Calculating and Reporting Rorschach Intercoder Agreement

Half day workshop given at the Annual Meeting of the Society for Personality Assessment, New Orleans, LA, March 2008

Harald Janson
Atferdssenteret, The Norwegian Center for Child Behavioral Development
P.O. Box 1565 Vika
N-0118 Oslo, Norway
E-mail harald.janson@atferdssenteret.no
Phone +47 24 14 79 08
Fax +47 24 14 79 46
Cell phone +47 92 61 08 78

Contents

Acknowlegdements

This work is a development of a previous version of this workshop, presented at the Midwinter Meeting of the Society for Personality Assessment in San Francisco, CA, March 2003. Parts of the present text were first presented as a paper presentation to the XVII International Congress of Rorschach and Projective Methods (Janson, Meyer, Exner, Tuset, &

Nakamura, 2002). I am thankful to Philip Erdberg, Anna Maria Carlsson, Gregory Meyer, John E. Exner, Ana Maria Tuset, Noriko Nakamura, and Thomas Lindgren for their input and contributions to this work.

Summary

This practically-oriented workshop will suggest choices for procedures in conducting a Rorschach reliability study, including how to report results, and give hands-on practice in calculating agreement on real-life Rorschach data sets. Calculating univariate and multivariate response-level agreement, as well as protocol-level agreement, for two or more coders, either the same for all protocols or different for different protocols, will be covered. Software for interrater agreement calculations will be demonstrated.

Theory

This section will briefly go over some basic concepts, give a sketchy overview over the issues about Rorschach intercoder agreement discussed in the recent literature, and end with pointing out some important choices for intercoder study procedures as well as for reporting agreement in publications.

Some terminology

Usage in the literature varies somewhat. When talking about the Rorschach, I prefer the term "coder" rather than "rater" or "judge" because it better describes the task of coding a client's verbalizations in a Rorschach protocol according to a manual. I use the term "clients" to refer to the examinees, (patients, research participants, etc.). I write "protocol" rather than "Rorschach record."

Some basics of reliability and interrater agreement

I am only going to make a few points here which are pertinent to the discussion of Rorschach intercoder agreement. For an introduction to reliability issues, see for example Pedhazur & Schmelkin (1991; Chapter 5, pp. 81-117). For the various issues involved in Rorschach agreement and reliability, see for example Meyer, Hilsenroth, Baxter et al. (2002), Acklin, McDowell, Verschell, & Chan (2000), and Meyer (1997a; 1997b).

Some reliability theory

There are several ways of conceptualizing reliability. In classical test theory, an *observed score* is the sum of two uncorrelated terms, a *true score* (which cannot be known) and *error*. The variance of the observed score can be decomposed into the variance of the true score and the variance of the error term. Most reliability estimates use an expression of the correlation between two measures (i.e., two observed scores) to estimate the proportion of the variance attributable to true score variance (or systematic variance). Although today strict classical test theory is not viewed as a realistic model for most actual data situations, its basic concepts are well known and heuristically useful.

Some important points about reliability that have consequences for Rorschach intercoder agreement are:

- Reliability (as validity) is a property of test scores, not tests (Messick, 1989)
- Reliability is specific to a population (in the case of interrater agreement, actually, populations of clients/protocols and coders)
- All other things being equal, the estimate of reliability increases with increasing variability in the variable of interest in the sample/population at hand

I will comment specifically on the consequences for Rorschach data of reliability being specific to a population. In the specific case of Rorschach data, agreement depends on both the population of clients (protocols) and the population of coders. A Rorschach protocol may be more difficult to code (resulting in lower observed agreement) if responses are complex, bizarre, poorly verbalized, poorly inquired, or if protocols are nonstandard in any way, such as archival records administered according to non-current guidelines, translated protocols, handwritten protocols, and so on.

Reliability measures depend on the variability in the sample. Thus (and maybe counter-intuitively), reliability estimates *may tend* to be higher in heterogeneous samples (such as mixed nonpatient and patient samples with varying ethnicities and ages) as compared to homogeneous samples (e.g., only white working male Caucasic nonpatients; or only patients fulfilling certain diagnostic criteria. This is because a homogeneous population *may often be expected to* have smaller variance in Rorschach scores. Because reliability and variability are interdependent, descriptive statistics are of interest for evaluating agreement measures.

As with any complex coding or rating procedure, agreement for Rorschach codings is expected to increase with the skill, training, and co-training of coders. Coders working under laboratory-like conditions (i.e., experts or students under supervision in a setting where research is the main objective) may be expected to perform differently than clinicians using the Rorschach as part of their daily routine. Additionally, it might perhaps be supposed that coders perform better if they are aware that they will be compared with others (as when it is known what protocols in a sample will be used for calculating intercoder agreement). In order to evaluate agreement results and draw inferences about the conditions these may generalize to, the reader needs to have relevant information about the nature of the populations of clients and coders.

Facets of reliability

Variance attributed to "error" may have many different sources. The correlation between, for example, two Rorschach tests some time apart, with different examiners, may be less than perfect because of systematic and random person effects (people develop gradually over time, and swing in mood, concentration, and so on), systematic and random environment effects (testing during the evening versus morning time might have a general effect, or a loud TV set in the next room might have a temporary influence), systematic or random examiner effects (some examiners inquire more, and examiners, too, have mood and attention fluctuations), and systematic and unsystematic coder effects (some coders code Form Quality Minus more often than others, and all make occasional errors). All these components, or *facets,* would be part of the "error" variance. Any single reliability measure, however, lumps together several sources of unreliability in the "error" term. Due to the design of a particular study, some sources of variability count as error, and some count as "true" variance. (A development in reliability theory, *generalizability theory* (see e.g., Shavelson & Webb, 1991), allows to take into account and estimate sevaral facets at one time. The present presentation does not allow to go into the details of this more refined approach.)

In calculating and report intercoder reliability (i.e., the amount of variance not due to coder error), we are only leaving two things in the "error" component: systematic and random coder error. (Some people are even reluctant to speak about intercoder "reliability" because they consider that the facet under study is too narrow to be telling about the reliability of a variable.) It is thus possible to demonstrate near perfect intercoder reliability for, for example, Rorschach variables, in a range of populations, without this supporting the reliability of the score(s) in any more general way (i.e., over test occasions, examiners, and so on).

The published literature strongly suggests that Rorschach Comprehensive System coding may be expected to be very reliable over a range of populations of protocols (clients) and appropriately trained coders (Meyer et al., 2002; Shaffer, Erdberg, & Meyer, 2007; Viglione, 2003). However, we must not let this general finding lead us to conclude that Rorschach coding is a trivial task to which we should devote less effort and attention. Rorschach coding is indeed an inherently challenging task, and as mentioned above, reliability is specific to populations. It will therefore remain to be of utmost importance to keep focus on quality of Rorschach coding. Sound practices in training coders, collecting, and reporting agreement data, must remain a central part of the continuing effort of Rorschach research (for a discussion of administration issues, see also Lis, Parolin, Calvo, Zennaro, & Meyer, 2007).

Agreement versus reliability

Sometimes "agreement" is mentioned as if it sere something altogether different from "reliability." The two concepts clearly have different histories and have been used in different contexts. Traditionally, "agreement" has been used to refer to observed, absolute, agreement, i.e., the extent to which two persons could apply the same codes to the objects in a sample. "Reliability" brings to mind the notion of an estimate the proportion of variance in a variable among a population that is not attributable to error (in the case of intercoder reliability, coder error).

