate lower ICC values; if it is too small, it will generate larger than expected ICC values. The second way chance will affect ICC values is through between-subjects variance. Relative to the population at large, if patients in a sample differ more than average on certain characteristics, ICC values for corresponding Rorschach variables will be larger than expected because the MSB will be larger than average. Conversely, to the extent patients in a sample happen to be quite similar on certain characteristics, ICC values for those characteristics will be smaller than expected. The latter processes affect rare scores more often than common scores.

Response-Level Reliability Versus Patient-Level Reliability

In the next set of analyses, we tested whether the reliability of scores assigned to each response present a more stringent standard than the reliability of summary scores evaluated across patients. This was accomplished by comparing response-level κ coefficients to their corresponding summary score ICC values. For this analysis, we limited the variables to those that were assigned at least one time by each rater across samples. We also only examined those scores for which a dichotomous scoring decision (i.e., present vs. absent) could be made at the response level and then compared to an equivalent score at the protocol level. The variable R was excluded because there are problems evaluating this score at the response level.

Specifically, even though it is easy to determine when both raters agree a response is present or when one rater believes a response is present and the other does not, it is impossible to count the number of times both raters agree a response is absent. Consequently, it is impossible to compute a meaningful kappa coefficient for R at the response level of analysis. In total, we evaluated 108 scores.

Because κ and the ICC are equivalent chance-corrected statistics for dichotomously assigned scores in increasingly large samples,[8] and because all 108 scores can be considered dichotomous assignments to each response (i.e., present vs. absent), the impact of aggregating scores across responses can be seen by the difference between summary score ICC coefficients and response-level κ coefficients for the same variables. Table 6 presents summary results for these analyses. The table reports the mean κ and ICC values across all 108 scores. It also indicates the mean and median difference between these two statistics (i.e., ICC – κ) and the percentage of times the ICC was larger than κ by a difference of .10 or more. The last two columns report the difference after converting both coefficients to Fisher's *Z* statistics (i.e., $Z_{ICC} - Z_{\kappa}$). Summary results are presented for the seven samples that had more than 50 participants. Sample 3 was excluded because, as we demonstrated already, its size ($N = 19$) would have produced less stable findings.

Focusing on the raw differences, it can be seen that the results from each sample support psychometric theory. Summary scores that have been aggregated across all the responses produced by a patient are more reliable than scores assigned to individual responses because random measurement errors tend to cancel out as information is accumulated across responses. Table 6 also indicates that the effect of aggregation is more salient when there is a greater degree of random scoring error. In the clinical samples, the greatest degree of difference was observed with the student raters, and the smallest differences were seen with the experienced raters. In the student sample, on average across these 108 scores the response-level κ coefficients were lower than the summary score ICC coefficients by a magnitude of .09 (i.e., $M = .69$ vs. $M = .78$, respectively). For 36% of the scores in this sample, response-level reliability was lower than summary score reliability by a magnitude of at least .10. In contrast, with the experienced raters the average difference was .02 (i.e., $M = .91$ vs. $M = .93$, respectively) and only 1% of the scores had a reliability difference of .10 or greater. The same pattern was observed for the samples containing 10%, 20%, and 30% random scoring error. As one moves across these samples, the average difference between response-level reliability and summary

[8] As Fleiss and Cohen (1973) indicated, ICC calculations divide between-subjects variation by $N - 1$ rather than N to obtain the MSB. To make the ICC exactly equivalent to κ, the MSB must be multiplied by $(N - 1)/N$. However, for large samples like those used here, the difference between the two ICC computations is trivial. In our composite sample, on average the reported ICC values (generated by MSB with $df = 218$) for all 164 summary scores were too large by a magnitude of .00042 (maximum = .00162) relative to the calculations that would make them exactly equivalent to κ. In our smallest clinical sample ($N = 65$), the average difference was .00089 (maximum = .00558).

TABLE 6
The Superior Reliability of Comprehensive System Summary Scores: A Summary of Response-Level Reliability Coefficients (κ) Versus Total Score Reliability Coefficients (ICC) for 108 Variables Across Samples

			Raw Difference		*% of Raw Difference*	*Fisher's Z Difference*	
Sample	*M* κ	*M ICC*	*M*	*Mdn*	≥ .10	*M*	*Mdn*
Clinical samples							
1. Student	.69	.78	.09	.06	36	.23	.21
2. Experienced	.91	.93	.02	.02	01	.27	.27
4. Applied	.89	.92	.03	.03	03	.37	.30
5. Composite	.84	.89	.05	.04	10	.29	.26
Forced random error samples							
6. 10% error	.88	.94	.06	.06	16	.42	.38
7. 20% error	.77	.88	.11	.11	57	.39	.39
8. 30% error	.66	.80	.14	.14	67	.36	.34

Note. Kappa (κ) was calculated across responses (*N* = 1,407; 1,299; 1,667; 4,761; 1,378; 1,378; and 1,378, respectively for each sample). ICC was calculated across protocols (*N* = 66, 65, 69, 219, 57, 57, and 57, respectively). ICC = intraclass correlation coefficient; raw difference = ICC – κ; *Z* difference = Fisher's Z_{ICC} – Fisher's Z_{κ}.

score reliability increased from .06 to .14, whereas the proportion of scores that had noticeably lower response-level reliability increased dramatically from 16% to 67%. Thus, the pattern supports psychometric theory. Like self-report scales that sum across items, random CS coding errors tend to cancel out when the scores are aggregated across individual responses, resulting in summary scales that are more reliable than the scores assigned to individual responses. Furthermore, the impact of aggregation is more pronounced in instances when the response-level scoring is less accurate. This suggests that although skilled raters benefit to some extent from summarizing scores across responses, it is relatively less proficient raters who benefit the most.

Table 7 presents results for specific scores to illustrate the difference between response-level and protocol-level reliability. Data are presented from the four samples that provide optimal contrasts, including the student coders versus the experienced raters and the sample containing 30% random error versus the sample containing 10% random error. Reviewing the student rater sample, it can be seen that aggregation across responses often leads to dramatic gains in reliability for those scores the raters found difficult to code on a response-by-response basis (e.g., FQxu, INC1, DR1). Considering the cognitive Special Scores (i.e., DV1 to FAB2), with the exception of DR2, the response-by-response coding would be classified as poor or fair although the summary score coding would be considered good or excellent. Thus, even though these coders could not synonymously clas-

TABLE 7
Illustrating the Difference Between Response Level Reliability (κ) and Summary Score Reliability (ICC) With Specific Results From Four of the Samples

