JOURNAL OF PERSONALITY ASSESSMENT, *72*(2), 230–255

ARTICLES

Simple Procedures to Estimate Chance Agreement and Kappa for the Interrater Reliability of Response Segments Using the Rorschach Comprehensive System

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage

When determining interrater reliability for scoring the Rorschach Comprehensive System (Exner, 1993), researchers often report coding agreement for response segments (i.e., Location, Developmental Quality, Determinants, etc.). Currently, however, it is difficult to calculate kappa coefficients for these segments because it is tedious to generate the chance agreement rates required for kappa computations. This study facilitated kappa calculations for response segments by developing and validating formulas to estimate chance agreement. Formulas were developed for 11 segments using 400 samples, cross-validated on 100 samples, and applied to the data from 5 reliability studies. On cross-validation, the validity of the prediction formulas ranged from .93 to 1.0 (M = .98). In the 5 reliability studies, the average difference between estimated and actual chance agreement rates was .00048 and the average difference between estimated and actual kappa values was .00011 (maximum = .0052). Thus, the regression formulas quite accurately predicted chance agreement rates and kappa coefficients for response segments.

Since Weiner (1991) published editorial guidelines for this journal, researchers have become more attentive to the interrater reliability of Rorschach scoring. Recently, the reliability of the Comprehensive System (CS; Exner, 1993) has been addressed in several reports (Erdberg & Cooper, 1998; Janson, 1998; McDowell & Acklin, 1996; Meyer, 1997a, 1997c; Meyer et al., 1999). When calculating reliabil-

ity, researchers face choices regarding the type of statistic to use and the level at which the analyses should focus. Potential statistics include: (a) percentage of exact agreement; (b) percentage of exact agreement determined only when at least one rater assigns a score (i.e., not counting agreement on the absence of a score); (c) measures of association, such as the Pearson correlation; (d) measures of "chance-corrected" agreement, such as kappa or the intraclass correlation; and (e) measures of deviation. Deviations measures are rarely discussed in psychology, although they are frequently used in medical research (cf. Bland & Altman, 1995).

Historically, Rorschach research has focused on either percentage agreement statistics or simple measures of association. These approaches can be criticized for not being sufficiently stringent and chance-corrected measures of agreement can be employed instead. Although it is conceptually appealing to correct agreement for chance, there are several ways to define chance and alternative definitions can have dramatic influences on the resulting reliability coefficients (Brennan & Prediger, 1981; Finn, 1970, 1972; Meyer, 1997a; Whitehurst, 1984). Furthermore, some chance-corrected reliability statistics are quite sensitive to a variable's distribution (Meyer, 1997a; Meyer et al., 1999) and may produce misleading results if they are not employed properly.

With respect to the level of data analysis, researchers can focus on (a) specific scores assigned to each response (e.g., agreement on *S* across all responses in a sample), (b) segments of scores assigned to each response (e.g., agreement on all location and space scores across all responses in a sample), or (c) summary scores that are aggregated across all responses in a protocol (e.g., agreement on the number of *S* responses across all the protocols in a sample). Based on the purpose of a study, Meyer et al. (1999) provided suggestions for when each level of analysis may be optimal. If the goal is to conduct a stringent analysis of a scoring system's clarity, it would be reasonable to focus on specific scores assigned to each and every response (i.e., Option a). For instance, if one wished to know the intrinsic reliability of the CS, chance-corrected reliability statistics could be calculated for the more than 90 score options that can be assigned to each response (assuming the scores are statistically stable). Alternatively, if the goal is to examine the reliability of scores as they are typically used in research or applied clinical practice, it would be appropriate to focus on the summary scores from a protocol (i.e., Option c). For instance, to examine the reliability of all CS scores as they are typically used, the researcher could generate coefficients for the 154 variables contained in a Structural Summary and evaluate them across patients (assuming the variables are statistically stable; see Meyer, 1997a; Meyer et al., 1999).

Either of the preceding options is a major undertaking that requires considerable time and many participants to be completed appropriately (Meyer et al., 1999). However, for many studies, a detailed examination of reliability is not a primary consideration. Rather, researchers often wish to document scoring reliability simply as a precursor to their main analyses that are focused on questions of Ror-

schach validity. Given this, a less intensive overview of scoring reliability may frequently be sufficient. In fact, the historical recommendation for examining CS reliability has been for researchers to report the general reliability of scores contained within the primary segments of a response (Exner, 1991; Weiner, 1991). These primary segments include: Location and Space, Developmental Quality, Determinants, Form Quality, Pairs, Content, Populars, *Z* Scores,[1] Cognitive Special Scores (e.g., *FAB1, DR2*), and Other Special Scores (e.g., *MOR, PER*). At times, the last two may be combined to form a segment of All Special Scores.

Calculating the percentage of exact agreement for these response segments is a simple and quick procedure. The researcher only needs to compare the sequence of scores from Rater 1 and Rater 2, noting each instance when any score in a response segment differs between the two raters. For instance, if Rater 1 scored a response as *Wv ma.YFo Fi, Id MOR, DR1,* and Rater 2 scored the same response as *Wv ma.Yu Fi MOR, DR2,* then there is agreement on the segments for Location, Developmental Quality, Pair (for its absence), Popular (for its absence), *Z* Frequency (for its absence), and Other Special Scores. There is disagreement on the segments for Determinants, Form Quality, Content, and Cognitive Special Scores (and the All Special Scores segment if it is used). Similar determinations are then made for all the other responses included in the reliability analyses. After all responses have been evaluated for segment agreement, the researcher simply counts the number of times there was complete agreement for each segment and then divides by the total number of responses evaluated to determine the percentage of exact agreement for each segment. (When raters differ on how many responses were contained in the sample, the average number of responses should be used as the denominator.)

Although it is easy to calculate exact agreement rates for a response segment, it is much more difficult to calculate Cohen's (1960) kappa for a segment. This is because Cohen's kappa defines chance agreement by the relative frequency (i.e., base rate) with which each rater assigns each score option. For instance, if two raters independently code 100 responses and Rater 1 assigns an *S* 27 times, then his or her base rate for assigning *S* is $p(S_1) = 27/100 = .27$, and his or her base rate for not assigning *S* is $p(\text{not-}S_1) = 73/100 = .73$. Similarly, if Rater 2 assigns *S* 30 times to

[1]For each of the 10 inkblots, distinct *Z* score values are assigned for whole responses (*ZW*), synthesized adjacent details (*ZA*), synthesized distant details (*ZD*), and synthesized white space (*ZS*). However, *Z* values are never tallied on a card-by-card basis according to these four categories, so it is currently impossible to determine the relative frequency for *ZW, ZA, ZD,* and *ZS* on each card from existing CS information. In turn, it is not possible to calculate estimates of chance agreement for specific *Z* values. Consequently, this article focuses on interrater agreement for assigning any *Z* value (i.e., *Zf* vs. not-*Zf*). Researchers wishing more precise reliability data on *Z* scores can either calculate exact kappa values from their sample information (either using the categories *ZW, ZA, ZD,* and *ZS* or categories for each possible *Z* value; i.e., 2.0, 2.5, etc.) or they can generate protocol-level intraclass correlations for *Zf* and *ZSum.*

the same 100 responses, his or her base rate for assigning *S* is p(S_2) = 30/100 = .30, and his or her relative frequency for not assigning *S* is p(not-S_2) = 70/100 = .70.

Following Cohen's definition of chance for kappa, chance agreement is determined by multiplying each rater's base rate for a score option and summing the product across all options in a score category. In this example, there are two score options in the Space category: *S* and not-*S*. Thus, chance agreement = [p(S_1) × p(S_2)] + [p(not-S_1) × p(not-S_2)], where p(i_1) indicates Rater 1's base rate for Option *i* and p(i_2) indicates Rater 2's base rate for Option *i*. Using the appropriate numbers, chance agreement = [.27 × .30] + [.73 × .70] = .081 + .511 = .592. Thus, in this sample of 100 responses, agreement is expected by chance alone on 59.2% of the responses.

With this information, kappa can be calculated. The formula is κ = (observed agreement – chance agreement)/(1 – chance agreement). If the raters had 95% agreement on *S* for these 100 responses, then κ = (.95 – .592)/(1.0 – .592) = .358/.408 = .87745. In contrast, if the raters only agreed 80% of the time, then κ = (.80 – .592)/(1.0 – .592) = .208/.408 = .5098.

Although it is relatively straightforward to calculate kappa for any particular score, the formula becomes much more complex when applied to response segments that contain many independent scores. Consider the Determinant segment. Table 1 lists the 10 mutually exclusive and exhaustive determinant categories in this segment, using the determinant *F* as the default score when no other determinants are present.[2] Using the same notation as before, Table 1 also provides the appropriate formula to calculate chance agreement for this segment.