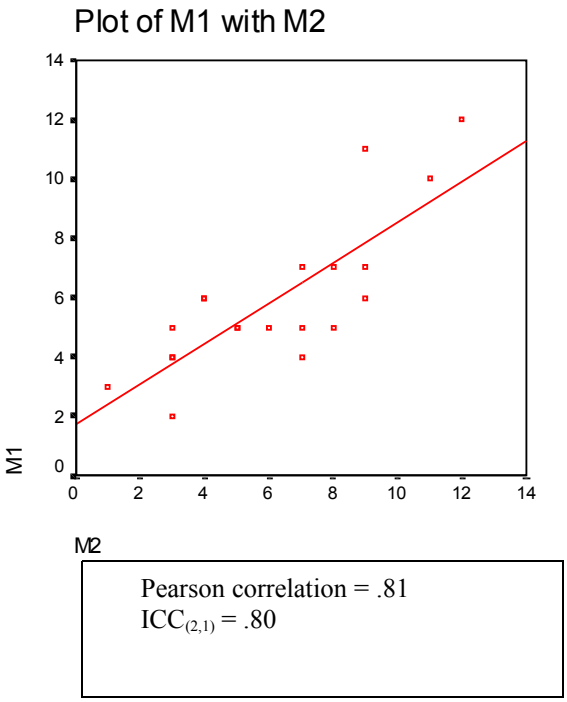
However, *chance-corrected agreement amounts to reliability when certain conditions are met.* Specifically, it has been shown that chance-corrected measures of agreement of the kind discussed in the following (i.e., *kappa* measures), are at the same time applicable to reliability measurement (e.g., Kraemer, Periyakoil, & Noda, 2002). On the other hand, some authors have avoided using the term "interrater reliability" because "reliability" is then only understood in a very narrow sense (i.e., the amount of variance that is not due to coder disagreement, while assessment of reliability of scores conventionally includes facets such as items, test forms, and measurement occasions) (e.g., Pedhazur & Schmelkin, 1991). Because chance-corrected agreement amounts to reliability when certain conditions are met, I will in the present text employ the terms "agreement" and "reliability" somewhat interchangeably throughout this presentation, although I do recognize the diverging connotations of the two terms. "Agreement," when not otherwise qualified, will be used to refer both to chance-corrected agreement--as estimated by kappa-type measures--and interrater reliability--in conventional terms, such as estimated by intraclass correlation coefficients--of scores.

Reliability and intraclass correlations

Intraclass correlations are a family of coefficients that estimate the proportion of variance in a variable that is not due to error. (In conventional reliability terms, intraclass correlations estimate the proportion of variance that is due to "true score" variation; in generalizability theory terms, the estimate is of the proportion of variance that can be generalized across the facets of reliability that are included in the design.)

One important difference between intraclass correlations and the Pearson correlation coefficient is that the Pearson correlation coefficient standardizes scores before computing the coefficient. Intraclass correlations, in contrast, may also take into account the absolute difference among coders, that is, in the "absolute agreement" varieties of intraclass correlations. To take an example, let's say two coders have scored 20 protocols for the number of human movement (*M*) responses and obtained the following results:

	M (Human Movement)	
Protocol	Coder 1	Coder 2
1	6	4
2	2	3
3	7	9
4	5	3
5	7	8
6	5	6
7	5	5
8	4	7
9	12	12
10	5	5
11	5	7
12	4	3
13	6	9
14	4	3
15	11	9
16	3	1
17	7	7
18	5	8
19	10	11
20	6	4



When, as in the example above, the two coders' distributions of scores are similar, the intraclass correlation $ICC_{(2,1)}$ does not differ greatly from the Pearson correlation. However, if for some reason Coder 2 were systematically to score two more *M*s for every protocol (M=6 for the first protocol, M=5 for the second protocol, and so on), the situation would be different. The Pearson correlation would remain unchanged by the mean difference, and still be .81. However, the intraclass correlation would fall to.60. This fall is the consequence of absolute agreement-type intraclass correlations accounting differences among the coders as disagreements.

There exist several forms of intraclass correlations, depending on the nature of the data. The illustrated example is a "two-way" data example, which means that the same set of coders has scored all of the objects in the sample. The term "two-way" comes from ANOVA terminology and reflects the fact that a two-way ANOVA is performed (with observations as one factor and coders as the other factor) in order to obtain the terms for the intraclass correlation.

When there are several coders and different objects have been coded by different coders, the appropriate intraclass correlation is the "one-way" intraclass correlation. The terms in this variety of correlation are obtained from a one-way ANOVA (with observations as factors). The one-way intraclass correlation is defined for varying numbers of coders for different objects (see e.g., Haggard, 1958).

Further, intraclass correlations have several other interesting properties. Intraclass correlations can account for the reliability among several coders (not just two).

In the example above, the intraclass correlation estimates the reliability of a single coder's scoring. This is denoted by the "1" in "$ICC_{(2,1)}$." In addition, there are forms of the intraclass correlation which estimate the reliability of the average of the several coders' scorings.

Definitional and computational formulas for the two most widely applied intraclass correlations, as well as procedures for obtaining these measures in SPSS, are given in Appendix 1 (see page 20).

Cohen's *kappa*

Cohen's (1960) *kappa* is the commonly used chance-corrected agreement measure for nominal data. It is defined for dichotomous and polytomous variables. Cohen's original measure applies to two-way data (with a set of two coders who have coded all of the observations). However, a number of extensions have been formulated, notably to interval data (Fleiss and Cohen, 1973; Janson & Olsson, 2001; 2004), one-way data, and several coders (Fleiss, 1971; 1981; Janson & Olsson, 2004).

It is important to realize that kappa is a reliability measure, although the original formulation of *kappa* was of an agreement measure corrected for chance (Kraemer, Periyakoil, & Noda, 2002). For dichotomous and interval variables, therefore, appropriate extensions of *kappa* equal the corresponding forms of the intraclass correlation. See Appendix 1 for computational formulas and SPSS syntax for obtaining *kappa*.

An extension of Cohen's kappa to multivariate data and several observers

Janson and Olsson's (2001) *iota*, an extension of Cohen's (1960) *kappa,* allows one to compute chance-corrected agreement for a multi-variable test scored by one set of two or more multiple raters. The measure is applicable to nominal or interval variables, and it can be used to examine agreement on individual responses or summary scores. The statistic's metric is conventional, and in applicable cases it is equivalent to existing extensions of the *kappa* coefficient to several observers, as well as to the two-way random intraclass correlation coefficient. More recently, Janson and Olsson (2004) extended their model to the case with varying numbers of different raters for different cases.

In brief, Janson and Olsson's *kappa* extension is enabled because Cohen's *kappa* (which was first expressed in terms of observed and expected *agreement*) can be reformulated in the following way:

$$\kappa = 1 - \frac{(\text{observed disagreement})}{(\text{expected disagreement})} \qquad (1)$$

Janson and Olsson (2001) used the notation d_o and d_e for observed and expected disagreement, respectively. To take a simple example (Figure 1), two psychologists may have coded eight Rorschach responses for absence or presence of human movement (*M*). For two responses, the two judges agree on the presence of *M*. For four responses, the coders agree that *M* is not present. In two cases they disagree about the coding of *M*. They are thus in disagreement in two cases out of eight, or in a proportion of .25 of the cases.

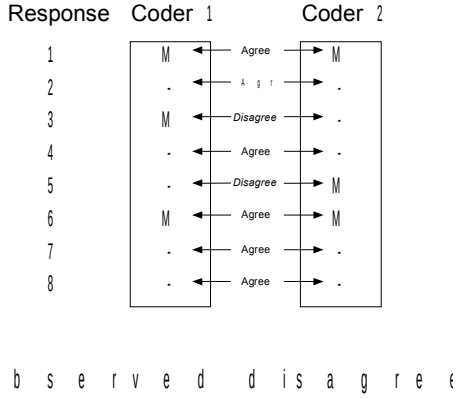


Figure 1. Observed Disagreement of Two Coders' Codings of Eight Rorschach Responses for Human Movement (M)

Chance-expected disagreement can be obtained by comparing each coding from the first coder with each of the other coder's codings. Figure 2 shows how, if we start with the first coder's coding of response 1, it agrees with the other coders' coding of response 1, disagrees with his coding of response 2, and so on. There are 8 by 8, that is, 64, possible such comparisons. The eight first are shown with arrows here. Of the total 64 comparisons, 30--or about 47%--are disagreements. This proportion, the average disagreement between one judges' rating of one target, and another judges' rating of any target, is also the average disagreement that could be expected if the responses were not matched (that is, scrambled within judges), and also the average of disagreement obtained in all possible matchings of responses.

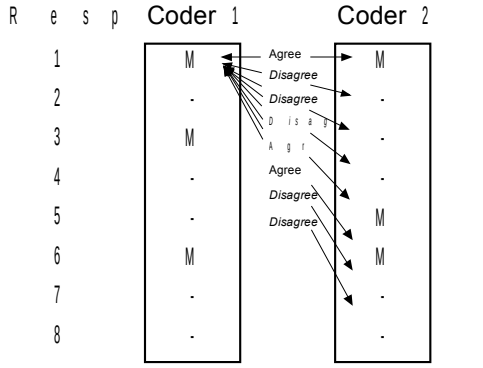


Figure 2. Chance Expected Disagreement

Janson and Olsson (2001) noted how the two terms of observed and expected agreement are informative in their own right: The proportion of observed disagreement (d_o) tells how great the disagreement was, on the average, between any pair of coders. The proportion of expected disagreement (d_e) tells how much disagreement might be expected by chance, when chance is defined as being equal to the distribution of responses in the sample at hand.