	Student			*Experienced*			*30% Error*			*10% Error*		
	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*
W	.95	0.99	0.04	1.00	1.00	.00	0.45	0.76	.31	0.81	0.96	.15
Wv	.72	0.72	0.00	0.90	0.94	.04	0.68	0.79	.12	0.87	0.92	.05
D	.88	0.97	0.09	0.99	1.00	.01	0.51	0.90	.39	0.83	0.98	.15
Dd	.85	0.96	0.11	0.99	1.00	.01	0.63	0.86	.23	0.90	0.97	.07
S	.83	0.87	0.04	0.95	0.97	.02	0.66	0.78	.12	0.89	0.97	.08
DQ+	.86	0.92	0.06	0.95	0.98	.02	0.45	0.81	.36	0.81	0.97	.15
DQo	.83	0.98	0.15	0.93	0.99	.07	0.39	0.84	.46	0.79	0.97	.18
DQv/+	.75	0.79	0.04	0.77	0.75	–.02	0.65	0.74	.09	0.94	0.96	.02
DQv	.75	0.78	0.03	0.84	0.91	.07	0.69	0.84	.15	0.86	0.95	.08
DQ+ & FQ–	.66	0.86	0.20	0.91	0.96	.05	0.59	0.80	.21	0.80	0.93	.13
DQo & FQ–	.73	0.91	0.18	0.91	0.96	.05	0.67	0.87	.20	0.91	0.98	.06
DQv & FQ–	.21	0.47	0.26	0.65	0.64	.00	0.56	0.69	.13	0.70	0.84	.14
FQxo	.84	0.94	0.11	0.98	0.99	.01	0.51	0.84	.33	0.84	0.94	.10
FQxu	.59	0.82	0.24	0.92	0.95	.03	0.57	0.90	.33	0.87	0.98	.11
FQx–	.72	0.88	0.17	0.96	0.98	.02	0.56	0.91	.35	0.83	0.97	.13
FQxNone	.62	0.78	0.16	1.00	1.00	.00	0.69	0.86	.17	0.88	0.94	.07
FQfo	.85	0.95	0.10	0.97	0.98	.01	0.65	0.72	.07	0.88	0.93	.05
FQfu	.62	0.85	0.23	0.92	0.95	.03	0.64	0.68	.04	0.91	0.93	.02
FQf–	.67	0.86	0.18	0.92	0.97	.05	0.64	0.84	.20	0.87	0.95	.07
MQo	.87	0.80	–0.07	0.97	0.99	.02	0.68	0.76	.08	0.91	0.94	.03
MQu	.66	0.72	0.06	0.87	0.81	–.07	0.65	0.77	.12	0.90	0.95	.05
MQ–	.71	0.87	0.16	0.96	0.98	.02	0.64	0.73	.09	0.84	0.91	.07
SQo	.76	0.79	0.03	0.93	0.93	.00	0.66	0.57	–.09	0.89	0.91	.02

(continued)

TABLE 7 (Continued)

	Student			*Experienced*			*30% Error*			*10% Error*		
	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*
SQu	.61	0.67	0.06	0.83	0.86	.03	0.67	0.67	.00	0.92	0.96	.04
SQ–	.69	0.79	0.10	0.94	0.96	.01	0.74	0.80	.05	0.90	0.95	.05
WDFQo	.84	0.92	0.09	0.98	0.99	.01	0.54	0.84	.29	0.85	0.94	.09
WDFQu	.61	0.72	0.11	0.94	0.96	.03	0.61	0.91	.31	0.88	0.98	.10
WDFQ–	.70	0.89	0.19	0.96	0.98	.02	0.57	0.85	.28	0.82	0.94	.12
WDFQNone	.63	0.76	0.13	1.00	1.00	.00	0.70	0.85	.15	0.89	0.94	.06
Ma	.89	0.93	0.04	0.96	0.96	.00	0.67	0.80	.14	0.90	0.96	.07
Mp	.69	0.75	0.06	0.97	0.99	.02	0.62	0.70	.08	0.87	0.91	.03
M	.90	0.93	0.03	0.98	0.98	.01	0.61	0.82	.21	0.87	0.96	.09
FM	.92	0.95	0.03	0.94	0.94	–.01	0.66	0.83	.17	0.89	0.95	.07
m	.80	0.83	0.03	0.90	0.92	.02	0.65	0.82	.17	0.89	0.96	.07
FC	.59	0.53	–0.06	0.88	0.93	.05	0.69	0.82	.13	0.91	0.96	.05
CF	.62	0.64	0.02	0.87	0.92	.04	0.71	0.94	.23	0.89	0.97	.08
C	.58	0.64	0.06	0.97	0.98	.01	0.69	0.84	.15	0.84	0.92	.08
CF + C + Cn	.74	0.77	0.02	0.92	0.97	.05	0.70	0.95	.24	0.88	0.97	.09
Sum C	.89	0.91	0.02	0.97	0.99	.02	0.67	0.93	.26	0.88	0.97	.09
FC'	.67	0.79	0.12	0.92	0.95	.03	0.65	0.82	.17	0.86	0.92	.06
C'F	.39	0.34	–0.05	0.83	0.88	.06	0.79	0.89	.10	0.96	0.98	.02
C'	.16	0.20	0.04	1.00	1.00	.00	0.78	0.74	–.04	0.78	0.74	–.04
Sum C'	.76	0.88	0.12	0.92	0.94	.01	0.69	0.86	.17	0.88	0.96	.08
FT	.61	0.51	–0.10	0.86	0.90	.04	0.58	0.76	.18	0.90	0.94	.04
TF	.45	0.51	0.06	0.67	0.52	–.14	0.70	0.63	–.07	0.75	0.68	–.06
Sum T	.76	0.85	0.09	0.84	0.89	.04	0.61	0.70	.08	0.85	0.88	.03
FV	.44	0.52	0.08	0.89	0.91	.02	0.74	0.89	.14	0.90	0.96	.06
VF	.61	0.82	0.21	0.67	0.66	.00	0.79	0.92	.14	0.97	0.99	.01
Sum V	.58	0.68	0.09	0.87	0.90	.03	0.74	0.96	.21	0.93	0.98	.05