Because a formula like this is so cumbersome, it is time consuming to use and prone to calculation errors. Thus, it would be useful to have a simple means for estimating chance agreement rates for CS response segments. I (1997a) offered sample-based estimates for this purpose. Based on severity of disturbance, the frequency of certain scores fluctuates from sample to sample. As a result, chance agreement rates also fluctuate across samples, so I provided estimates for five different types of samples. Although these estimates produce accurate results when they are employed correctly (Meyer et al., 1999), it is not always easy to decide which sample should be used to generate estimates of chance. For instance, when examining a sample of patients diagnosed with borderline personality disorder, a sample of patients with psychosomatic conditions, or a sample of outpatients diagnosed with schizophrenia, the researcher would have to guess which of the five reference samples provided the best approximation.

[2]Theoretically, it is possible to assign *F* in a blend with some other determinant. However, the instances when this is appropriate are remarkably few. Furthermore, if *F* were treated as an independent score in its own right (i.e., one that could be assigned in combination with any other score), it would lead to dramatic underestimates of chance agreement. Thus, for the purpose of estimating chance agreement, it is best to treat *F* as the default score when no other determinant scores are assigned.

TABLE 1
Score Categories and Options Within the Determinant Response Segment and the Formula to Calculate Chance Agreement for the Segment

Score Category	*Score Options*	*Chance Agreement Formula*
Human movement	*Ma*, *M*p, *Ma*-p, no *M*	$(\{[p(Ma_1) \times p(Ma_2)] + [p(Mp_1) \times p(Mp_2)] + [p(Ma\text{-}p_1) \times p(Ma\text{-}p_2)] + [p(\text{not-}M_1) \times p(\text{not-}M_2)]\} \times$
Animal movement	*FMa*, *FM*p, *FMa*-p, no *FM*	$\{[p(FMa_1) \times p(FMa_2)] + [p(FMp_1) \times p(FMp_2)] + [p(FMa\text{-}p_1) \times p(FMa\text{-}p_2)] + [p(\text{not-}FM_1) \times p(\text{not-}FM_2)]\} \times$
Inanimate movement	*ma*, *m*p, *ma*-p, no *m*	$\{[p(ma_1) \times p(ma_2)] + [p(mp_1) \times p(mp_2)] + [p(ma\text{-}p_1) \times p(ma\text{-}p_2)] + [p(\text{not-}m_1) \times p(\text{not-}m_2)]\} \times$
Color	*Cn*, *C*, *CF*, *FC*, no *C*	$\{[p(Cn_1) \times p(Cn_2)] + [p(C_1) \times p(C_2)] + [p(CF_1) \times p(CF_2)] + [p(FC_1) \times p(FC_2)] + [p(\text{not-}C_1) \times p(\text{not-}C_2)]\} \times$
Achromatic color	*C´*, *C´F*, *FC´*, no *C´*	$\{[p(C'_1) \times p(C'_2)] + [p(C'F_1) \times p(C'F_2)] + [p(FC'_1) \times p(FC'_2)] + [p(\text{not-}C'_1) \times p(\text{not-}C'_2)]\} \times$
Diffuse shading	*Y*, *YF*, *FY*, no *Y*	$\{[p(Y_1) \times p(Y_2)] + [p(YF_1) \times p(YF_2)] + [p(FY_1) \times p(FY_2)] + [p(\text{not-}Y_1) \times p(\text{not-}Y_2)]\} \times$
Texture	*T*, *TF*, *FT*, no *T*	$\{[p(T_1) \times p(T_2)] + [p(TF_1) \times p(TF_2)] + [p(FT_1) \times p(FT_2)] + [p(\text{not-}T_1) \times p(\text{not-}T_2)]\} \times$
Vista	*V*, *VF*, *FV*, no *V*	$\{[p(V_1) \times p(V_2)] + [p(VF_1) \times p(VF_2)] + [p(FV_1) \times p(FV_2)] + [p(\text{not-}V_1) \times p(\text{not-}V_2)]\} \times$
Form dimension	*FD*, no *FD*	$\{[p(FD_1) \times p(FD_2)] + [p(\text{not-}FD_1) \times p(\text{not-}FD_2)]\} \times$
Reflections	*rF*, *Fr*, no *r*	$\{[p(rF_1) \times p(rF_2)] + [p(Fr_1) \times p(Fr_2)] + [p(\text{not-}r_1) \times p(\text{not-}r_2)]\})$
If no determinants	*F*	

Note. The notation $p(i_1)$ indicates Rater 1's base rate for score option *i* and $p(i_2)$ indicates Rater 2's base rate for score option *i*.

To rectify this limitation, simple procedures for calculating segment chance agreement rates are presented. The procedures use information that should be readily available from a CS reliability study, and they can be employed with any sample of participants. Furthermore, as shown, the procedures are quite accurate.

METHOD

Four steps were taken to develop and validate chance agreement estimates. First, a large number of CS protocols were used to generate chance agreement rates. To generalize, the protocols used at this stage were selected to encompass the data likely to be found in virtually all samples and settings. Second, multiple regression was used to predict the chance agreement rates observed in Step 1 from a subset of CS scores. Third, the regression formulas were cross-validated in new samples. Finally, the cross-validated estimates of chance agreement were applied to five samples of actual reliability data and their accuracy was evaluated.

Because chance agreement rates depend on the base rate of score options in a sample, to generate chance rates that would match those in a typical reliability study, base rate variability was obtained from 400 samples of CS protocols. Each of these samples was randomly selected from two larger "populations" of CS data. The first population consisted of 443 psychiatric inpatients and outpatients (cf. Meyer, 1997b), and the second consisted of 190 nonpatients from the CS normative pool (Exner, 1993). Within each population, two sets of random draws were completed. In the first set, 10 records were randomly selected 100 times from the psychiatric population and then from the nonpatient population. In the second set, 20 records were randomly selected 100 times from each population. Thus, a total of 400 samples were produced from the two populations; 200 contained 20 records, and 200 contained 10 records. Because the populations contained individuals who ranged from quite healthy to quite impaired, repeatedly selecting a small number of protocols from within them ensured that the base rate for all score options would encompass those likely to be encountered in virtually all reliability studies. Also, it should be noted that random samples of 10 protocols were used for these analyses so that the final set of 400 samples would have considerable variability in base rates, not because this reflects a "good" number of protocols to include in a reliability study.

For each of the 400 samples, chance agreement rates were calculated for the 11 response segments. This was accomplished by using appropriate variations of the determinant formula presented in Table 1, with one modification. Rather than using base rates from two independent raters, it was assumed that each rater assigned each score option at the same relative frequency.[3] Although this will al-

[3]This procedure is also equivalent to using the average base rate across raters, which Zwick (1988) noted is actually the definition of chance agreement for Scott's (1955) π coefficient.

most never happen in practice, making this assumption has a trivial impact on realistic segment data from a CS interrater reliability study. Furthermore, to the extent that this assumption makes a difference, it leads to an overestimate of chance agreement and an underestimate of kappa (Zwick, 1988). Thus, it is a conservative assumption.

The next step entailed selecting CS variables that would accurately predict the segment chance agreement rates for each of the 400 samples. In several instances the "segment" consisted of a single dichotomous decision in which the score was either present or absent (i.e., Pair, Popular, and Z Frequency), so the choice of a predictor was obvious. For the other more complex segments, the choice of a suitable predictor was less clear, so different alternatives were tested using curve estimation regression procedures. The most optimal predictor was selected based on precision and ease of calculation. A significant consideration in this regard was my desire to have a single, relatively simple predictor variable that could be calculated from the CS data, as this allowed scattergrams to be generated that would visually convert the CS predictor to its corresponding chance agreement rate. Once the optimal predictor variable was selected, scattergrams and regression equations were developed.

To test the adequacy of each prediction equation, three validation steps were taken. First, each regression formula was cross-validated by using a new population of 320 Rorschach protocols to generate 100 randomly selected subsamples, each of which contained 20 patients. Using the same procedures as before, segment chance agreement rates were calculated for each of these 100 samples. Subsequently, estimated chance agreement rates were produced using the regression formulas. The actual and estimated chance agreement rates were then correlated to assess the validity of the regression formulas. The population of 320 protocols from which these 100 random samples were drawn came from archived records at Rorschach Workshops. The population included records from schizophrenic patients ($n = 80$), patients who ultimately committed suicide ($n = 80$), outpatients ($n = 80$), and nonpatients ($n = 80$).