Cohen's *kappa* for this example, according to Equation 1, is calculated as one minus the ratio of observed disagreement to expected disagreement--in the example .53. With this kind of data, Equation 1 is mathematically equivalent to Cohen's (1960) original formula.

The reformulation of *kappa* as in Equation 1 enables straightforward extension to interval-level data, to multivariate data, and to several observers. Interval-level data can be accommodated if the squared Euclidean distance is used for the measure of disagreement between two observations (following Fleiss & Cohen, 1973). The resulting *kappa* is then equivalent to an intraclass correlation (except for a small term resulting from the use of *n* as the degrees of freedom for cases in *kappa* instead of *n*-1 as in the intraclass correlation). Multivariate observations can be addressed because the terms of observed and expected disagreement can readily address multivariate distances (when with nominal data, the simple number of disagreements over variables is used as the disagreement measure, and with interval data, the squared Euclidean distance in multivariate space is used). The terms of observed and expected disagreement may also be averaged over any number of observers, not just two. Janson and Olsson (2001) used the Greek letter *iota* (ι) to distinguish the *kappa* extension they proposed from Cohen's (1960) original measure. It should be noted that there exist alternative formulas for calculating the disagreement terms. These alternative equations--which make use of ANOVA sums of squares--involve much fewer calculations (See Appendix 1).

Cohen's (1960) *kappa* and Janson and Olsson's (2001) extension apply to the case when there is one same set of raters for all observations. Following an approach by Fleiss (1971; 1981), Janson and Olsson (2004) presented definitions of observed and expected disagreement that extend the multivariate *kappa* to the one-way case, that is, when there are different raters (not necessarily equal in number) for different observations.

Overview of critical issues for Rorschach intercoder agreement

Several suggestions concerning standardized procedures for calculating and reporting agreement that expand on the first recommendations (Weiner, 1991) have been put forward over the past decade or so (e.g., Acklin, McDowell, Verschell, & Chan, 2000; McDowell & Acklin, 1996; Meyer, 1999; Meyer, Hilsenroth, Baxter, et al., 2002). The following sections attempt to summarize some of the central issues that have been discussed over these years that concern choices to consider in designing an interrater agreement study, and computing and reporting these results.

Protocol versus response level agreement

Rorschach data is coded at the level of single responses, and summed to protocol-level variables. Agreement can be accounted on several levels: For protocol-level summary variables, as well as for response-level dichotomous coding decisions, complex coding decisions, response segments, or even whole responses. For a discussion of the various considerations involved, see for example, Acklin et al. (2000) and Meyer et al. (2002).

Response-level agreement is informative about the precision with which coders were able to apply the same categorical coding scheme to responses, and is useful for monitoring training and practice. The first set of JPA recommendations for publishing intercoder agreement (Weiner, 1991) recommended only response-level agreement for "response segments," and these are still the most commonly reported measures (Meyer, 1999; Viglione & Taylor, 2003).

Reliability for protocol-level summary variables is most informative about the reliability of data at the level that it is used for clinical decisions or research about people. Wood, Nezworsky, and Stejskal (1996a; 1996b; 1997) pointed out that this kind of data had been sparsely reported until then, and went as far as to suggest that the reliability of the Comprehensive System was at best unknown. However, the number of published studies which have reported protocol-level agreement results for a range of Roschach Comprehensive System variables in a wide range of populations is now rather impressing, and suggests overall very high levels of reliability among appropriately trained coders (Meyer et al., 2002; Viglione & Taylor, 2003).

Overall (per cent) versus chance-corrected agreement for single decisions and response segments

Until the 1990s, agreement on the response-level categorical coding variables was usually reported as per cent agreement (observed agreement; overall proportion of agreement). However, the distribution of Rorschach coding categories is for the most part very uneven, with low rates of "present" codings for many variables, which yields high agreement estimates as a simple result of many agreements for "absent" codings.

Recent writers agree that agreement among coders should preferably be reported by means of a chance-corrected measure, such as Cohen's kappa or an intraclass correlation coefficient (Acklin et al., 2000; McDowell & Acklin, 1996; Meyer et al., 2002). These two types of measures are essentially equivalent, have conventional metrics, and are also directly applicable to reliability measurement.

Chance-corrected measures take into account the sample base rate of a category, and report the amount of agreement above this chance expectation. This brings with it two considerations. First, the definitions of chance underlying each measure, play a role in the expression of agreement. Second, and relatedly, chance-corrected agreement measures may be variable, and biased, when the base rate of the phenomena under consideration is low.

Low base rates

The problems with applying chance-corrected measures to variables with low population base rates (or low population variances) have been one focus of the Rorschach reliability literature (Acklin et al., 2000; Meyer et al., 2002). The meaning of the phenomenon is that the chance agreement term becomes a poor estimate of the actual chance agreement when a coding category occurs very infrequently. As a result, when a category appears very infrequently in a sample, little confidence may be placed in the agreement/reliability coefficient.

Empirical findings (e.g., Acklin et al., 2000) suggest that low base rates more often result in too low estimates of reliability of ratings, than too high. If appropriate estimates of the variability of the agreement measures (i.e., standard errors and/or confidence intervals for intraclass correlations and for *kappa*) were available, such variability estimates would help in interpreting the expected loss of precision in the coefficient which was the result of a low base rate. Unfortunately, the published (asymptotic) expressions for the variability of intraclass correlations and *kappa* (which are implemented in statistical software packages such as SPSS) are not applicable when base rates are very low. The available estimates of variability of agreement coefficients yield too small standard errors, and too small confidence intervals. Bootstrap or jackknife estimates might be better, but are not implemented in the major statistical software packages.

A consensus regarding a resolution of this problem has not been reached. Pending further evidence, it is recommended that agreement be reported for single coding categories only if the code appeared with a certain frequency. Minimum base rates of .01 or .05 for reporting agreement have been suggested in the literature, but a consensus has not been reached concerning this issue (see Acklin et al. 2000; Meyer et al., 2002).

Response level agreement for response segments versus single coding decisions

Response segments. Response-level agreement for so-called coding segments as well as for whole responses is informative about the overall agreement among coders for a complete multivariate coding task. Some varying approaches have been applied to estimating agreement on this level. The approach that I suggest has the advantages of clear interpretation and more straightforward calculation than some of the alternatives. I suggest to treat different coding categories as several separate descriptive dimensions, and to use Janson and Olsson's (2001; 2004) multivariate approach to calculate measures of coders' joint agreement on all dimensions. Janson and Olsson's (2001; 2004) extensions of Cohen's *kappa* are useful for calculating Rorschach intercoder agreement, precisely because this approach addresses multivariate data like Rorschach response segments, and several coders--which may vary among protocols and need not be the same in number.

(Alternative procedures for calculating chance-corrected agreement for response-segments that have been proposed [McDowell & Acklin, 1996; Meyer, 1999], instead calculate *kappa* for a very large cross table of all occurring combinations for a segment [i.e., Determinants]. This necessitates complex data treatment, and the results are also not quite easily interpreted. Because only an exact correspondence between two coder's codings is counted as an agreement, the proportion of disagreements on the combined variable tends to increase with the number of coding decisions.[1] Estimates of observed and expected agreement for multi-decision segments such as Rorschach Determinants and Contents are then based on a relatively small proportion of agreements. Furthermore, the proposed procedures for calculations of chance-corrected agreement for Rorschach response segments also differ in some details among authors.)

Single coding decisions. While agreement for response segments are informative about the overall level of agreement for all of the decisions included in the "segment," the results are not informative of the reliability of single coding decisions (Janson & Olsson, 2001).[2] When a study relies heavily on certain single coding decisions, it might be preferable to report the response-level agreement for these decisions along with protocol-level agreement for the corresponding variables. This way, it will be as clear as possible to which degree coding decisions on the level of individual responses mattered for the results of the study (see e.g., Meyer, 1997a; 1997b). In addition, response-level agreement data for single coding decisions is useful for training purposes, for monitoring ongoing projects (because estimates of response-level agreement requires fewer protocols except for categories with very low base rates), as well as for accumulating evidence pertinent to the applicability of coding definitions, for future refinements to the coding system.