FY	.51	0.78	0.27	0.93	0.97	.05	0.65	0.83	.19	0.88	0.95	.07
YF	.46	0.60	0.14	0.91	0.93	.02	0.74	0.85	.11	0.89	0.93	.05
Y	.60	0.47	–0.13	0.89	0.92	.03	0.86	0.88	.03	1.00	1.00	.00
Sum Y	.61	0.79	0.17	0.92	0.97	.05	0.68	0.89	.22	0.88	0.96	.08
Fr	.88	0.94	0.06	0.96	0.98	.02	0.63	0.53	–.10	0.90	0.91	.01
rF	.00	0.49	0.49	0.87	0.94	.07	0.56	0.75	.20	0.78	0.88	.10
Fr + rF	.94	0.96	0.03	0.97	0.99	.02	0.60	0.71	.11	0.86	0.93	.07
FD	.71	0.76	0.04	0.90	0.93	.03	0.77	0.85	.09	0.90	0.95	.05
F	.87	0.97	0.10	0.96	0.99	.03	0.51	0.77	.26	0.85	0.95	.11
Blends	.79	0.89	0.10	0.91	0.98	.07	0.64	0.92	.28	0.87	0.97	.11
C – Shd Blend	.74	0.83	0.10	0.92	0.94	.02	0.76	0.89	.13	0.89	0.96	.06
Pairs	.83	0.94	0.10	0.98	1.00	.01	0.52	0.80	.27	0.85	0.95	.10
H	.88	0.92	0.04	0.98	0.99	.01	0.63	0.73	.10	0.88	0.92	.04
(H)	.81	0.87	0.06	0.96	0.97	.01	0.65	0.83	.18	0.87	0.95	.08
Hd	.75	0.74	–0.01	0.94	0.97	.04	0.69	0.88	.19	0.89	0.97	.08
(Hd)	.64	0.67	0.03	0.92	0.96	.04	0.67	0.75	.09	0.90	0.93	.03
Hx	.26	0.40	0.13	0.97	0.98	.02	0.62	0.77	.16	0.85	0.92	.07
A	.92	0.96	0.04	0.98	0.99	.01	0.56	0.88	.32	0.84	0.96	.12
(A)	.71	0.78	0.08	0.92	0.96	.04	0.60	0.66	.06	0.86	0.90	.04
Ad	.81	0.90	0.09	0.97	0.99	.02	0.62	0.78	.16	0.90	0.95	.05
(Ad)	.44	0.44	0.00	0.86	0.85	–.01	0.81	0.87	.05	0.97	0.98	.01
An	.81	0.61	–0.20	0.98	0.99	.01	0.76	0.84	.08	0.90	0.95	.05
Art	.73	0.86	0.14	0.94	0.94	.00	0.66	0.86	.20	0.91	0.98	.07
Ay	.72	0.70	–0.02	0.95	0.97	.02	0.68	0.69	.02	0.88	0.90	.03
Bl	.89	0.92	0.02	0.99	0.99	.01	0.82	0.90	.08	0.89	0.93	.04
Bt	.90	0.93	0.03	0.97	0.99	.01	0.58	0.76	.18	0.88	0.94	.06
Cg	.85	0.89	0.04	0.92	0.95	.04	0.62	0.79	.17	0.88	0.95	.07
Cl	.91	0.90	–0.01	0.75	0.78	.04	0.80	0.84	.04	0.80	0.84	.04
Ex	.70	0.69	–0.01	1.00	1.00	.00	0.72	0.79	.07	0.97	0.98	.01
Food	.88	0.90	0.01	0.95	0.97	.02	0.71	0.83	.12	0.94	0.96	.02
Fi	.91	0.88	–0.03	0.91	0.93	.02	0.74	0.87	.14	0.91	0.96	.06
Geog	.82	0.80	–0.02	1.00	1.00	.00	0.86	0.84	–.02	0.92	0.91	–.01
Hh	.73	0.64	–0.09	0.95	0.96	.01	0.64	0.74	.10	0.86	0.92	.05

(continued)

TABLE 7 (Continued)

	Student			*Experienced*			*30% Error*			*10% Error*		
	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*	κ	*ICC*	*Raw Difference*
Ls	.68	0.72	0.04	0.94	0.95	.02	0.70	0.87	.17	0.89	0.94	.05
Na	.75	0.86	0.11	0.88	0.88	.00	0.64	0.72	.07	0.90	0.94	.04
Sc	.78	0.85	0.07	0.96	0.97	.01	0.67	0.78	.12	0.90	0.94	.04
Sx	.74	0.89	0.15	0.94	0.96	.02	0.69	0.82	.12	0.86	0.95	.09
Xy	.93	0.95	0.02	1.00	1.00	.00	0.83	0.87	.03	1.00	1.00	.00
Idio	.50	0.66	0.16	0.96	0.97	.02	0.48	0.51	.03	0.77	0.82	.05
Popular	.84	0.80	–0.04	0.98	0.98	.01	0.56	0.59	.03	0.84	0.80	–.04
Zf	.86	0.95	0.08	0.96	0.99	.03	0.55	0.89	.34	0.85	0.98	.13
DV1	.58	0.72	0.13	0.81	0.85	.04	0.63	0.58	–.04	0.89	0.93	.04
INC1	.35	0.61	0.26	0.73	0.76	.03	0.65	0.85	.21	0.89	0.93	.05
DR1	.57	0.82	0.26	0.83	0.90	.07	0.67	0.91	.24	0.91	0.98	.07
FAB1	.53	0.62	0.09	0.83	0.91	.07	0.73	0.87	.15	0.90	0.97	.06
ALOG	.58	0.65	0.07	0.66	0.86	.19	0.69	0.86	.16	0.88	0.94	.07
CONTAM	.00	1.00	1.00	0.67	0.65	–.02	1.00	1.00	.00	1.00	1.00	.00
DV2	.61	0.84	0.23	0.67	0.49	–.18	0.63	0.68	.06	0.78	0.81	.04
INC2	.57	0.84	0.28	0.86	0.86	.00	0.62	0.76	.13	0.86	0.93	.07
DR2	.25	0.28	0.04	1.00	1.00	.00	0.74	0.90	.16	0.90	0.98	.08
FAB2	.53	0.72	0.18	0.94	0.93	–.01	0.67	0.70	.04	0.90	0.93	.04
AB	.64	0.67	0.03	0.97	0.98	.01	0.36	0.28	–.08	0.80	0.88	.08
AG	.78	0.83	0.05	0.85	0.87	.02	0.76	0.89	.13	0.92	0.97	.05
COP	.79	0.77	–0.02	0.88	0.88	–.01	0.66	0.76	.10	0.86	0.91	.05
GHR	.83	0.81	–0.02	0.94	0.94	.01	0.62	0.74	.12	0.88	0.91	.04
PHR	.76	0.78	0.02	0.91	0.97	.06	0.62	0.89	.27	0.86	0.97	.11
MOR	.81	0.89	0.07	0.90	0.95	.05	0.72	0.90	.18	0.88	0.95	.07
PER	.83	0.94	0.12	0.90	0.94	.04	0.67	0.91	.24	0.88	0.98	.10
PSV	.80	0.87	0.08	0.81	0.82	.01	0.73	0.67	–.07	0.95	0.95	.00

Note. ICC = intraclass correlation coefficient.

sify all responses on these dimensions, they were much more able to agree on which protocols contained many or few scores for disrupted thought processes. As an extreme example, for a protocol one rater could have assigned the scores DR1, FAB2, DR2, and FAB1 to responses 3, 5, 7, and 9, respectively. A second rater could assign DR2, FAB1, DR1, and FAB2 to the same responses. The raters would have no response-level agreement for these scores although they would have perfect summary score agreement.