The second set of analyses evaluated the extent of formula bias that may result from assuming each rater had a common base rate. As noted earlier, actual raters would almost never have equivalent base rates. To test the extent of bias introduced by this assumption, two sets of "rater samples" were created, with each sample having a distinct base rate. Specifically, the 100 subsamples used in the prior analysis were treated as the data generated by Rater 1. Each sample was then paired with a randomly selected alternative sample, which was treated as the data from Rater 2. Thus, the data set mimics 100 instances when 20 protocols were independently rated by two people, yet each rater assigned wildly different scores to each sample. To exemplify, one sample of 20 protocols had 507 total responses and this was paired with a randomly selected alternative sample that contained only 389 responses. Another sample containing 190 pure *F* responses was paired with a sample containing 99

pure *F* responses, and a sample containing 62 cognitive special scores was paired with a sample containing 137. As would be expected, CS variables were uncorrelated across these 100 pairs of subsamples. For instance, the correlations between "Rater 1" and "Rater 2" for *R*, *F*, *Sum6*, *Pair*, *Popular*, and *Zf* were –.09, –.05, –.05, –.05, .00, and –.04, respectively (all $p > .60$, $N = 100$). Clearly, by pairing each sample with a randomly selected second sample, the data are very unrealistic. However, they provide an extreme test of how well the prediction formulas work when each rater employs a very different base rate.

Finally, the regression formulas were tested in five large reliability samples. These samples were analyzed in detail for interrater agreement, with a focus on intraclass correlations calculated on summary scores (see Meyer et al., 1999). Sample 1 contained 51 protocols ($R = 1,047$) that were independently scored by two predominantly novice coders. Sample 2 contained 55 protocols ($R = 1,125$) rated by two experienced clinicians. Sample 3 was a survey sample containing 19 protocols ($R = 388$), each of which were scored five times by a total of 95 different clinicians. For the current analysis, only two randomly selected ratings were used for each protocol. Sample 4 contained 69 protocols ($R = 1,667$) that were initially scored as part of clinical practice and then rescored by researchers. Finally, Sample 5 contained 57 protocols ($R = 1,378$). The original scoring for these records was compared to experimentally manipulated scoring in which 20% of all the scores in each record had been replaced with randomly generated erroneous scores. Across these samples, the reliability of 118 to 123 statistically stable CS Structural Summary scores was excellent, with median intraclass correlations of .95, .97, .97, .95, and .89, respectively.

In this study, estimates of chance agreement for each segment in each sample were compared to the exact chance agreement rates calculated from the interrater data. In addition, because the sole purpose of chance agreement rates is to generate kappa coefficients, in each sample, the exact kappa values for each response segment were compared to kappa values derived from the estimated chance agreement rates.

RESULTS

Listed in the initial columns of Table 2 are the 11 response segments, the CS predictor variables, the regression formulas to estimate chance agreement from the predictor variables, the adjusted R^2 between the predicted and observed chance agreement rates in the 400 development samples, and the correlation between predicted and observed chance agreement rates in the 100 cross-validation samples. The validity coefficients obtained on cross-validation ranged from a low of .93 to a high of 1.0 ($M = .98$). The cross-validation sample produced 1,100 comparisons between observed and estimated chance agreement rates (i.e., 100 samples × 11 segments). The aver-

TABLE 2
Formulas to Calculate Chance Agreement for Comprehensive System Response Segments and Their Performance in the Derivation and Cross-Validation Samples and Under the Extreme Condition When Scores From One Rater Are Randomly Paired With Scores From Another

Segment	*Predictor (x)*	*Formula*	*Derivation Adusted* R^{2a}	*Cross-Validation* r^b	*Validation r When Rater and Base Rates Are Randomly Paired*[b]
Location and Space	(*Dd* + *S*)/*R*	.51 – .92(*x*) + .66(x^2)	.982	.934	.955
DQ	(*DQo* – *DQv*)/*R*	.29 + .19(*x*) + .46(x^3)	.960	.987	.970
Dets	Sum of All Non-*F Dets*/*R*	.64 – .63(*x*) + .12(x^3)	.998	1.000	.998
FQ	(*FQo* – *FQ–*)/*R*	.31 + .07(*x*) + .21(x^2) + .39(x^3)	.996	.948	.940
Pair	Pair/*R*	1 – 2(*x*) + 2(x^2)	1.000	1.000	.996
Content	Sum of all Contents/*R*	.48 – .37(*x*) + .04(x^3)	.987	.978	.979
Popular	*Pop*/*R*	1 – 2(*x*) + 2(x^2)	1.000	1.000	.994
Z Frequency	*Zf*/*R*	1 – 2(*x*) + 2(x^2)	1.000	1.000	.962
Cognitive SpcSc	*Sum6*/*R*	1 – 1.96(*x*) + 1.7(x^2) – .64(x^3)	1.000	1.000	.998
Other SpcSc	Sum of All Other SpcSc/*R*	.995 – 1.93(*x*) + 1.63(x^2) – .52(x^3)	1.000	1.000	.997
All Special Scores	Sum of All SpcSc/*R*	.98 – 1.81(*x*) + 1.38(x^2) – .41(x^3)	1.000	1.000	.997

Note. For every predictor variable (e.g., *Dd, S, DQo, R*), the quantity to be included in the equation consists of the total number of scores assigned by Rater 1 plus the total number of scores assigned by Rater 2. *DQ* = Developmental Quality; *Dets* = Determinants; *FQ* = Form Quality; SpcSc = Special Scores. The predictor for the *Det* segment consists of the total number of determinants other than *F* assigned by both raters (i.e., all *M, FM, m, C, C´*, etc.). Similarly, the predictor for the Content segment consists of the total number of content scores assigned by both raters (i.e., all *H, [H], Hd, [Hd], Hx, A*, etc.), the predictor for the Other Special Score segment consists of the total number of special scores that are not part of the *Sum6* (i.e., all *AB, AG, CFB, COP*, etc.), and the predictor for the All Special Scores segment is simply the sum of the previous two segments (i.e., *Sum6* plus the sum of all other special scores).

[a]n = 400. [b]n = 100.

age difference between observed and predicted values was .00072 and the largest absolute difference was ± .0373. Thus, the data suggest that each prediction formula allows one to calculate chance agreement rates with a high degree of accuracy.

Figures 1 through 11 provide scatterplots that correspond to the Table 2 equations. Each plot contains the line of best fit between the predictor variable and the chance agreement rate. These figures allow researchers to quickly determine approximate chance agreement rates for the 11 segments. Although the regression equations provide the most specific estimates, the precision of these figures can be enhanced by using a copy machine to enlarge each graph.

Figures 5, 7, and 8 reveal that a quadratic formula perfectly predicts the observed chance agreement rates, and Figures 9, 10, and 11 indicate that a cubic formula achieves perfect prediction. These figures are slightly misleading in that they were generated from data in which both raters were assumed to have the same base rates. Plots of realistic interrater data (in which equivalence would be rare) would reveal slight dispersion below the best fit lines. For instance, using either the *Zf* formula in Table 2 or Figure 8, it can be seen that chance agreement will be .50 when both raters assign *Zf* scores 50% of the time. Using the exact chance agreement formula presented earlier, where chance agreement = $[p(Zf_1) \times p(Zf_2)] + [p(\text{not-}Zf_1) \times p(\text{not-}Zf_2)] = [.50 \times .50] + [.50 \times .50] = .50$, we can see that this estimate is entirely accurate. However, as previously noted, when the raters do not have equivalent base rates, the formula from Table 2 (or Figure 8) slightly overestimates chance agreement. If Rater 1 assigned *Zf* to 55% of the responses in a reliability sample and Rater 2 assigned *Zf* to 45% of the responses, actual chance agreement would be $(.55 \times .45) + (.45 \times .55) = .4950$, even though the estimated value would still be .5000.

The differences become larger as the base rates for Rater 1 and Rater 2 become increasingly discrepant. For example, across the 400 samples containing 10 to 20 protocols that were used to generate Figure 8, the lowest base rate observed for *Zf* was about .425, whereas the highest rate was about .755. Although it is extremely unlikely that two raters scoring the same records in an actual reliability study would generate such marked discrepancies (and if they did, estimating chance agreement should be the least of anyone's worries), these two values can be used to examine a "worst case" scenario for the prediction equation. Assuming Rater 1 had the base rate of .425 and Rater 2 had the rate of .755, the *Zf* prediction value that should be used to estimate chance agreement is .59.[4] The value of .59 produces an estimated chance agreement rate of approximately .52 from Figure 8 and a more specific value of .5162 from the formula in Table 2. However, the true chance agreement rate would be $(.425 \times .755) + (.575 \times .245) = .46175$. Thus, the estimate of chance agreement is too large by a magnitude of .05445. As a result, kappa val-

[4]The average base rate (i.e., [.455 + .755]/2) is equivalent to the sum of *Zf* across both raters divided by the sum of *R* across both raters.