One set of two (or several) coders for all of the protocols, or different coders for different protocols

One may wish to calculate agreement among two or several coders, and Rorschachs may have been coded by one set of coders, or by different coders for different protocols. Different measures (in fact, different estimates of chance agreement) are applicable in these different cases.

The term "two-way data" describes the case when there is one set of coders for all protocols. This name comes from the fact that a two-way ANOVA layout is used to derive the appropriate intraclass correlation (i.e., $ICC_{(2,1)}$) in this case. As is clear from its definition, Cohen's (1960) original *kappa* is also a two-way measure. Two-way measures should only be applied when the coders are the same for all of the observations. (Until recently, computation of two-way measures in common statistical software packages like SPSS strictly required that all of the coders had coded each of the observations.

---

[1] This relationship may not be easy to understand intuitively. To exemplify, assume that 2 raters have rated a large number of objects on k uncorrelated dichotomous variables, and that each rater's base rate on each variable is .5, and that disagreements between the raters occur randomly on every 100th decision. For each of the single dichotomous variables, the expected values of Cohen's p_o, p_e, and *kappa* are then .99, .50, and .98. Compute a new variable which has as many categories as there are possible combinations of the k basic variables. (With k=2, there are 4 possible combinations--00, 01, 10, and 11). The corresponding expectations when k=2 are .98, .25, and .97. With k=27 (the number of Rorschach content categories), the expectations are .37, 8×10^{-9}, and .37. That is, as k increases, p_o, p_e, and *kappa* decrease (and, as Meyer [1999] noted, *kappa* comes closer to p_o). In contrast, the expectation for *iota* in this example is .98 regardless of k, and the expectations for p_o and p_e are proportional to k.

[2] Kraemer et al., 2002, go as far as to discourage from use of even multi-category variable *kappa* measures, because such measures don't express the reliability of any variable. I don't agree with this view, and have shown that multi-category as well as multivariate *kappa* measures have a meaningful interpretation; see Janson & Olsson, 2001.

The variance components estimation procedures available in more recent versions of SPSS enable estimation of two-way reliability even with some missing observations--see Appendix 1, page 20.)

When there are different coders--not necessarily equal in number--for different protocols, this is similar to a one-way ANOVA layout. When one-way data is at hand, one-way intraclass correlation or a modification of Cohen's kappa appropriate for one-way data (e.g., Fleiss, 1971; 1981) should be applied. One-way data ideally assumes that a rather large number of different coders each have performed a small number of observations. However, applying a one-way measure to a data set where the same coders have scored a fair proportion (but not nearly all) of the data does not normally constitute a serious violation of the assumptions. Normally, applying a one-way agreement measure can in this case be expected to slightly underestimate the reliability. The situation where one set of coders have scored a great proportion (but not all) of the protocols in a study, should be avoided if possible, because this situation does not fit any of the sets of assumptions underlying the commonly used agreement measures.

The most commonly used forms of the intraclass correlation assume that the coders are sampled from a larger population of possible coders. (The alternative assumption that the coders in the analysis are the only coders of interest, probably very seldom is tenable.) Reliability measures use the correlation between two measures to estimate the amount of variance due to error (versus systematic or true variance). A maybe surprising consequence of these assumptions and procedures, is that the agreement/reliability results may not be easily interpreted if the coders are not of roughly the same training and experience. When coders have roughly an equal level of training, co-training, and experience, the results may be expected to represent the reliability in a population of similar coders. On the contrary, when coders are very unequal in proficiency, such as one senior expert and one novice student, it is very difficult to say what the results represent, and what situations they would generalize to (an agreement in the middle range may, or may not, result from one of the coders having almost perfect reliability and almost all of the errors being committed by the other).

Suggestions for procedures for collecting intercoder agreement data

The following are my suggestions for choices to be considered in conducting a Rorschach intercoder agreement study.

Protocol considerations

Consider whether you can have your complete sample of protocols coded by two or more coders. If this is feasible it might be preferable on two grounds:

- Recent findings suggest that chance-corrected agreement measures may be very variable, and on the average biased downwards, when the number of protocols is small and the base rates of phenomena are low. A minimum of 20 protocols, which has until now been a common sample size, may be inadequate.
- If all of your protocols have been coded by the same number of coders, you might want to consider using average Comprehensive System scores as your research data (particularly if your study draws on more or less continuous Rorschach variables). The average score will be more reliable than a single coder's score, and the reliability of the average score can be estimated by applying special forms of the intraclass correlation coefficient (see e.g., Shrout & Fleiss, 1979; McGraw & Wong, 1996; SPSS reports these estimates as the "average measure" reliability).

If it is not feasible to have several coders for all of your protocols, you should try to sample a number exceeding than 20 protocols (preferably much more) from your study protocols if you plan to report protocol-level agreement or response-level agreement for low base rate codes.

If your client sample is homogeneous, it is preferable to select intercoder protocols by randomization. If you have a sample composed of several subgroups, a stratified randomization design might be preferable (i.e., randomly draw a certain number from each subgroup).

It is preferable if coders cannot tell which protocols will be used for calculating intercoder agreement. Intercoder agreement protocols should distributed to coders along with other protocols, and should not be distinguishable from the other protocols.

Coder considerations

Due to the mechanics of the commonly used agreement and reliability measures (*kappa* as well as intraclass correlation coefficients), it is preferable that the protocols in the intercoder agreement study be coded:

- *either* by one set of two or several coders who score all of the protocols,
- *or* by a rather large number of coders who each score a small proportion of the protocols.

This is because "two-way" agreement measures suppose that there is a common set of coders for all responses/protocols, while "one-way" agreement measures suppose that there are a large number of coders who have contributed observations (although it is not a very serious violation to have several observations from each coder). The condition to be avoided is to have a majority of the protocols coded by one same set of coders, and a few protocols coded by other coders. There is no reliability study design that matches such a condition.

Because the assumption that coders are sampled from a larger population of possible coders is either explicit or implicit in the common measures of reliability, it is preferable that the coders in your study have very roughly a similar level of proficiency. Avoid to have two (or a small number of) coders who are very different in coding proficiency as your group of coders.

Ensure that coders work independently, that is, stay completely ignorant about any others' codings of the same protocols.

Suggestions for reporting agreement

Lately, some consensus concerning the minimal necessary reporting of agreement seems to be emerging, as reflected not least in the special issue of *JPA* devoted to international Rorschach Comprehensive System reference samples (Shaffer, Erdberg, & Meyer, 1997). The following are my recommendations, which to a great extent concur with this emerging consensus. However, the following recommendations in some respects are more far-reaching than the minimal necessary level. I suggest that researchers should carefully consider how extensive and detailed their report should be--different formats might be appropriate depending on the study at hand.

Data pertinent to your study variables' reliability

- Provide information relevant to the heterogeneity of your sample of clients/protocols (e.g., homogeneous diagnostic group vs. undifferentiated psychiatric patients vs. nonpatients vs. population sample)
- Provide descriptive statistics for the variables you use in your substantive study (i.e., the main variables of your study)
- Provide information about the difficulty of protocols (e.g., archival, handwritten, translated, etc.)
- Provide information about the amount of training and experience of the coders in your study
- Provide information about the amount of co-training that the coders received
- Describe whether the study was laboratory-like (i.e., expert or student coders under supervision) or field-like (i.e., clinicians under normal work conditions)
- State that the coders worked independently, or describe *fully* any possible departure from this requirement
- If your intercoder study did not include your complete sample, describe how intercoder protocols were selected (i.e., using randomization; stratified randomization; systematically; accidentally) and distributed to coders (e.g., were coders aware which protocols were for the intercoder agreement study; were there some coders who only contributed intercoder study protocols?)
- Describe the freedom that coders were allowed in delimitation of responses (see below, page 14)
- If coders in your study delimited responses differently, and as a consequence some responses were excluded from response-level agreement calculations, report the number of times that this occurred, as well as how it was decided which unmatched response(s) should be omitted

Response segment agreement

For compatibility with earlier published results, and because it is an efficient means of communicating overall coding precision, report response level agreement for so-called coding segments (Location & Space; Developmental Quality; Determinants; Form Quality; Pairs; Contents; Popular; Z-Score; Special Scores). As noted above, some of these segments are multi-category variables, and some are multivariate profiles of several categorical variables. Report chance-corrected agreement using an appropriate measure of the *kappa* family. Report which procedure you used for calculating chance-corrected agreement (e.g., the procedure applying Janson & Olsson's [2001; 2004] multivariate approach, which is implemented in the *RRU* program, or McDowell & Acklin's [1996] procedure).