Table 7 also reveals occasions in which the data do not fit the expected pattern. For instance, FT, Y, and An in the student coder sample and TF and DV2 in the experienced rater sample all have ICC values that are noticeably lower than their corresponding κ values. Some of these unexpected findings reflect very minor differences (e.g., Hd or Cl in the student sample). However, the other instances illustrate an important point about Rorschach scoring. All scoring errors are not random events. Rather, coders may have a peculiar interpretation of some coding rules or they may have lapses in which they consistently fail to attend to a certain variable. Both of these problems can result in rater-specific systematic scoring errors rather than random scoring errors. Consider FT, Y, and An in the student rater sample and TF and DV2 in the experienced rater sample. For each score, a single protocol was an outlier for the ICC calculations. In each instance, one scorer did not code the variable (i.e., assigned a score of 0), whereas the other coded it more frequently (assigning scores of 3, 2, 8, 2, and 2, respectively). When the single outlier protocol was removed for each variable, the summary score ICC value then equaled or exceeded the response-level kappa value, as would be expected by psychometric theory.

The most dramatic ICC versus κ disparity occurred for the An score in the student sample. For one protocol, the target patient produced eight sexual responses. Both raters always agreed that sex (Sx) was the primary content category for these responses. However, one rater consistently coded human detail (Hd) as the secondary content for these responses, whereas the other rater always coded anatomy (An) as the secondary content. With this outlier protocol removed, the ICC jumped from .61 to .86. However, even smaller disparities can have a substantial impact on the ICC values. This is particularly true when the score under consideration is rare. For instance, in the experienced coder sample, both raters agreed that one protocol should have a DV score on two of the responses. However, Rater 1 assigned a DV1 to both responses, whereas Rater 2 assigned a DV2 to both responses. The net result was that the DV2 score for this protocol was 0 for Rater 1 and 2 for Rater 2. This is a small difference. However, DV2 scores are quite rare and this became the largest discrepancy in the whole sample of 65 protocols. With this one case removed, the ICC value for DV2 scores jumped from .49 to 1.00.

Overall, although aggregating scores across responses allows random errors of measurement to cancel out, these data highlight a critical principle. Not all CS scoring disparities are due to random error. At times, systematic scoring error may

be present on the part of one rater. This can occur when a rater consistently neglects a score, misunderstands the scoring rules, is consistently conservative in assigning scores, or is consistently liberal when assigning scores. When these kinds of processes are present in some test protocols, summary scores will not necessarily be more reliable than response-level scores. Barring this, however, the data firmly support classical reliability theory and indicate CS summary scores are more reliable than the scores assigned to individual responses.

Finally, we should make note of the κ values for the random scoring error samples in Tables 6 and 7. Had we simply replaced all of the original scores in these samples with randomly generated scores (i.e., allowing existing scores to be replaced with the same "new" score), the theoretical κ value for each score would have been exactly .90 in the 10% error sample, .80 in the 20% error sample, and .70 in the 30% error sample (in the absence of any sampling error). However, because we never replaced existing scores with the same new score, we actually introduced somewhat more error into each altered response. This resulted in κ values that theoretically should have been less than the nominal values of .90, .80, and .70 for every score (again assuming no sampling error). The results clearly supported this expectation. Table 6 indicates that the mean κ for each sample was slightly lower than the nominal value. In the 10% error sample, the mean κ was .88 rather than .90; in the 20% error sample, the mean κ was .77 rather than .80; and in the 30% error sample, the mean κ was .66 rather than .70. Furthermore, because common scores were replaced with random alternatives more often than rare scores and because common scores were also inserted into the protocols more often as random errors, in these samples κ tended to be lower than the nominal value for the more common scores (e.g., D, DQ+, F, 2). Specifically, the correlation between score base rate and the observed kappa values was –.30, –.55, and –.61 in the 10% error, 20% error, and 30% error samples, respectively. Again, this pattern simply reflects the strategy we used to alter existing scores in these data sets.

Reliability of Response Segments

The next set of analyses determined whether the estimation formula for κ used in Meyer's (1997a) meta-analysis produced accurate results in our five clinical samples. Meyer (1999) subsequently provided more accurate procedures to estimate κ for response segments and Janson and Olsson (2001) recently presented even more sophisticated and appropriate reliability procedures for multivariate observations like these. Nonetheless, at present we evaluate the old estimation procedures simply to determine whether the previous meta-analysis generated accurate results.

For each sample, Table 8 presents three statistics: (a) the percent of exact agreement on all scores assigned within a response segment, (b) exact κ values for the seg-

TABLE 8
Percentage Agreement, Actual Kappa, and Estimated Kappa Values for the Reliability of Comprehensive System Response Segments in the Clinical Samples

Segment (No. of Scores)	*Sample 1 Students*[a]			*Sample 2 Experienced*[b]			*Sample 3 (k = 2) Clinician*[c]			*Sample 4 Applied*[d]			*Sample 5 Composite*[e]		
	%A	κ	*Estimated* κ	*%A*	κ	*Estimated* κ	*%A*	κ	*Estimated* κ	*%A*	κ	*Estimated* κ	*%A*	κ	*Estimated* κ
Location & Space (4)	.90	.86	.86	.98	.98	.98	1.00	1.00	1.00	.97	.96	.96	.96	.94	.94
Develop Quality (4)	.92	.83	.86	.96	.92	.93	0.98	0.97	0.97	.99	.97	.97	.96	.92	.93
Determinants (28[f])	.73	.67	.65	.91	.89	.89	0.89	0.88	0.87	.83	.80	.80	.83	.80	.80
Form Quality (5)	.82	.72	.70	.97	.96	.96	0.90	0.83	0.85	.97	.95	.95	.92	.88	.88
Pairs (1)	.93	.83	.82	.99	.98	.98	1.00	0.99	0.99	.98	.96	.96	.97	.93	.93
Contents (27)	.76	.74	.70	.93	.93	.92	0.97	0.96	0.96	.93	.92	.91	.88	.87	.86
Populars (1)	.95	.84	.87	.99	.98	.98	0.99	0.99	0.98	.99	.97	.97	.98	.93	.93
Cog Sp Sc (10)	.89	.58	.63	.94	.82	.89	0.90	0.79	0.80	.89	.75	.77	.90	.74	.80
Other Sp Sc (10)	.86	.78	—	.93	.90	—	0.93	0.91	—	.90	.86	—	.90	.85	—
Other Sp Sc–Old (8)	.91	.77	.76	.95	.88	.88	0.95	0.89	0.88	.93	.83	.84	.93	.83	.84
All Sp Sc (20)	.78	.71	—	.89	.86	—	0.86	0.83	—	.82	.79	—	.83	.79	—
All Sp Sc–Old (18)	.82	.67	.68	.90	.84	.86	0.86	0.80	0.80	.84	.75	.77	.85	.76	.79
M	.86	.75	.75	.95	.92	.93	0.94	0.91	0.91	.93	.89	.89	.92	.86	.87
M difference:															
Actual κ – estimated κ			.0011			–.0095			–.0003			–.0037			–.0097

Note. Values were rounded to two decimal places after all calculations were completed. The mean and mean difference were computed from just the segments with estimated κ values. %A = percentage agreement; κ = kappa calculated directly from the sample data; estimated κ = kappa estimated using the formula and chance agreement rates presented in Meyer (1997a, Table 1); Develop = Developmental; Cog = Cognitive; Sp Sc = Special Scores; Old = segment does not include GHR and PHR.