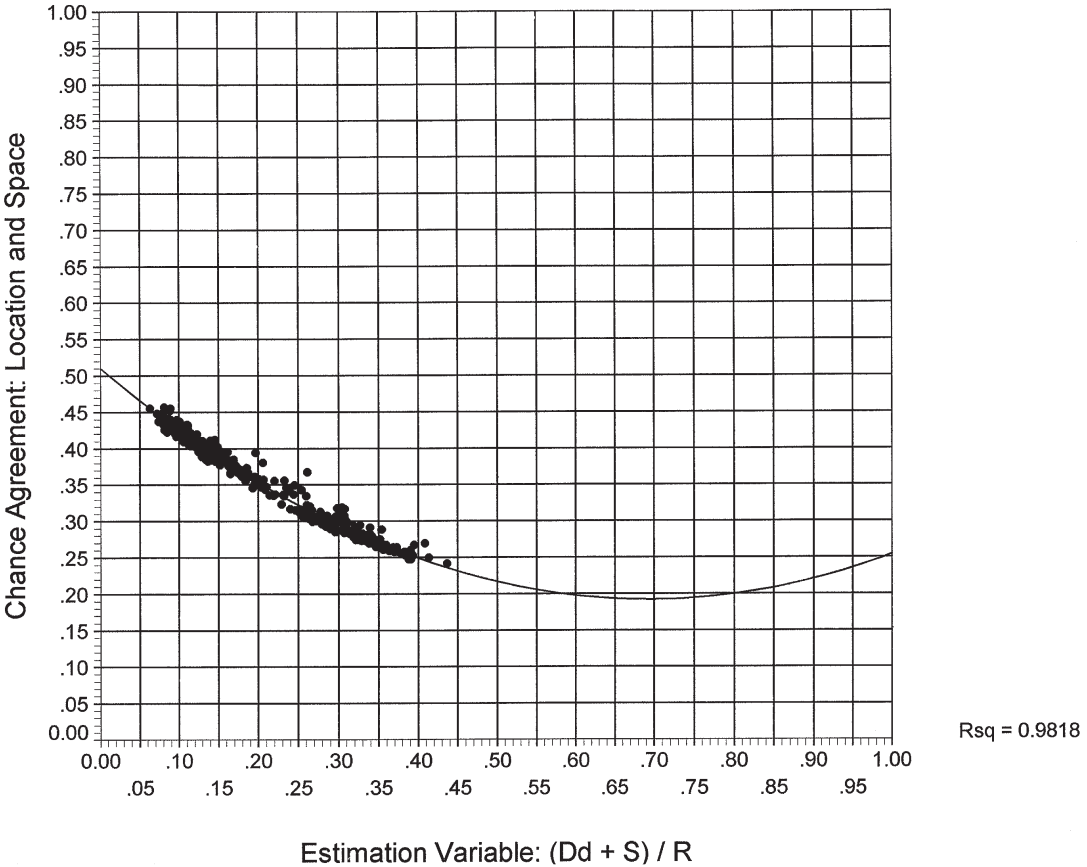

FIGURE 1 Estimates of chance agreement for the Location and Space segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

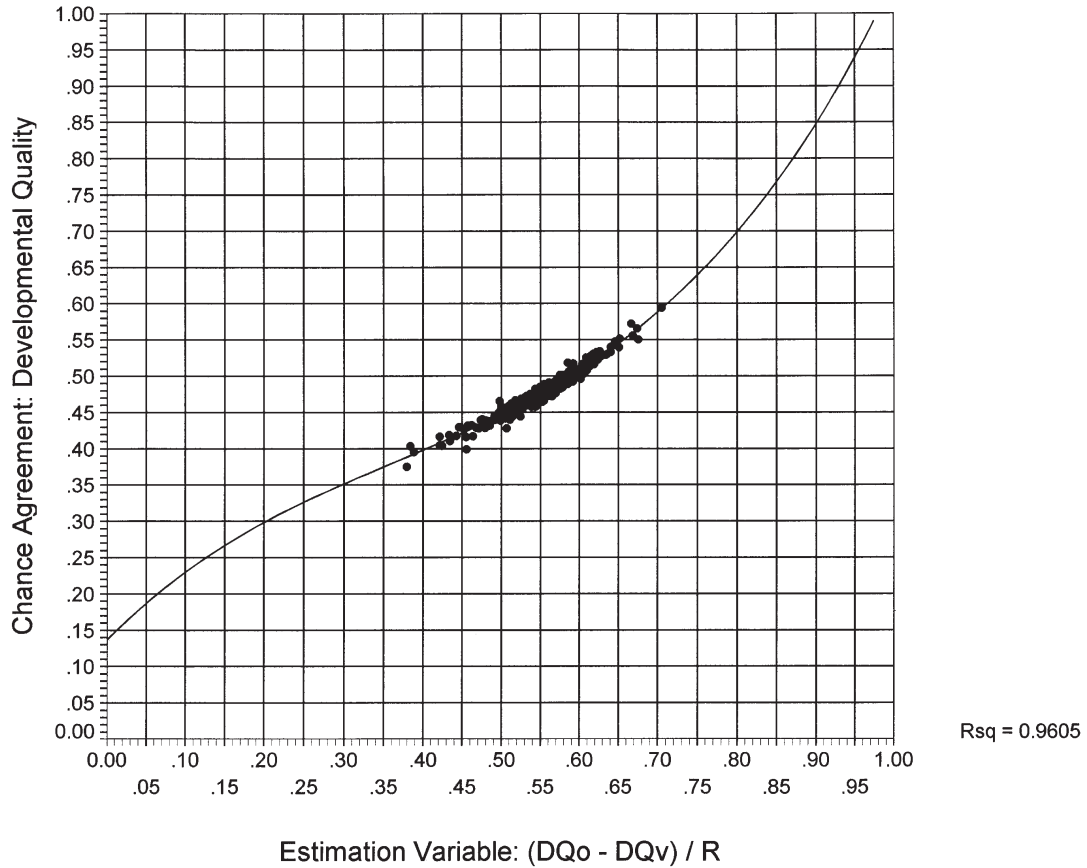

FIGURE 2 Estimates of chance agreement for the Developmental Quality segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

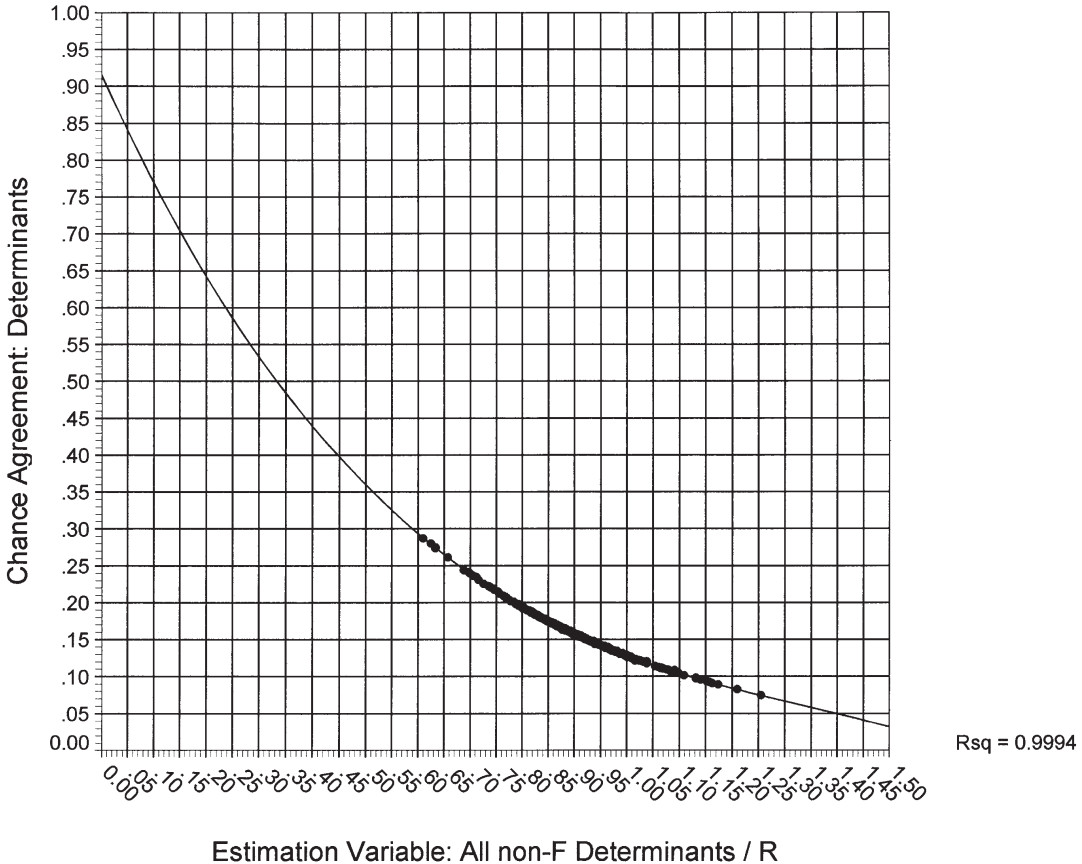

FIGURE 3 Estimates of chance agreement for the Determinant segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