Whole response agreement

In addition, you might want to report multivariate response level agreement for whole responses taken as a whole using the procedures of Janson & Olsson (2001; 2004). Along with this overall measure on the response level, you should report the terms of observed and expected disagreement for entire CS responses.

Detailed reliability results for single coding decisions or protocol-level variables

Unless your study focuses especially on reliability findings, there is no need, and no room, for reporting agreement results for a large number of single coding decisions or protocol-level variables.

However, protocol-level chance-corrected agreement for the central variables in your study is of interest to readers (because the reliability of your variables is informative for evaluating your validity findings). Therefore, preferably report protocol-level chance-corrected agreement--along with descriptive statistics--for the central Rorschach variables in your study. Use the appropriate intraclass correlation as the chance-corrected agreement measure for continuous protocol-level variables, or the appropriate measure of the *kappa* family for categorical protocol-level variables (e.g. HVI=Positive; Fr+rF>0).

If your study involves Rorschach variables which rely heavily on some particular coding decisions, you might also want to report response-level agreement for those coding variables or single coding categories that are of importance in your study. Use the appropriate measure of the *kappa* family to report agreement. However, do not report agreement for single coding decisions with a very low base rate in your sample (less than .01, or less than .05). Instead, list the coding decisions where this applied, along with the overall proportion of agreement and base rate for these coding decisions.

Practice

The Practice section uses a *"Rorschach Research Utilities" (RRU)*, a Windows program that calculates agreement from RIAP2, RIAP3, or Plain Text (.TXT) files. (RIAP4 and RIAP5 users can export the Sequence of Scores report for each protocol in RTF format, which the *RRU* program can read; ROR-Scan users may export a kind of file that is readable by *RRU*). The program may be freely distributed, without charge, in its present form, according to the conditions listed in the software documentation. I shall use real-life data examples to exemplify the steps involved in preparing data and calculating agreement using the *RRU* program.

Organizing Rorschach data for agreement calculations

One of the first choices you need to think about is to determine if a "one-way" or "two-way" analysis is proper for your data set. This choice has implications for how to array your data, and for what agreement measures to use.

A two-way analysis is proper when the same set of two or more coders have coded all of the protocols. For two-way data Cohen's (1960) *kappa* or Janson and Olsson's (2001) *iota* are proper chance-corrected agreement measures.

A one-way analysis is proper when different protocols have been coded by different coders (which may vary in number). Proper agreement measures are then Fleiss' (1971; 1981) modifications of kappa for one-way data, and Janson and Olsson's (2004) iota with one-way data.

The *RRU* program allows you to state whether you have one-way or two-way data, and the organization of your files and the choice of agreement measure will follow automatically.

Entering the scorings into computer files

It is common practice in clinical work and research to use an error-correcting program in entering Rorschach scores. If agreement is calculated on data which has been prepared in such a way, the results represent the agreement in the usual case when such error-correction has taken place at the time of entering the data into computer.

An alternative to using an error-corretcting function would be to enter Rorschach codings into a Rorschach scoring program with the error checking feature disabled (some versions of RIAP allowed for this), or directly into a text file. This kind of data would be more representative of hand-written Rorschach codings. (This alternative is probably most often of less interest.)

It might be preferable that codings are entered into computer files by each coder independently and individually. That way, each coder is prompted to take action to the possible errors reported by the program in the way s/he would normally have done, which is probably most representative of usual Rorschach data. If hand scored Sequences of Scores are entered at a research unit, care has to be taken to make explicit guidelines for the staff who enter the codings into computer.

Organizing a table of coding files

Once all of the files are entered into computer, you should list your coding files according to the appropriate data layout. For a two-way data set, the layout might be as follows (entries are coding file names):

	Coder		
Protocol	1	2	3
1	P01C1.r3	P01C2.r3	P01C3.r3
2	P02C1.r3	P02C2.r3	P02C3.r3
3	P03C1.r3	P03C2.r3	P03C3.r3
...	...	...	...

For a one-way data set, the typical layout would look like this:

Protocol	Coder	File name
1	1	P01C1.txt
1	2	P01C2.txt
2	3	P02C3.txt
2	4	P02C4.txt
2	5	P02C5.txt
3	1	P03C1.txt
3	2	P03C2.txt
...	...	...

Using the RRU program. Your files may have any file extension as long as they are files that are readable by *RRU.* The simplest way to tell the *RRU* program what files are for what coder and protocol, is to name your files as in the examples above, and place all of the files in one directory on your computer (this directory should not contain any other files). As in the example, your file name should start with a "P", continue with the protocol number, a "C", and the coder number. (The "C" and coder number may also precede the "P" and protocol number.) Your files can have any extension as long as they are readable by the RRU program. (You might want to use leading zeros in your numbering of protocols and coders--i.e., 01, 02..., rather than 1, 2,... --this will result in the protocols being listed in *RRU* according to the numbers you have assigned.)

Checking responses—What to do when coders have different numbers of responses

A problem that must be dealt with in working with response-level Rorschach agreement data concerns instances when different coders disagree about what constitutes a response. For example, one coder may decide that two people and a butterfly on Card III should be treated as one synthesis response, while a second coder may decide that there was nothing in the verbalization to suggest a meaningful relation between the people and the butterfly and so codes each as separate responses. (This is usually very infrequent in typical samples of protocols.) When coders have disagreed about what constitutes a response, their responses will not match, and *all of the responses involved must be excluded from your sample before calculating agreement*. If there are many codings of one protocol, and only one or two coders are in disagreement with the rest, an alternative way of dealing with the problem is to exclude those coder's codings of the complete protocol from calculations. Regardless of how one chooses to deal with the problem, the number of times a response or a protocol was omitted from calculations should be reported along with the agreement results.

Proposed procedure. Go through all of the protocols, comparing all of the coders' sequence of scores. In the rare case when one coder has divided the verbalizations into into responses that do not correspond to the responses assigned by the other coders, you need to delete, for all of the coders, all of the responses affected. For example, if one coder has coded two people and a butterfly on Card III as two separate responses, and all of the other coders have coded that as one synthesis response, you need to delete all of these responses from the comparison, so that only responses that are quite comparable remain. An alternative to this procedure that may be considered when the number of coders is large and only one or two have diverged on the delimitation of responses for a protocol, is to exclude the complete protocol from those one or two coders from the analysis. Keep a count of how often some responses (or protocols) were deleted. This count should be reported together with the agreement results.

Using the RRU program. The *RRU* program allows you to review the different codings for each protocol, and notifies you if the Card numbers for a certain response do not match among coders. (This check most probably catches the vast majority of such discrepancies). You can then mark as "skipped" the responses you want to skip in the response-level analyses. (The "skipped" responses are not included in the response-level analysis, but they are not deleted from the coding files, and will be included in protocol-level analyses).

Converting Rorschach scorings for single responses to numeric values for computation

Proposed procedure. Using Table 1, convert the Rorschach Comprehensive System scores for each coder's response into numbers that are then placed in a table with 59 columns. For example, the response

Wo M^a.FC.FD_u 2 H,Bt,Hh P 3.0 COP,DR

is converted to the 59 entries

```
1 0 2 1 0 0 3 0 0 0 0 1 0 0 3 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
0 0 0 0 0 1 3.0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 .
```

(Good Human Response and Poor Human Response may or may not be entered as the 60th and 61st column in the table. However, these codes are completely derived from other codes, and should according to my view not be included when calculating multivariate agreement for the Special Scores segment or whole Rorschach responses.) If you are working by hand (which is not recommended because the risk of error is great), enter each response on a separate row in your table. Start by entering the first coder's scores for the first response in Row 1, continue in Row 2 with the second coder's scores for the first response, and so on. Note the Protocol and Coder number for each response.