[a]Responses = 1,407. [b]Responses = 1,299. [c]Responses = 388. [d]Responses = 1,667. [e]Responses = 4,761. [f]Pure Form was considered the default option when no other determinants were scored rather than as a code that could be assigned independently and in conjunction with all the other determinants.

ment calculated from within each sample, and (c) κ values that were estimated using the formula and chance agreement rates proposed by Meyer (1997a). For Sample 3, we again used data randomly selected from two of the five raters.

To estimate κ for each response segment, chance agreement (CA) rates were obtained from the five types of samples reported in Table 1 of Meyer (1997a). Because CA rates are determined by the base rate for all score options in a response segment and because base rates and CA rates both change as a function of psychopathology, it is essential to know the sample under consideration to accurately employ those CA rates. The outpatient CA rates were used for Sample 1. For Samples 2, 3, 4, and 5, the CA rates corresponding to the mixed psychiatric (inpatient and outpatient) population were used. Superficially, Sample 3 was an outpatient sample. However, because each patient in the sample carried a personality disorder diagnosis and because only complex records were selected for the scoring study, the frequency of otherwise rare scores would be substantially increased in the sample. Thus, it would have been inappropriate to estimate κ with standard outpatient base rates. Doing so would have led to artificially low estimates of κ.[9]

Table 8 contains data on 12 response segments. The Special Score segments deserve comment. Although the cognitive Special Scores that contribute to the Sum6 and WSum6 did not change, the other Special Score segment did. Currently, good and poor human representations (GHR and PHR, respectively) are coded as Special Scores. These variables were not part of the CS when Meyer (1997a) conducted his meta-analysis. Thus, Table 8 reports two sets of results for the Other Special Scores and All Special Scores segments. The first corresponds to current scoring practices and the second (designated as Old) corresponds to the scoring practices that were in place when Meyer (1997a) computed the chance agreement rates to estimate κ for response segments.

Perusing Table 8, it can be seen that the estimation formula produced segment κ values that were quite consistent with the observed values. As indicated in the Mean Difference row, when averaged across response segments, the κ estimates derived from the formula matched the observed values almost perfectly, with deviations of .0011, –.0095, –.0003, –.0037, and –.0097 in Samples 1 through 5, respectively. Across all 50 calculations, the average difference between the estimated and actual κ values was –.0044, a trivial magnitude. Given this and given that no deviation was ever greater than .06, it can be seen that the formula Meyer (1997a) used in his meta-analysis provided an accurate means for predicting response segment κ from the percent agreement rates in the published literature.

[9]Specifically, if the outpatient base rates are used, the estimated κ values are too low by a magnitude of .021 on average.

Results From This Study Compared to the Published Literature on CS Reliability

Finally, we compare the results from our investigations to the sample weighted results obtained from all the other published studies we know of that have reported chance-corrected interrater reliability coefficients for CS scores. These studies were identified by consulting articles that reviewed prior research on CS reliability (e.g., Acklin et al., 2000; McDowell & Acklin, 1996; Meyer, 1997a; Viglione & Hilsenroth, 2001) and by conducting a thorough PsycINFO database search that covered the recent literature. The latter identified all articles with the word *Rorschach* in the title or abstract published in the 4-year period from 1997 through December, 2000. The 247 abstracts identified through this search were reviewed and winnowed by excluding those in a language other than English, reviews and nonempirical articles, and those that clearly addressed scoring systems other than the CS. The 70 articles that remained were then manually inspected to see if they provided chance-corrected (i.e., κ or ICC) reliability coefficients for CS scores.[10]

Table 9 presents the relevant information for summary scores, individual scores at the response level, and response segments. For our study, the average summary score was obtained from Table 2, the average coefficient for individual scores at the response level was obtained from Table 6 (using the composite clinical sample), and the average for response segments was obtained from Table 8 (again using the composite clinical sample).

---

[10]A study by Grønnerød (1999) initially appeared relevant. It reported reliability at the summary score level and at the response level for single scores and segments. Even though scores were labeled using CS terms, a Klopfer-based Norwegian system was actually used to score the protocols (p. 117). Because scoring did not follow standard CS guidelings (i.e., definitions, examples, practice items), the data were not included in our summary of the literature. However, Grønnerød's study has been cited as evidence of CS reliability (e.g., Lilienfeld et al., 2000), so several other complications should be noted. Grønnerød reported results for 10 Klopfer-based scores that were said to be defined in the same way as CS scores (i.e., W, C'F, FV, rF, A, (Hd), Sex, Hx, Cg, and INC). However, it was not clear why these specific scores were equated with the CS, while many similar scores (e.g., FC', SumC', SumV, Fr) were not and did not have reliability results reported. It also was not clear why response segment data was presented for some categories that apparently did not have overlapping definitions with the CS (e.g., Form Quality). In addition, Grønnerød computed response segment reliability by averaging results across the individual scores in a segment. This procedure is not commensurate with other studies in our review. Finally, summary scores were treated as nominal categories rather than as continuous dimensions and reliability was computed using unweighted κ rather than the ICC. As a consequence, a minor disparity between two raters (e.g., W = 9 vs. W = 8) would be treated as an error that was as severe as a major disparity (e.g., W = 9 vs. W = 3 or W = 16 vs. W = 2). This approach is not typical and should produce lower estimates of reliability than the standard procedures used elsewhere in the literature. Given these computational differences, at most our review could have included response level κ results for the 10 Klopfer-based scores that were said to be comparable to CS scores. The four scores with a base rate > .05 had *M* κ = .87; the six scores with a base rate < .05 had *M* κ = .41. Had we included all 10 scores in Table 9, the overall *M* κ for individual scores would be .80 (*N* = 15,139) rather than .83 (*N* = 13,159).

TABLE 9
Comprehensive System Reliability in the Current Samples Relative to the Published Literature

	This Study			*Published Literature*			*All Available Data*		
	No. Var	*N*	*M κ/ICC*	*No. Var*	*N*[a]	*κ/ICC*	*No. Var*	*N*[a]	*M κ/ICC*
Summary scores	164	219	.90	2–85	455	.93	2–164	674	.92
Individual scores	108	4,761	.84	2–88	8,398	.83	2–108	13,159	.83
Response segments	10	4,761	.86	4–10	7,247[b]	.85	4– 10	12,008[c]	.86

Note. No. Var = number of scores or segments considered in each study; ICC = intraclass correlation coefficient.