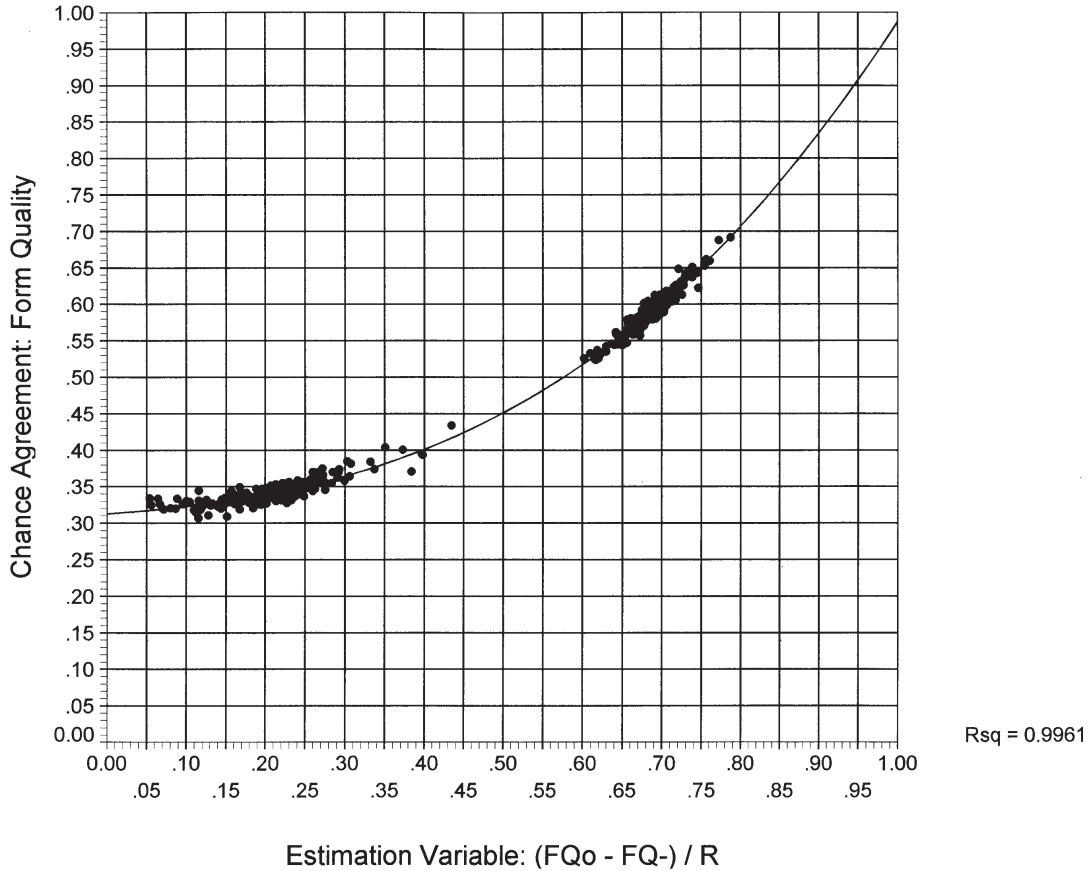

FIGURE 4 Estimates of chance agreement for the Form Quality segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

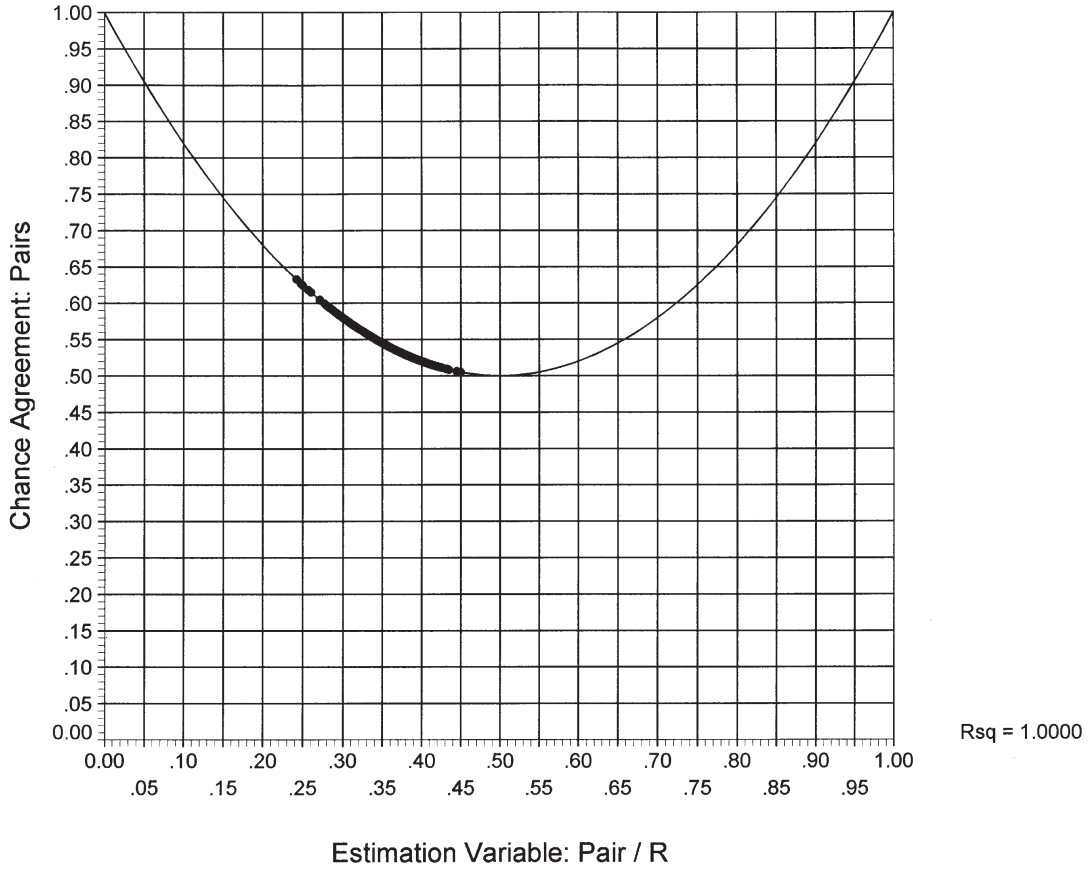

FIGURE 5 Estimates of chance agreement for the Pair segment (Comprehensive Score estimation variables refer to the sum of scores from both raters).