Using the RRU program. The *RRU* program performs a conversion to numeric values automatically each time your Rorschach coding files are read by the program. (The *RRU* program reads the coding files anew for each operation, so the most recent changes to your coding files are always reflected.) *RRU* accounts agreement for Good Human Response and Poor Human Response, but does not include these scores when calculating multivariate agreement for the Special Scores segment or for whole Rorschach responses. Note that you may use the "Export Data" function to export a numeric data table for use in other programs such as Excel or SPSS.

Table 1 *Coding Scheme for Rorschach Comprehensive System Codings of Single Responses*[a]

Column	Variable	Code	Category
1	Location	1	Whole (*W*)
		2	Common Detail (*D*)
		3	Unusual Detail (*Dd*)
2	Space	0	Absent
		1	Space (*S*)
3	Developmental Quality	1	Synthesized (+)
		2	Ordinary (*o*)
		3	Vague/Synthesized (*v*/+)
		4	Vague (*v*)
4	Human Movement	0	Absent
		1	Active (M^a)
		2	Passive (M^p)
		3	Active-passive ($M^{a\text{-}p}$)
5	Animal Movement	0	Absent
		1	Active (FM^a)
		2	Passive (FM^p)
		3	Active-passive ($FM^{a\text{-}p}$)
6	Inanimate Movement	0	Absent
		1	Active (m^a)
		2	Passive (m^p)
		3	Active-passive ($m^{a\text{-}p}$)
7	Color	0	Absent
		1	Pure Color (*C*)
		2	Color-Form (*CF*)
		3	Form-Color (*FC*)
		4	Color Naming (*Cn*)
8	Achromatic color	0	Absent
		1	Pure Achromatic Color (*C′*)
		2	Achromatic Color-Form (*C′F*)
		3	Achromatic Form-Color (*FC′*)
9	Shading-Texture	0	Absent
		1	Pure Texture (*T*)
		2	Texture-Form (*TF*)
		3	Form-Texture (*FT*)
10	Shading-Dimension	0	Absent
		1	Pure Vista (*V*)
		2	Vista-Form (*VF*)
		3	Form-Vista (*FV*)
11	Shading-Diffuse	0	Absent

		1	Pure Shading (*Y*)
		2	Shading-Form (*YF*)
		3	Form-Shading (*FY*)
12	Form Dimension	0	Absent
		1	Present (*FD*)
13	Reflection	0	Absent
		1	Reflection-Form (*rF*)
		2	Form-Reflection (*Fr*)
14	Pure Form	0	Absent
		1	Present (*F*)
15	Form Quality	0	None
		1	Superior-Overelaborated (+)
		2	Ordinary (*o*)
		3	Unusual (*u*)
		4	Minus (-)
16	Pairs	0	Absent
		1	Present (*2*)
17	Whole Human	0	Absent
		1	Present (*H*)
18	Whole Human, Fictional or Mythological	0	Absent
		1	Present (*(H)*)
19	Human Detail	0	Absent
		1	Present (*Hd*)
20	Human Detail, Fictional or Mythological	0	Absent
		1	Present (*(Hd)*)
21	Human Experience	0	Absent
		1	Present (*Hx*)
22	Whole Animal	0	Absent
		1	Present (*A*)
23	Whole Animal, Fictional or Mythological	0	Absent
		1	Present (*(A)*)
24	Animal Detail	0	Absent
		1	Present (*Ad*)
25	Animal Detail, Fictional or Mythological	0	Absent
		1	Present (*(Ad)*)
26	Anatomy	0	Absent
		1	Present (*An*)
27	Art	0	Absent
		1	Present (*Art*)
28	Anthropology	0	Absent
		1	Present (*Ay*)
29	Blood	0	Absent
		1	Present (*Bl*)
30	Botany	0	Absent
		1	Present (*Bt*)
31	Clothing	0	Absent
		1	Present (*Cg*)
32	Clouds	0	Absent
		1	Present (*Cl*)
33	Explosion	0	Absent
		1	Present (*Ex*)
34	Fire	0	Absent
		1	Present (*Fi*)
35	Food	0	Absent
		1	Present (*Fd*)
36	Geography	0	Absent
		1	Present (*Ge*)
37	Household	0	Absent
		1	Present (*Hh*)
38	Landscape	0	Absent
		1	Present (*Ls*)
39	Nature	0	Absent
		1	Present (*Na*)
40	Science	0	Absent
		1	Present (*Sc*)
41	Sex	0	Absent
		1	Present (*Sx*)
42	X-ray	0	Absent
		1	Present (*Xy*)
43	Idiographic	0	Absent
		1	Present (*Id*)
44	Popular	0	Absent
		1	Present (*P*)
45	Z-Score	0	Absent
		(Value)	(Value)
46	Deviant Verbalization	0	Absent
		1	Level 1 (*DV1*)
		2	Level 2 (*DV2*)
47	Incongruent Combination	0	Absent
		1	Level 1 (*INC1*)
		2	Level 2 (*INC2*)
48	Deviant Response	0	Absent
		1	Level 1 (*DR1*)
		2	Level 2 (*DR2*)
49	Fabulized Combination	0	Absent
		1	Level 1 (*FAB1*)
		2	Level 2 (*FAB2*)
50	Inappropriate Logic	0	Absent
		1	Present (*ALOG*)
51	Contamination	0	Absent
		1	Present (*CONTAM*)
52	Perseveration	0	Absent
		1	Present (*PSV*)
53	Confabulation	0	Absent
		1	Present (*CONFAB*)
54	Abstract Content	0	Absent
		1	Present (*AB*)
55	Aggressive Movement	0	Absent
		1	Present (*AG*)
56	Cooperative Movement	0	Absent
		1	Present (*COP*)
57	Morbid Content	0	Absent
		1	Present (*MOR*)
58	Personal	0	Absent
		1	Present (*PER*)
59	Color Projection	0	Absent
		1	Present (*CP*)

Note. The coding for each response is converted to numeric values by inserting the proper numeric value into each of 59 columns (i.e., one column for each variable listed).

[a] Good Human Representation and Poor Human Representation may be entered as the 60th and 61st column, each with a code of 0 for Absent and 1 for Present. However, these codes are completely derived from other codes in the Comprehensive System. The *RRU* program, like current Rorschach scoring software, calculates these codes directly from the other codes in a response. *RRU* accounts

agreement for Good Human Representation and Poor Human Representation, but does not include these codes when calculating multivariate agreement for Special Scores or whole Rorschach responses.

Calculating response-level agreement for two or more coders

Single coding decisions

Proposed procedure. Agreement for single coding categories is calculated using data from one of the columns in the numeric data table. For example, agreement for the content *Bl* is calculated as the agreement among coders for Variable 29 in the data table, and Form Quality Minus is calculated as the agreement for Category 4 on the 15th variable in the data table. The agreement for absence [or presence] of a variable like *M* (which can take several values if present--Active, Passive, or Active-Passive) can be calculated as the agreement for Category 0 (i.e., *M* Absent) on the 4th variable in the data table.

Using the RRU program. The *RRU* program reports agreement for single coding decisions according to the proposed procedure, as part of the agreement results for response level data.

Multi-category coding variables

Sometimes, it is informative to report agreement for the coding of a categorical variable with several categories, for example Form Quality which can be coded into the five categories None, Minus, Unusual, Ordinary, or Plus (Table 1). Calculating a kappa coefficient for a 5 x 5 crosstable is not different than for a 2 x 2 crosstable; it is the interpretation that differs. The interpretation of the agreement for a multi-category coding variable is how well coders agreed on this task as a whole, but the coefficient is not necessarily informative of the agreement for each of the categories. Which is more appropriate to report depends on the use of the data in a study.

Proposed procedure. When total agreement is considered for a variable that has several possible categories (such as *Form Quality*, which has five possible values: *None*, *Plus*, *Ordinary*, *Unusual*, or *Minus*), data from a single column in the numeric data table is treated as one nominal, multi-category variable in calculating agreement. The basic calculations then involve constructing a two-way multi-category crosstable for Cohen's kappa (see Cohen, 1960), or performing a dummy transformation, calculating agreement separately for each category, and summing terms over categories following Janson and Olsson (2001; 2004).

Using the RRU program. The *RRU* program reports agreement for multi-category variables according to the proposed procedure, as part of the agreement results for response level data.