[a]Studies reporting pairwise agreement rates or multirater coefficients were treated as containing replicated samples across the rater pairs. As such, the *N* for these studies was calculated as the number of target protocols or responses coded times the number of unique observer pairs. [b]This is the average number of responses coded across studies (range from 1,400 to 14,003 responses per segment; *Mdn* = 6,221). [c] This is the average number of responses coded across studies (range from 6,161 to 18,764 responses per segment; *Mdn* = 10,982).

From the published literature, summary score coefficients were obtained from Acklin et al.'s (2000) clinical and nonpatient samples; Franklin and Cornell (1997); Greco and Cornell (1992); Netter and Viglione (1994); Ornduff, Centeno, and Kelsey (1999); and Perry and Viglione (1991). Results from Meyer et al. (2000) were not used because their reliability protocols were already part of our study. From Acklin et al. we only used the statistically stable scores. Greco and Cornell just reported that all ICC values were > .85. Conservatively, the value of .86 was used as the mean for this study. Netter and Viglione reported that one score had an ICC of .90, whereas the remaining six scores had ICC values > .95. Conservatively, the latter were assumed to average .96. For Ornduff et al. it was unclear whether intraclass correlations were reported for two raters or four raters. Given the ambiguity, we conservatively assumed just two raters.

For individual scores at the response level, reliability coefficients were obtained from Acklin et al. (2000); Baity and Hilsenroth (1999); Hilsenroth, Fowler, and Padawer (1998); Krishnamurthy, Archer, and House (1996; Archer & Krishnamurthy, 1997); Perry and Braff (1994); Perry, Potterat, Auslander, Kaplan, and Jeste (1996); Shaffer, Erdberg, and Haroian (1999); and Young, Justice, and Erdberg (1999). Krishnamurthy et al. reported the mean κ value for individual scores that fell within response segments. Perry and Braff just reported that κ values were between .88 and .97, so the mean of these two values was used in calculations. The total number of responses (R) was estimated by assuming 20 responses per protocol. Perry et al. (1996) reported that κ values were between .74 and .82, so the mean of these two values was used in calculations. Total R was estimated from their Table 3. Young et al. reported that κ values were between .75 and

1.0. The mean of these two values was used in calculations and total R was estimated by assuming 20 responses per protocol.

For response segments, reliability coefficients were obtained from McDowell and Acklin (1996); the 16 samples summarized in Meyer (1997a); Perry, McDougall, and Viglione (1995); and Perry, Sprock, et al. (1995).[11] Perry, McDougall, et al. (1995) reported that κ ranged from .63 to .89, so the mean of these values was used in calculations. Total R was estimated by assuming 20 responses per protocol. Perry, Sprock, et al. (1995) reported that κ ranged from .71 to .82. The mean of these two values was used in calculations and total R was estimated from their Table 1.

Although this investigation examined more scores than other studies in the published literature, our average results are essentially equivalent to those observed by other investigators. The average summary score ICC from this study closely matched the weighted average derived from seven other CS samples in the published literature (.90 vs. .93). Similarly, our response-level reliability for individual scores was almost identical to the average κ value reported in nine other samples (.84 vs. .83). Finally, our average response segment κ value was virtually the same as the weighted average derived from up to 19 other samples reported in the literature (.86 vs. .85).[12]

DISCUSSION

In this study, we presented detailed findings on the interrater reliability of the Rorschach CS in eight large samples. Five of these samples were based on clinical records: (a) a novice sample containing 2 independent ratings of 66 protocols (R = 1,407), (b) an expert sample containing 2 independent ratings of 65 protocols (R = 1,299), (c) a clinician survey sample containing 95 independent ratings of 19 protocols (i.e., 5 ratings per protocol; R = 388), (d) an applied reliability sample contain-

[11]It is not clear that McDowell and Acklin (1996), Perry, McDougall, et al. (1995), and Perry, Sprock, et al. (1995) computed segment chance agreement rates in the same manner as our study and Meyer's (1997a) meta-analysis. To the extent that they did not take into account the base rate for all score options within a segment, their segment κ values will underestimate reliability.

[12]Acklin et al. (2000) provided an extensive array of reliability data from two small samples. In their analyses, protocols had been scored by two students with some advanced training. As such, their results can be compared to the findings we reported for our student raters. Acklin et al.'s average ICC values were .78 and .80 across 82 to 85 summary scores. These values are slightly lower than the average of .82 we found across 133 variables in our Sample 1. Acklin et al.'s average response-level reliability coefficients were κ = .73 and .78 across 88 to 89 variables. These values are higher than the average κ = .69 across 108 variables in our Sample 1. In addition, McDowell and Acklin (1996) reported κ values for response segments in one of their samples. They found an average κ = .79. Using the same response segments, the average κ was .78 in our student sample.

ing 69 protocols initially scored as part of everyday clinical practice and then rescored by researchers (R = 1,667), and (e) a composite clinical sample containing 2 independent ratings of 219 protocols (R = 4,761) that were derived from the four previous samples. The remaining three samples examined 57 experimentally manipulated protocols (R = 1,378) in which 10%, 20%, and 30% of all scores were replaced with randomly generated erroneous scores. Across all samples, reliability coefficients were generally excellent, with median ICC coefficients of .85, .96, .97, .95, .93, .95, .89, and .82, respectively, across the 133 to 143 statistically stable summary scores in each sample.[13] Not surprisingly, the sample containing 30% random error produced the lowest reliability coefficients. However, even in this sample reliability remained quite high.

When the four distinct clinical samples were combined to form a single composite sample of 219 protocols and 4,761 responses, the median and mean interrater reliability coefficients were .92 and .90, respectively across all 164 structural summary variables that could be evaluated. No variables had poor reliability. Instead, 1 variable (TF) had fair reliability, 7 were classified as having good interrater agreement, and the remaining 95% of the variables were classified as having excellent reliability. These data indicate unequivocally that CS scoring rules are sufficiently clear and unambiguous to produce highly reliable summary scores when reasonably trained raters independently code the same responses.

Previously, Wood, Nezworski, and Stejskal (1996, 1997; also see Garb et al., 2001; Lilienfeld et al., 2000; Wood & Lilienfeld, 1999) suggested CS reliability may be poor and they asserted that proper reliability studies would demonstrate how some CS scores were reliable, whereas others were not. Unfortunately, there was never evidence to support these suppositions. Rather, all the available data had suggested Rorschach scoring was reliable (cf. Table 9 and Meyer, 1997a, 1997c). When the historical literature is considered in conjunction with the evidence assembled here, one must conclude that the assertions of poor reliability were erroneous. Because the Rorschach has held a contentious place in psychology's history, claims of poor reliability may have emerged from negative attitudes toward the Rorschach as a method of assessment rather than from an understanding of the instrument and an appreciation of the available empirical literature.