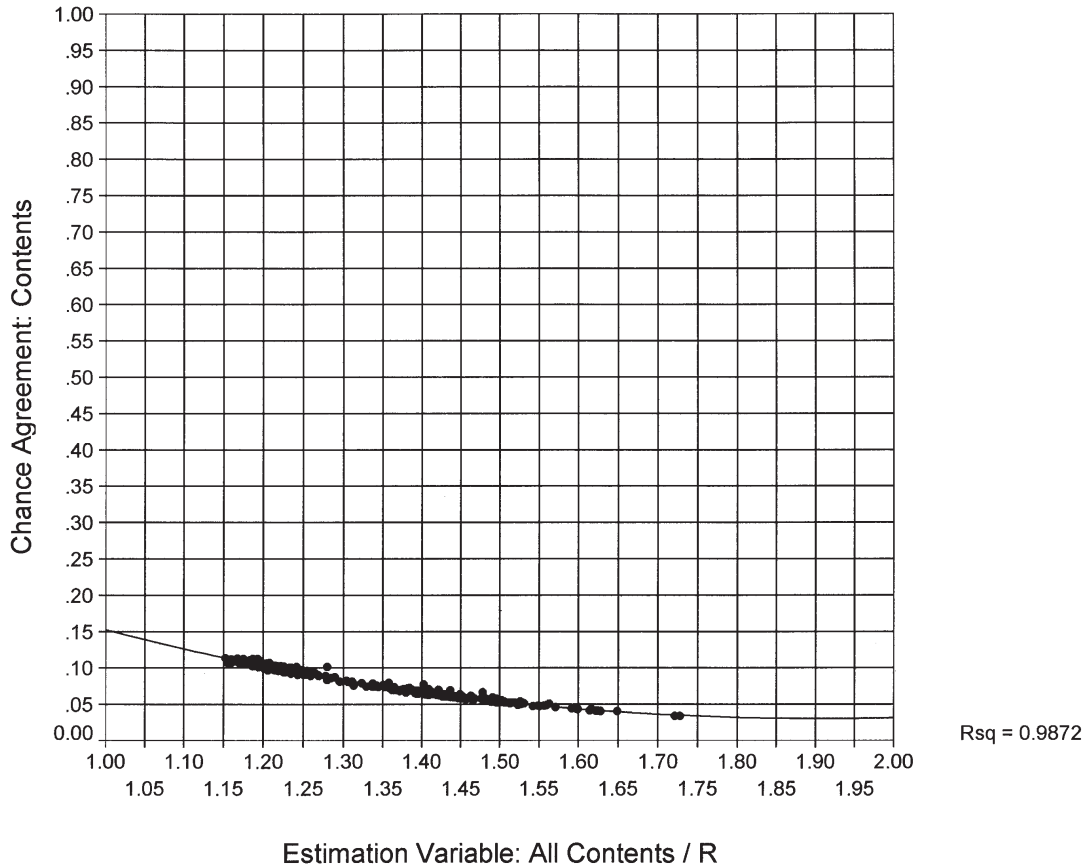

FIGURE 6 Estimates of chance agreement for the Content segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

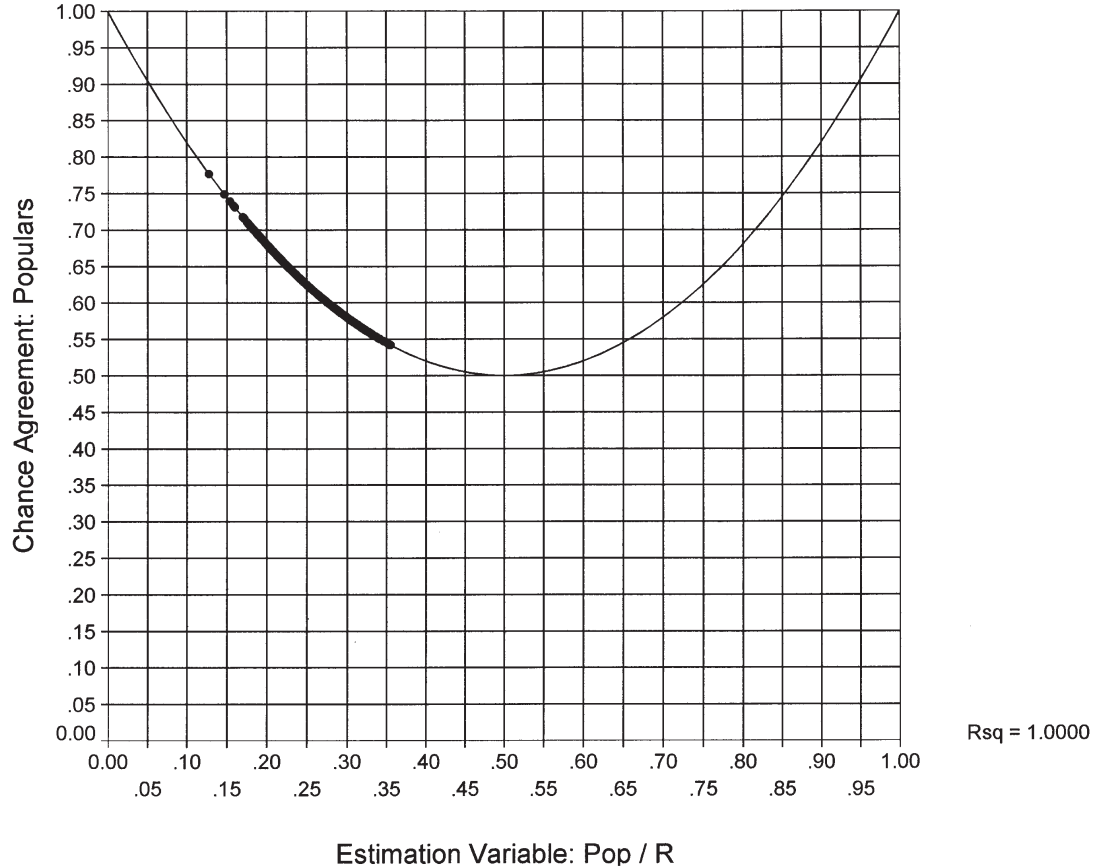

FIGURE 7 Estimates of chance agreement for the Popular segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

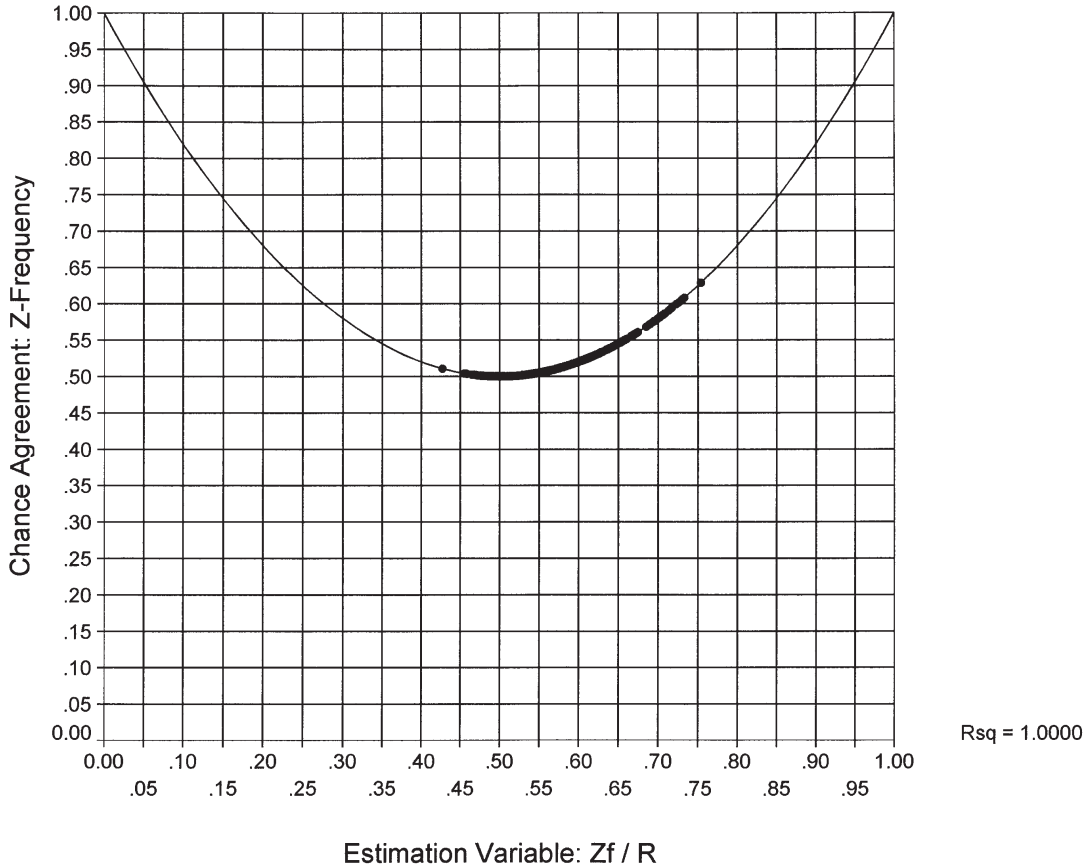

FIGURE 8 Estimates of chance agreement for the Z Frequency segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