Response segments

Traditionally, Rorschach agreement has been reported in the form of overall proportion of agreement for the five major categories of Rorschach codes, or "response segments" (i.e., Location and Space, Developmental Quality, Determinants, Form Quality, Pairs, Contents, Populars, Z-Score, and Special Scores). Upon scrutiny, 5 of these 9 "segments" are single variables, either single coding decisions (Pairs; Populars) or multi-category variables (Developmental Quality, Form Quality, and Z-Score). How to calculate *kappa* agreement for these is outlined above. The remaining 4 segments are made up of several coding variables. I propose that you calculate response-segment chance-corrected agreement employing Janson and Olsson's (2001; 2004) multivariate agreement measure. This approach accounts the overall level of agreement on a multivariate coding task as a whole. The measure is based on terms of observed and expected disagreement which have a meaningful interpretation in their own right, being the average number of coding decisions which coders disagree on when codings of same responses, or of any responses, are compared. I suggest that accounting agreement for response-level coding segments using Janson and Olsson's (2001; 2004) multivariate approach, is efficient for communicating the overall level of coding agreement among studies or for training purposes. Janson and Olsson's (2001; 2004) approach has several advantages: It computes chance-corrected agreement that is equivalent to *kappa* and the intraclass correlation (except for a small term which diminishes with increasing sample size). This approach to calculating agreement for response segments (and whole responses) has several advantages over previously used approaches for calculating chance-corrected agreement for complex response segments. In Janson and Olsson's approch, the terms of expected and observed disagreement have straightforward interpretations.

Proposed procedure. Multivariate agreement (using Janson and Olsson's [2001; 2004] iota statistic) for coding segments can be calculated using parts of the table with converted Rorschach data that you have created (see page 15). For the Location and Space segment, treat Columns 1-2 as two nominal variables. For the Determinants segment, treat Columns 4-

14 as 11 nominal variables. For Contents, treat Columns 17-43 as 27 nominal variables. For Special Scores, treat Columns 46-59 as 14 nominal variables.

Using the RRU program. The *RRU* program reports multivariate agreement for response segments according to the proposed procedure, as part of the agreement results for response level data. (For two-way data, the program can optionally also compute "total configuration" statistics based on a crosstabulation of all possible combinations.)

Whole responses

The overall multivariate chance-corrected agreement for whole Rorschach responses provides information about the overall response-level coding precision. The measure is an expression of how well coders agreed on the multivariate coding task as a whole. This measure provides an efficient means for comparing the overall coder agreement among studies, and it can also be very useful in training settings. It should be noted that the multivariate agreement measure does not equal the average of the univariate agreement measures. This is because the terms of expected and observed disagreement are summed separately over variables, and the multivariate agreement calculated on these sums. The multivariate agreement coefficient may as a result be lower or higher than the average of the univariate agreement coefficients.

Proposed procedure. The overall multivariate chance-corrected agreement for whole responses is calculated using the complete vector of 59 nominal variables (Table 1) as data.

Using the RRU program. The *RRU* program reports multivariate agreement for whole Rorschach responses according to the proposed procedure, as part of the agreement results for response level data.

Calculating protocol-level variable agreement for two or more coders

While agreement on the response level conveys important information about the precision of coding at the level at which it was performed, the data that are used in clinical decision making and in research, are variables that are sums over a whole protocol of codes (or derivatives thereof). Protocol-level agreement is of primary interest in any study where protocol-level variables make up a study's results, because it is directly informative about (one facet of) the reliability of the research variables. Many protocol-level Rorschach variables are more or less continuous. While in actuality their distribution often departs to a certain degree from the normal distribution, for the purpose of communicating agreement, a measure of agreement applicable to interval-scale data may most often be used. Intraclass correlations are widely known and accepted reliability measures for interval-level data; intraclass correlations also correspond closely with kappa-type measures. The varieties of intraclass correlations that are most applicable to Rorschach data are the one-way, single-measure, absolute agreement coefficient ($ICC_{(1,1)}$ in the terminology of Fleiss and Shrout [1979]) when observations are by different coders for each case, and the two-way, random model, single-measure, absolute agreement coefficient ($ICC_{(2,1)}$ in the terminology of Fleiss and Shrout [1979]) when observations are by the same set of coders for each case. Formulas for calculating these coefficients from a data tabulated for one variable and two or more coders are given in Appendix 1, and further directions are given for example in Fleiss and Shrout (1979), and in McGraw & Wong (1996).

In many cases, protocol-level Rorschach variables are considered on a categorical level (e.g, HVI Positive; Lambda>.99) and in that case applying a categorical agreement measure of the kappa family is appropriate.

Proposed procedure. Calculate the protocol-level scores for the central variables in your study (most of Rorschach scoring programs allow for data export to a format which can be read by a statistic software package), and set up a data table with separate columns for each coders. Use a statistical software package such as SPSS to calculate intraclass correlations (for two-way interval data, two-way, random, absolute agreement, single measure intraclass correlation, "$ICC_{(2,1)}$" in the notation of Shrout & Fleiss, 1979), or *kappa* (for two-way categorical data). If the coders are not the same set for all of the protocols, you have "one-way" data and need to use "one-way" measures to report agreement. The appropriate intraclass correlation is the one-way, single measure, absolute agreement, intraclass correlation ("$ICC_{(1,1)}$" in the notation of Shrout & Fleiss, 1979; see also McGraw & Wong, 1996). Janson and Olsson's (2004) *kappa* extension is an appropriate form of *kappa* for one-way data.

Using the RRU program. The *RRU* program calculates protocol-level agreement directly from your scoring files for a number of Comprehensive system variables as part of the protocol-level agreement output. You may also use the "Export data" function in *RRU* to export protocol-level data for use in programs such as Excel or SPSS.

References

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

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

Exner, J. E., Jr. (1996). A comment on "The Comprehensive System for the Rorschach: A critical examination." *Psychological Science, 7,* 11-13.

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.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance.* New York: Dryden Press.

Janson, H. (1999). *Projective methods and longitudinal developmental research: Considerations of data's nature and reliability.* Stockholm University: Department of Psychology. [Doctoral dissertation.]

Janson, H., Meyer, G. J., Exner, J. E., Jr., Nakamura, N., & Tuset, A. M. (2002, September). *Computing Rorschach agreement among many coders on various data levels: An illustration with three international data sets.* Paper presented at the XVII International Congress of Rorschach and Projective Methods, Rome, Italy. September 2002

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

Janson, H., & Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement, 64,* 62-70.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Tutorial in biostatistics: Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109-2129.

Lis, A., Parolin, L., Calvo, V., Zennaro, A., & Meyer, G. (2007). The impact of administration and inquiry on Rorschach Comprehensive System protocols in a national reference sample. *Journal of Personality Assessment, 89*(S1), S193-S200.

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.

Messick, S. J. (1989: Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 11-103). New York: Macmillan.

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

Meyer, G. (1997b). Thinking clearly about reliability: More critical corrections regarding the Rorschach comprehensive system. *Psychological Assessment, 9,* 495-498.

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

Pedhazur, E., Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Shaffer, T. W., Erdberg, P., & Meyer, G. J. (Eds.). (2007). International reference samples for the Rorschach Comprehensive System [Special issue]. *Journal of Personality Assessment*, 89(Supplement 1).

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

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

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

Weiner, I. B. (1991). Editor's Note: Interscorer Agreement in Rorschach Research. *Journal of Personality Assessment, 56,* 1.

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

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996b). Thinking critically about the Comprehensive System for the Rorschach: A reply to Exner. *Psychological Science, 7,* 14-17.