The analyses we report here demonstrate several additional points. First, response-level reliability coefficients provide more conservative estimates of CS interrater agreement than summary score coefficients. Second, across samples it was evident that practicing clinicians produced excellent interrater reliability coefficients that were equivalent to those of researchers. Third, Rorschach summary scores function like other types of psychometric scales. When scores are aggre-

[13] At the conclusion of this study, we realized we had omitted the new score that combines W and D locations. This is a statistically stable score. Across all the samples, the reliability of the WD score was .97, 1.0, 1.0, 1.0, 1.0, 1.0, .99, and .98, for Samples 1 through 8, respectively.

gated across responses, random item-level errors tend to cancel out making the summary values more reliable. This effect appears most pronounced in samples that initially contain more response-level random errors.

Some may wonder which type of coefficient is the most appropriate for CS reliability: the ICC for summary scores, κ for response segments, or κ for individual scores? We believe the answer depends on the goal of the reliability analysis. If the analysis is designed specifically to examine the extent to which two raters understand and agree on the CS scoring rules, it would be appropriate to focus on the response-level reliability of individual scores (although see Meyer, 1997a, and Table 4 and 5 for cautions when using small samples). Conversely, if the goal is to understand the applied reliability of the CS for research or practice, it would be most appropriate to focus on summary score ICCs because summary scores are generally used for statistical analyses or clinical decision making.

If researchers are interested in documenting that CS scoring is reliable simply as a precursor to investigating validity (cf. Weiner, 1991), it may be most appropriate to report the reliability for response segments. Although reliability coefficients derived from response segments do not provide differentiated information on each CS score, we recommend their use for several reasons. First, it is very time consuming to calculate intraclass correlations for 165 summary scores or to calculate κ coefficients for the 100+ score options that can be assigned to individual responses. Although computerized programs can be of assistance, programs are not yet widely available to generate all the necessary calculations from a computer-entered CS sequence of scores (although see Janson & Olsson, 2001). In this study, we relied on computerized tabulation for all analyses.[14] However, it takes many, many hours to organize the relevant data, import it into a statistical program, and write the appropriate syntax to transform and analyze the findings.

Some researchers may think it is appropriate to invest many hours computing the reliability for all CS scores, even when a study is really designed to address questions of validity for a limited subset of scores. However, there is another complication that must be considered. Tables 4 and 5 demonstrate how sampling error confounds the reliability statistics derived from small samples (e.g., $n = 20$). As a result, small reliability samples are prone to generate at least some misleading statistics for individual scores. This problem is not addressed by CIs. Table 5 and Figures 1 and 2 reveal that CIs are often grossly inaccurate and tend to underestimate true reliability, particularly for the relatively rare CS scores. Thus, true reliability is often higher than what is reported in a small sample study. In general, research-

[14]In a prior version of this manuscript, we relied on manual tabulation or data transformation for some of the analyses. Consistent with other reports (Meyer, 1998), once all the data were managed and tabulated electronically, we discovered we had initially made a number of relatively minor errors. The most serious error resulted from a misaligned tab in an output file that was subsequently read to compute ICC values. The error caused the Science content code in one sample to be computed as having a reliability of .03 rather than the correct value of .91.

ers should probably calculate κ or ICC coefficients on individual scores only when they have sufficiently large samples (e.g., 50 to 60 protocols) to ensure adequate and representative between-subjects variance across all the scores under consideration. Of course, increasing the size of a reliability sample also increases the time required to compute κ or ICC coefficients for individual scores.

One advantage of computing reliability coefficients for response segments is that they are less affected by relatively small samples. The reason is a function of chance agreement rates. Even in small clinical samples, it is quite unlikely for raters to agree by chance alone on all the Determinant scores, all the Content scores, all the Special Scores, and so forth (Meyer, 1997a). By contrast, it would not be surprising for a sample of 20 protocols to have, for example, only one C'F score, one (Ad) score, or one FAB2 score. Assuming each protocol had an average of 20 responses, these rare scores would then have a kappa-defined chance agreement rate of .995. This base rate leads to estimates of chance-corrected reliability that are statistically very unstable.

This study also demonstrated that the formula used in Meyer's (1997a) meta-analysis to estimate response segment κ values was accurate. Thus, the results from that study are stable and generalizable. However, researchers interested in generating κ for response segments should use the more precise and simple steps developed by Meyer (1999) or the even more sophisticated procedures developed by Janson and Olsson (2001).

Although this study reports important positive conclusions for CS reliability, readers should be equally clear about conclusions that are not warranted by these data. In particular, it is not the case that anyone who uses the CS is automatically a reliable scorer. Such a conclusion is patently false. The CS is a complex coding system that requires knowledge of many rules and benchmarks for accurate discrimination. These rules and benchmarks are only acquired through systematic training and practice; therefore, only well-trained individuals will score accurately. It is reassuring to know that coders have at their disposal a classification system that can be reliably implemented. However, the act of classifying Rorschach responses ultimately depends on the coder, not the scoring system. Thus, reliable use of the CS is dependent on the skill of the clinician or researcher using the instrument. Even within the highly reliable data presented here, there were instances when one coder in a pair systematically differed from the other. Consequently, even well-trained coders must maintain a vigilant stance toward proper scoring to avoid lapses and errors. We strongly encourage anyone who uses the CS in clinical practice (or any other instrument with complex scoring) to conscientiously evaluate their scoring accuracy to ensure the test is being employed in a reliable, clinically defensible manner.

ACKNOWLEDGMENTS

We thank Theresa Kiolbasa, Linda Arsenault, Rob Riethmiller, Gina Brooks, Bill Benoit, and Margot Holaday for their Rorschach scoring that contributed to this ar-

ticle. We also appreciate the input from five reviewers who commented on previous versions of this article.

REFERENCES

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,* 15–47.

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

Andreasen, N. C., Grove, W. M., Shapiro, R. W., Keller, M. B., Hirschfeld, R. M. A., & McDonald-Scott, P. (1981). Reliability of lifetime diagnosis: A multicenter collaborative perspective. *Archives of General Psychiatry, 38,* 400–405.

Archer, R. P., & Krishnamurthy, R. (1997). MMPI–A and Rorschach indices related to depression and conduct disorder: An evaluation of the incremental validity hypothesis. *Journal of Personality Assessment, 69,* 517–533.

Baity, M. R., & Hilsenroth, M. J. (1999). Rorschach aggression variables: A study of reliability and validity. *Journal of Personality Assessment, 72,* 93–110.

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

Buysse, D. J., Reynolds, C. F., III, Hauri, P. J., Roth, T., Stepanski, E. J., Thorpy, M. J., et al. (1994). Diagnostic concordance for *DSM–IV* sleep disorders: A report from the APA/NIMH *DSM–IV* field trials. *American Journal of Psychiatry, 151,* 1351–1360.

Carroll, R. M., & Faden, V. B. (1978). Some sampling characteristics of three estimators of the intraclass correlation. *Educational and Psychological Measurement, 38,* 855–863.