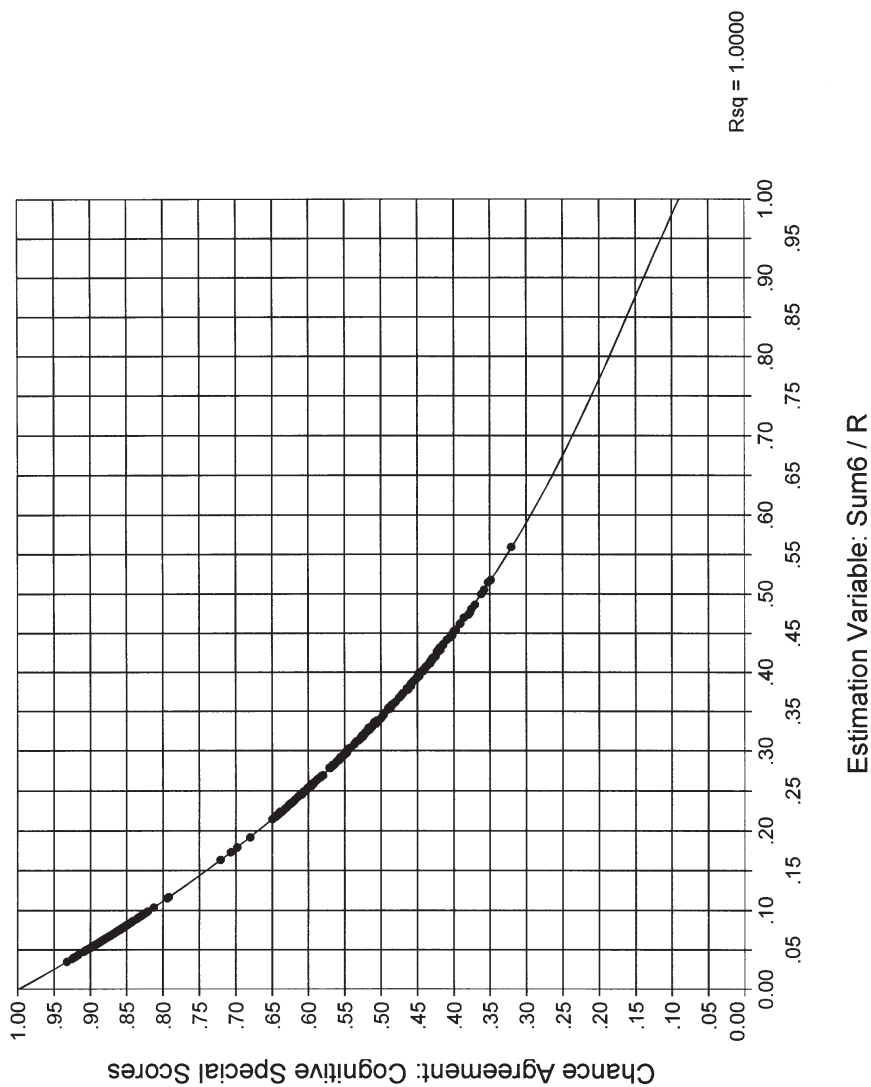

FIGURE 9 Estimates of chance agreement for the Cognitive Special Score segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

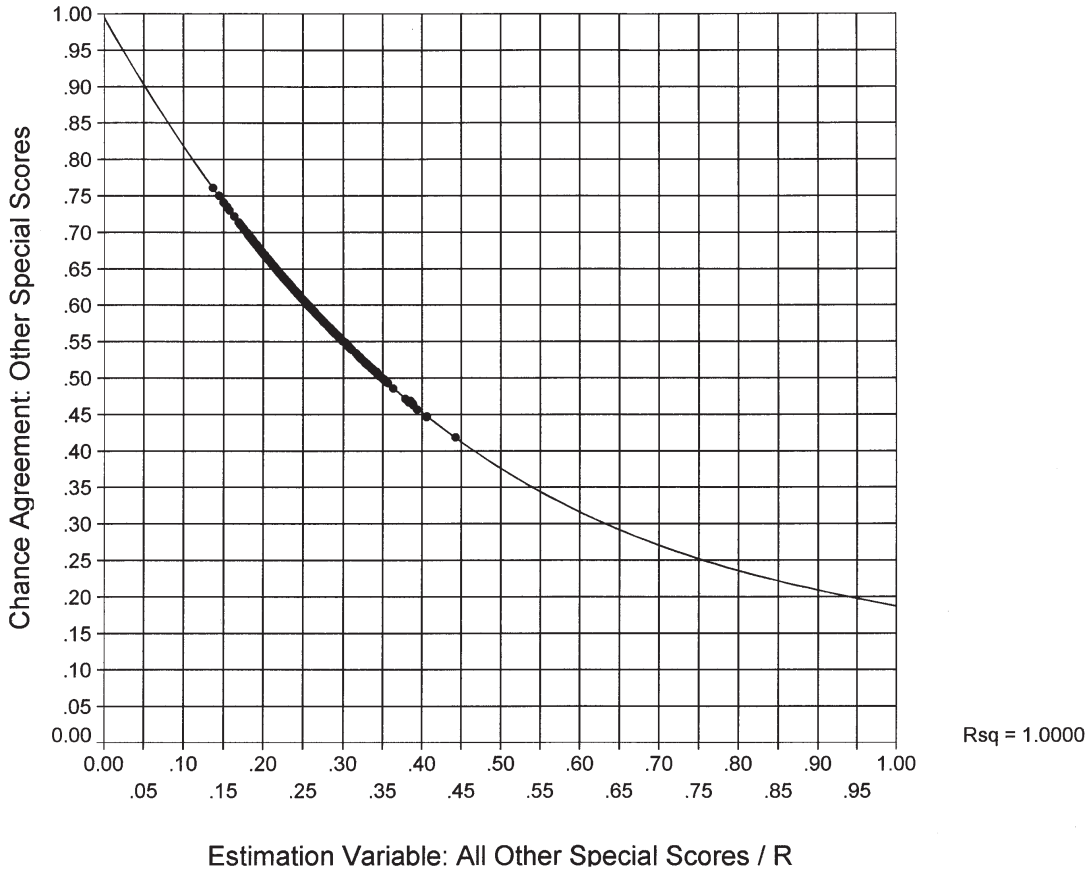

FIGURE 10 Estimates of chance agreement for the Other Special Scores segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

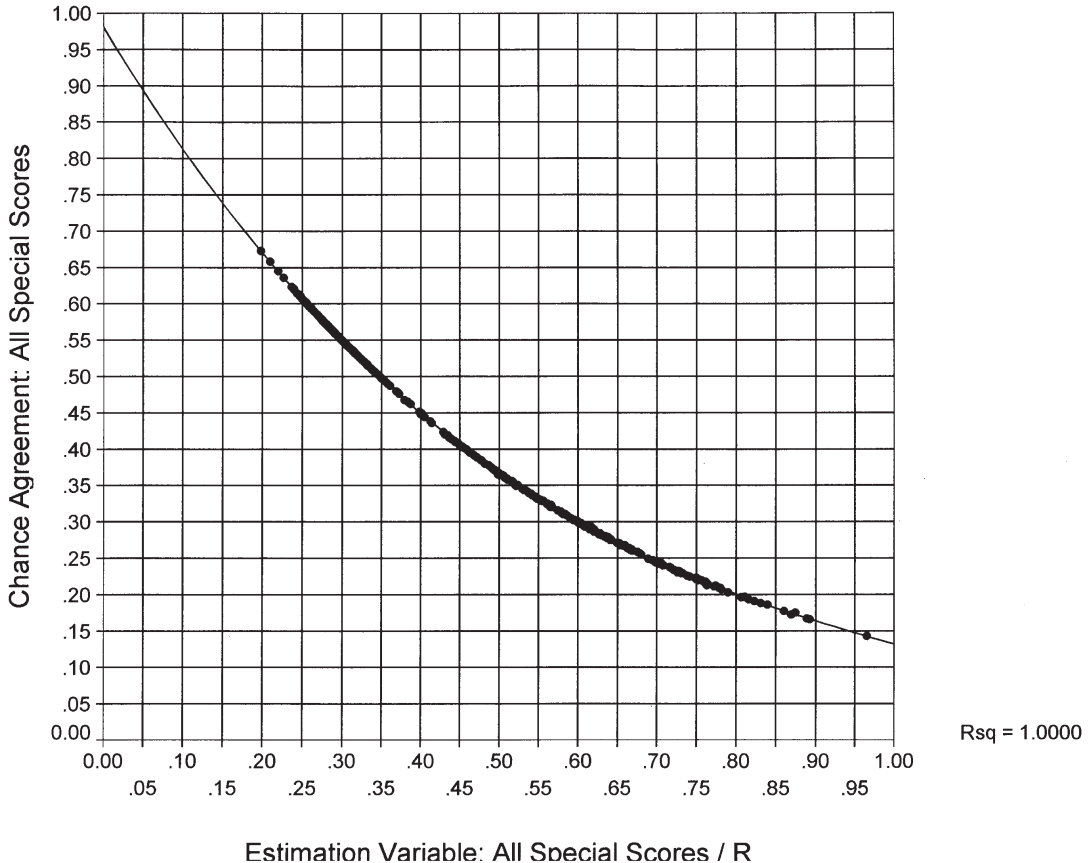

FIGURE 11 Estimates of chance agreement for the All Special Scores segment (Comprehensive System estimation variables refer to the sum of scores from both raters).