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

Appendix 1. Formulas for agreement measures

Intraclass correlations

Two-way data. Intraclass correlations are a family of measures that estimate the proportion of variance in a variable not due to measurement error. There exist several forms of intraclass correlations, depending on the nature of the data. When the same set of coders have scored all of the objects in the sample, this is labeled "two-way" data. The term "two-way" comes from ANOVA terminology and reflects the fact that a two-way ANOVA is performed in order to obtain the terms for the intraclass correlation (with observations as one factor and coders as the other factor). With "two-way" data, the measure which estimates the reliability of a single coder's scoring is the two-way, random, absolute agreement, single-measure intraclass correlation, $ICC_{(2,1)}$ in the notation of Shrout and Fleiss (1979). This coefficient estimates the proportion of variance in the population that is due to variance between observations, out of the total variance, which is composed of variance between observations, between judges, and the interaction among judges and observations:

(1) $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2)$

where the subscripts to the variance terms stand for: B=between observations, J=between judges, I=observation×judge interaction, and E=error (Shrout & Fleiss, 1979). The classical formulations for estimating this coefficient uses mean square terms from a two-way ANOVA with clients (cases) and coders as factors. The resulting Mean Square Between (*MSB*), Mean Square within Judges (i.e., coders, *MSJ*), and Mean Square Error (*MSE*) into the following formula where *b* is the number of coders, and *n* is the number of clients (cases):

(2) $$ICC_{(2,1)} = \frac{MSB - MSE}{MSB + (b-1)MSE + b(MSJ - MSE)/n}$$

In SPSS, intraclass correlations may be obtained as part of the Reliability procedure. If there are two coders whose scores for one variable are stored in the variables M1 and M2, the syntax for obtaining a two-way, random intraclass correlation coefficient is:

```
reliability
 /variables=M1 M2
 /statistics=anova
 /icc=model(random) type(absolute) cin=95 testval=0 .
```

Today, several possibilities for estimating variance components are available even in common statistical software packages like SPSS. These possibilities include various estimation options (including Maximum Likelihood rather than ordinary least squares), and missing data options. Provided that you have each observation on a new row in your data file, you may use, for example, the following SPSS syntax in order to obtain estimates the variance components that the classical formula is based on:

```
VARCOMP
 M  BY Protocol Coder
 /RANDOM = Protocol Coder
 /METHOD = MINQUE (1)
 /DESIGN
 /INTERCEPT = INCLUDE.
```

You must then take the variance components estimates from the SPSS output, and insert the estimates into the formula (1) above to calculate the intraclass correlation. The resulting estimate may differ somewhat from the classical ANOVA-based intraclass correlation, depending on the options you choose for the variance components estimation. (The syntax example specifies MINQUE estimation--the default in SPSS--which may not result in exactly the same coefficient as the classical formulation.)

One-way data. When there are several coders and different objects have been coded by different coders, the appropriate intraclass correlation is the "one-way" intraclass correlation. The terms in this variety of correlation are obtained from a one-way ANOVA (with clients/cases as the factor). With "one-way" data, the measure which estimates the reliability of a single coder's scoring is the two-way, random, absolute agreement, single-measure intraclass correlation, $ICC_{(1,1)}$ in the notation of Shrout and Fleiss (1979). The coefficient estimates the ratio of agreement

(3) $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_W^2)$

where the subscripts to the variance terms stand for: B=between observations and W=within observations. In order to obtain this coefficient, perform a one-way ANOVA with responses (or protocols, as the case may be) as the one factor, and insert the resulting Mean Square Between (*MSB*) and Mean Square Within (*MSW*) terms into the following formula where *k* is the number of coders:

(4) $$ICC_{(1,1)} = \frac{MSB - MSW}{MSB + (b-1)MSW},$$

where *b* is the number of coders. It is easily recognized that this formula only is defined when the number of coders, *b*, is the same for all the responses (or protocols). The one-way intraclass correlation is defined for varying numbers of coders for different objects (see e.g., Haggard, 1958 for a form of the $ICC_{(1,1)}$ which accommodates for this). However, the Reliability procedure in SPSS only allows to compute intraclass correlations when the number of coders is the same for all observations. If the codings of three coders per client (not the same for all of the clients) are stored in the variables M1, M2, and M3, the SPSS syntax for obtaining a one-way intraclass correlation coefficient is:

```
reliability
 /variables=M1 M2 M3
 /format=nolabels
 /icc=model(oneway) cin=95 testval=0 .
```

You may also use variance components estimation procedures in SPSS to obtain an equivalent of the one-way intraclass correlation. This option will allow varying numbers of judges per object. Provided that you have each observation on a new row in your data file, you may use, for example, the following SPSS syntax in order to obtain a one-way intraclass correlation that estimates the relevant variance components:

```
VARCOMP
  M  BY Protocol
  /RANDOM = Protocol
  /METHOD = SSTYPE (3)
  /DESIGN
  /INTERCEPT = INCLUDE .
```

You must then take the variance components estimates from the SPSS output, and insert the estimates into the formula (3) above to calculate the intraclass correlation. Agan, the resulting estimate may differ somewhat from the classical ANOVA-based correlation, depending on the options you choose for the variance components estimation. (The syntax example specifies ANOVA estimation, which should result in exactly the same estimate as the classical formulation in e.g., Haggard [1958].)

Cohen's *kappa*

Cohen's (1960) *kappa* is the commonly used chance-corrected agreement measure for nominal data. It is defined for dichotomous and polytomous variables. Cohen's original measure applies to two-way data (with a set of two coders who have coded all of the observations). Cohen's *kappa* is calculated from two terms. The first term is p_o , the observed proportion of agreement (i.e. the same as the overall proportion of agreement, or per cent agreement, which should always include agreements for presence as well as for absence of a code). The second term is an expression of chance agreement in the sample. This second term, p_e , the expected agreement, is calculated from the (marginal) distributions of the categories over the two coders. For example, if one coder categorized a proportion of .30 of the cases as Present, and the other coder categorized a proportion of .40 cases as Present, the chance expectation for a coding of Present is .30×.40, or .12. As for the observed agreement term, chance expectations for Present and Absent are summed to the total p_e term. The p_o and p_e terms are then inserted into the equation

(5) $$\kappa = \frac{p_o - p_e}{1 - p_e}.$$

The SPSS syntax for obtaining kappa statistics in a two-way crosstable of two coders' codings, for example if one coder's coding is in the variable H1, and the other coders' coding is in the variable H2, is:

```
crosstabs
 /tables = H1 by H2
 /cells = count total
 /statistics=kappa .
```

Janson and Olsson's *iota*

Janson and Olsson's (2001) extension of *kappa*, *iota,* builds on a mathematically equivalent reformulation of *kappa* (see pages 7 and following). Instead of observed and expected proportions of agreement, as in *kappa*, Janson and Olsson's extension builds on terms of observed *dis*agreement (d_o) and expected *dis*agreement (d_e). This reformulation extends *kappa* to use with several observers and interval or nominal multivariate data.

Observed disagreement, d_o, is defined as the amount of disagreement, or multivariate distance, between two observations of the same object.

Expected disagreement, d_e, is defined as the amount of disagreement, or multivariate distance, between two observations of any objects.

The chance-corrected agreement measure is calculated as

$$(6) \quad \iota = 1 - \frac{d_o}{d_e} .$$

The terms of observed and expected disagreement may be obtained by taking the average of all possible comparisons among observations. However, computational formula that make use of ANOVA terms are mathematically equivalent, and require many fewer calculations.

For two-way data, the Sum of Squares Total (SS_T) Sum of Squares Within (SS_W), and Sum of Squares between Judges (i.e., Coders) (SS_J) terms from from *c* (the number of variables) two-way ANOVAs with clients and coders as factors are inserted into the equations

$$(7) \quad d_o = \left[n\binom{b}{2} \right]^{-1} b \sum_{k=1}^{c} SS_{W_k}$$

and

$$(8) \quad d_e = \left[n\binom{b}{2} \right]^{-1} \sum_{k=1}^{c} \left[(b-1) SS_{T_k} + SS_{J_k} \right] ,$$

where *b* is the number of coders and *n* is the number of clients, to obtain the terms for *iota* (Janson & Olsson, 2001).

For one-way data, the Sum of Squares Total (SS_T) and Sum of Squares Within (SS_W) terms from from *c* (the number of variables) one-way, unequal *n*s ANOVAs with responses (or protocols, as the case may be) as the one factor are inserted into the equations

$$(9) \quad d_o = \frac{2}{n-t} \sum_{k=1}^{c} SS_{W_k}$$

and

$$(10) \quad d_e = \frac{2}{n} \sum_{k=1}^{c} SS_{T_k} \quad ,$$

where *t* is the number of responses/protocols, and *n* is the total number of observations (Janson & Olsson, 2004).

iota for two-way data (Janson & Olsson, 2001) exactly equals *kappa* with one categorical variable. The two-way form of *iota* equals the two-way intraclass correlation for one interval or dichotomous variable, except for a small term which derives from the use of *n* as degrees of freedom for cases in *iota*, but *n-1* in intraclass correlations (the difference is unimportant whenever the number of observations is at all large, e.g., >20). Similarly, the one-way form of *iota* equals the one-way intraclass correlation for one interval or dichotomous variable, except for a small term.