Cicchetti, D. V. (1985). A critique of Whitehurst's "Interrater agreement for journal manuscript reviews": De omnibus disputandem est. *American Psychologist, 40,* 563–568.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14,* 119–186.

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

Cicchetti, D. V., & Prusoff, B. A. (1983). Reliability of depression and associated clinical symptoms. *Archives of General Psychiatry, 40,* 987–990.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency, 86,* 127–137.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Exner, J. E., Jr. (1995). *A Rorschach workbook for the comprehensive system* (4th ed.). Asheville, NC: Rorschach Workshops.

Exner, J. E., Jr. (2000). *A primer for Rorschach interpretation.* Asheville, NC: Rorschach Workshops.

Exner, J. E., Jr. (2001). *A Rorschach workbook for the comprehensive system* (5th ed.). Asheville, NC: Rorschach Workshops.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30,* 71–76.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 32,* 255–265.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., Davies, M., Borus, J., et al. (1995). The Structured Clinical Interview for *DSM–III–R* Personality Disorders (SCID–II): Part 2. Multi-site test–retest reliability study. *Journal of Personality Disorders, 9,* 92–104.

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

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

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin, 86,* 974–977.

Fowler, J. C., Piers, C., Hilsenroth, M. J., Holdwick, D. J., Jr., & Padawer, J. R. (2001). The Rorschach Suicide Constellation: Assessing various degrees of lethality. *Journal of Personality Assessment, 76,* 333–351.

Franklin, K. W., & Cornell, D. G. (1997). Rorschach interpretation with high-ability adolescent females: Psychopathology or creative thinking? *Journal of Personality Assessment, 68,* 184–196.

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Towards a resolution of the Rorschach controversy. *Psychological Assessment, 13,* 433–448.

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

Greene, R. L. (2000) *The MMPI–2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Grønnerød, C. (1999). Rorschach interrater agreement estimates: An empirical evaluation. *Scandinavian Journal of Psychology, 40,* 115–120.

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

Hilsenroth, M. J., Fowler, J. C., & Padawer, J. R. (1998). The Rorschach Schizophrenia Index (SCZI): An examination of reliability, validity, and diagnostic efficiency. *Journal of Personality Assessment, 70,* 514–534.

Holaday, M. (1998). Rorschach protocols of children and adolescents with severe burns: A follow-up study. *Journal of Personality Assessment, 71,* 306–321.

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

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61,* 277–289.

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

Keller, M. B., Klein, D. N., Hirschfeld, R. M. A., Kocsis, J. H., McCullough, J. P., Miller, I., et al. (1995). Results of the *DSM–IV* mood disorders field trial. *American Journal of Psychiatry, 152,* 843–849.

Krishnamurthy, R., Archer, R. P., & House, J. J. (1996). The MMPI–A and Rorschach: A failure to establish convergent validity. *Assessment, 3,* 179–191.

Lahey, M., Downey, R., & Saal, F. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin, 93,* 586–595.

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

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Loranger, A. W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., et al. (1994). The international personality disorder examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration international pilot study of personality disorders. *Archives of General Psychiatry, 51,* 215–224.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

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

McNulty, J. L., Graham, J. R., Ben-Porath, Y., & Stein, L. A. R. (1997). Comparative validity of MMPI–2 scores of African American and Caucasian mental health center clients. *Psychological Assessment, 9,* 464–470.

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

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

Meyer, G. J. (1997c). Thinking clearly about reliability: More critical corrections regarding the Rorschach Comprehensive System. *Psychological Assessment, 9,* 495–498.

Meyer, G. J. (1998). Error in research and assessment data with an erratum for Meyer (1993). *Journal of Personality Assessment, 71,* 195–211.

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. (2002). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment, 78,* 104–129.

Meyer, G. J., Riethmiller, R. J., Brooks, G. D., Benoit, W. A., & Handler, L. (2000). A replication of Rorschach and MMPI–2 convergent validity. *Journal of Personality Assessment, 74,* 175–215.

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

Nichols, D. P. (n.d.). *ICCSF macros.* Retrieved March 26, 2002, from ftp://ftp.spss.com/pub/spss/statistics/nichols/macros/iccsf.txt

Nichols, D. P. (1998). *Choosing an intraclass correlation coefficient.* Retrieved March 26, 2002, from: ftp://ftp.spss.com/pub/spss/statistics/nichols/articles/whichicc.txt

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

Ornduff, S. R., Centeno, L., & Kelsey, R. M. (1999). Rorschach assessment of malevolence in sexually abused girls. *Journal of Personality Assessment, 73,* 100–109.

Perry, W., & Braff, D. L. (1994). Information processing deficits and thought disorder in schizophrenia. *American Journal of Psychiatry, 151,* 363–367.

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

Perry, W., Potterat, E., Auslander, L., Kaplan, E., & Jeste, D. (1996). A neuropsychological approach to the Rorschach in patients with dementia of the Alzheimer type. *Assessment, 3,* 351–363.

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

Perry, W., & Viglione, D. J., Jr. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56,* 487–501.

Selvage, R. (1976). Comments on the analysis of variance strategy for the computation of intraclass reliability. *Educational and Psychological Measurement, 36,* 605–609.

Shaffer, T. W., Erdberg, P., & Haroian, J. (1999). Current nonpatient data for the Rorschach, WAIS–R, and MMPI–2. *Journal of Personality Assessment, 73,* 305–316.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86,* 420–428.

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

Timbrook, R. E., & Graham, J. R. (1994). Ethnic differences on the MMPI? *Psychological Assessment, 6,* 212–217.

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

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fiction, and future. *Psychological Assessment, 13,* 452–471.

Volkmar, F. R., Klin, A., Siegel, B., Szatmari, P., Lord, C., Campbell, M., et al. (1994). Field trial for autistic disorder in *DSM–IV. American Journal of Psychiatry, 151,* 1361–1367.

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

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39,* 22–28.

Whitehurst, G. J. (1985). On lies, damned lies, and statistics: Measuring interrater agreement. *American Psychologist, 40,* 568–569.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., et al. (1992). The Structured Clinical Interview for *DSM–III–R* (SCID): 2. Multi-site test-retest reliability. *Archives of General Psychiatry, 49,* 630–636.

Wood, J. M., & Lilienfeld, S. O. (1999). The Rorschach Inkblot Test: A case of overstatement? *Assessment, 6,* 341–349.

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

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

World Health Organization. (1992). *ICD–10: International statistical classification of diseases and related health problems, 10th revision. Vol. 1.* Geneva, Switzerland: Author.

Young, M. H., Justice, J., & Erdberg, P. (1999). Risk factors for violent behavior among incarcerated male psychiatric patients: A multimethod approach. *Assessment, 6,* 243–258.

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

Gregory J. Meyer
Department of Psychology
University of Alaska, Anchorage
3211 Providence Drive
Anchorage, AK 99508
E-mail: afgjm@uaa.alaska.edu

Received January 30, 2001
Revised October 10, 2001