TABLE 3
Differences Between Estimated and Observed Chance Agreement Rates and Kappa Coefficients for 11 Comprehensive System Response Segments in Five Reliability Samples

		Response Segments										
Sample	*R*	*Loc*	*DQ*	*Det*	*FQ*	*Pair*	*Cont*	*Pop*	*Z Freq*	*C-SS*	*O-SS*	*A-SS*
Differences in chance agreement rates												
1	1,047	–.0128	.0015	.0116	–.0001	.0000	–.0019	.0000	.0000	–.0003	–.0003	.0004
2	1,125	–.0099	.0020	.0345	.0007	.0000	.0012	.0000	.0000	–.0012	.0004	.0004
3	338	–.0066	.0234	.0020	.0010	.0000	.0030	.0000	–.0001	–.0002	–.0002	.0003
4	1,667	–.0097	.0003	.0055	.0007	.0000	.0012	.0000	.0000	–.0021	–.0002	–.0016
5	1,378	–.0085	–.0007	–.0052	.0008	–.0002	–.0033	.0000	–.0001	.0004	.0000	.0004
M		–.0095	.0053	.0097	.0006	.0000	.0000	.0000	.0000	–.0007	–.0001	.0000
Differences in final kappa calculations												
1	1,047	.0016	–.0005	–.0052	.0000	.0000	.0004	.0000	.0000	.0004	.0002	–.0002
2	1,125	.0002	–.0003	–.0015	–.0001	.0000	–.0001	.0000	.0000	.0005	–.0001	–.0001
3	338	.0000	–.0010	–.0005	–.0001	.0000	–.0001	.0000	.0000	.0001	.0000	–.0001
4	1,667	.0004	.0000	–.0003	–.0001	.0000	–.0001	.0000	.0000	.0012	.0001	.0006
5	1,378	.0032	.0005	.0023	–.0002	.0001	.0008	.0000	.0001	–.0002	.0000	–.0001
M		.0011	–.0003	–.0010	–.0001	.0000	.0002	.0000	.0000	.0004	.0000	.0000

Note. *Loc* = Location; *DQ* = Developmental Quality; *Det* = Determinant; *FQ* = Form Quality; *Cont* = Content; *Pop* = Popular; *Z Freq* = *Z* Frequency; *C-SS* = Cognitive Special Scores; *O-SS* = Other Special Scores; *A-SS* = All Special Scores.

ues based on this estimate will be too low by a small margin. For instance, assuming an observed agreement rate of .65, the actual kappa coefficient should be .3497 but the estimated value would be .2766. Thus, under the worst of plausible circumstances, when one rater obtains a *Zf* base rate that is the lowest observed across 400 samples containing 10 to 20 records and the other rater obtains a *Zf* base rate that is the highest seen in 400 samples of records, the estimation formula will cause the final kappa value to be too low by a magnitude of .0731.

In general, it is unlikely one will see such extreme differences in rater base rates. Nonetheless, the final column in Table 2 presents data on the validity of the prediction formulas using the similarly stringent test imposed when each of the 100 cross-validation samples were paired with another randomly selected sample. Again, this data set evaluates what happens across 100 reliability studies (i.e., the 100 samples from Rater 1 paired with the 100 samples from Rater 2) when one rater produces base rates (and scores) that have no meaningful relation to those produced by the other rater. Even under these highly unusual circumstances, the formulas are still quite accurate. The validity correlations between observed and predicted chance agreement rates range from .940 to .998 (M = .98). Furthermore, across the 1,100 comparisons in this analysis, the average difference between observed and expected chance rates was –.0012 and the largest absolute difference was –.0236. Thus, even when rater base rates diverge substantially across a reasonable range of values, the prediction formulas remain accurate.

Turning to actual interrater data, Table 3 provides validity results from the five genuine samples. The average difference between the predicted and observed chance agreement rate was .00048 across the 55 calculations. More important, in terms of final kappa values, the average difference between the predicted and observed coefficients was only .00011. Across the 55 comparisons, at no time did the predicted and actual kappa values differ by more than .0052. Thus, not only do the regression formulas accurately predict chance agreement rates for genuine reliability samples, but they also quite accurately predict kappa coefficients for response segments.

DISCUSSION

This article is designed for researchers who wish to calculate kappa for traditional CS response segments. Segment reliability coefficients tend to produce conservative estimates of interrater agreement because no credit is given to partial agreement (Janson, 1998). However, because they summarize agreement for global segments of CS scores, they cannot simultaneously provide differentiated information on individual scores. If the latter is required, alternative procedures should be used. Nonetheless, unlike kappa or intraclass correlations calculated on individual scores, a virtue of the segment coefficients is their stability in relatively small sam-

TABLE 4
Illustration of the Estimation Procedures to Calculate Kappa for the Location Segment

1. Two raters independently score 23 protocols containing 500 responses. The raters unanimously agree on the exact location and space scoring for 481 responses. Thus, observed agreement is 481/500 = .9620.
2. Rater 1 assigns 66 *Dd* scores and 58 *S* scores. Rater 2 assigns 67 *Dd* and 63 *S* scores. Using Table 2, the chance agreement Comprehensive System predictor variable is computed as the sum for both raters of (*Dd* + *S*)/*R* = (66 + 67 + 58 + 63)/(500 + 500) = 254/1000 = .254.
3. Using the formula from Table 2 (or Figure 1), the predictor variable produces an estimated chance agreement rate of .51 – .92(.254) + .66(.254^2) = .51 – .23368 + .04258056 = .3189.
4. Because κ is defined as (observed agreement – chance agreement)/(1 – chance agreement), κ for the Location and Space segment in this reliability sample is (.9620 – .3189)/(1 – .3189) = .6431/.6811 = .9442.
5. Common interpretive guidelines (Cicchetti, 1994) for kappa are as follows:

Values	Interpretation
< .40	Poor
.40–.59	Fair
.60–.74	Good
> .74	Excellent

Consequently, in this example, the raters demonstrated "excellent" chance-corrected interrater agreement.

ples. This makes segment-level analyses suitable when researchers wish to document their general reliability as a precursor to reporting validity findings.

To generate kappa for a response segment, four steps must be followed. First, one must determine the extent to which two raters agree exactly on the scoring for a segment.[5] Second, one must generate the CS predictor variable that will be used to estimate chance agreement. This is a matter of simply computing the scores specified in the second column of Table 2. Third, one must use the predictor variable with either the regression formulas or Figures 1 through 11 to determine the chance agreement rate. Finally, one must calculate kappa for the response segment. This is accomplished by inserting the percentage of observed agreement (from Step 1) and the percentage of chance agreement (from Step 3) into the kappa formula. Table 4 illustrates these steps with a hypothetical example. Given that researchers frequently report percentage agreement for response segments, they should now also be able to quickly and accurately estimate kappa.

The results from these analyses should readily generalize to reliability samples containing 20 or more protocols. Statistical modeling not detailed here indicated that the results hold with only minor decrements in accuracy for reliability samples

[5]Agreement for the *Zf* segment is unlike the others. For *Zf*, agreement is counted whenever both raters assign any *Z* value or whenever they both agree no *Z* value is warranted. A disagreement is counted whenever one rater assigns a *Z* score and the other does not.

as small as 5 or 10 protocols, so long as the protocols are randomly selected from the study population. However, the estimation formulas start to break down when a reliability sample contains less than five protocols because very small samples may contain unusual base rates for some score options. For example, it is more possible that a sample of two records will have almost 100% animal contents or 0% special scores. Under these circumstances, the estimation formulas for the Content and Special Score segments may be less accurate. This problem is rectified by following Weiner's (1991) recommendation to select at least 20 records for a reliability study.

ACKNOWLEDGMENT

I thank John Exner for providing many of the Rorschach protocols that contributed to these analyses.

REFERENCES

Bland, J. M., & Altman, D. G. (1995). Comparing two methods of clinical measurement: A personal history. *International Journal of Epidemiology, 24,* S7–S14.

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

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

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

Erdberg, P., & Cooper, B. A. (1998, February). *Assessing intercoder agreement on the Rorschach.* Paper presented at the annual meeting of the Society for Personality Assessment, Boston.

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.

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.

Janson, H. (1998, February). *Coefficient iota: A chance-corrected agreement measure for multivariate observations.* Paper presented at the annual meeting of the Society for Personality Assessment, Boston.

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

Meyer, G. J. (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., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Hilsenroth, M. J., Piers, C. C., & Resnick, J. (1999). *An examination of interrater reliability for scoring the Rorschach Comprehensive System in five data sets.* Manuscript submitted for publication.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321–325.

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.

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 22, 1999