Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(1), p. 23-34.

Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial

Kevin A. Hallgren
University of New Mexico

Many research designs require the assessment of inter-rater reliability (IRR) to demonstrate consistency among observational ratings provided by multiple coders. However, many studies use incorrect statistical procedures, fail to fully report the information necessary to interpret their results, or do not address how IRR affects the power of their subsequent analyses for hypothesis testing. This paper provides an overview of methodological issues related to the assessment of IRR with a focus on study design, selection of appropriate statistics, and the computation, interpretation, and reporting of some commonly-used IRR statistics. Computational examples include SPSS and R syntax for computing Cohen's kappa and intra-class correlations to assess IRR.

The assessment of inter-rater reliability (IRR, also called inter-rater agreement) is often necessary for research designs where data are collected through ratings provided by trained or untrained coders. However, many studies use incorrect statistical analyses to compute IRR, misinterpret the results from IRR analyses, or fail to consider the implications that IRR estimates have on statistical power for subsequent analyses.

This paper will provide an overview of methodological issues related to the assessment of IRR, including aspects of study design, selection and computation of appropriate IRR statistics, and interpreting and reporting results. Computational examples include SPSS and R syntax for computing Cohen's kappa for nominal variables and intra-class correlations (ICCs) for ordinal, interval, and ratio variables. Although it is beyond the scope of the current paper to provide a comprehensive review of the many IRR statistics that are available, references will be provided to other IRR statistics suitable for designs not covered in this tutorial.

A Primer on IRR

The assessment of IRR provides a way of quantifying the degree of agreement between two or more coders who make independent ratings about the features of a set of subjects. In this paper, *subjects* will be used as a generic term for the people, things, or events that are rated in a study, such as the number of times a child reaches for a caregiver, the level of empathy displayed by an interviewer, or the presence or absence of a psychological diagnosis. *Coders* will be used as a generic term for the individuals who assign ratings in a study, such as trained research assistants or randomly-selected participants.

In classical test theory (Lord, 1959; Novick, 1966), observed scores (X) from psychometric instruments are thought to be composed of a true score (T) that represents the subject's score that would be obtained if there were no measurement error, and an error component (E) that is due to measurement error (also called noise), such that

$$\text{Observed Score} = \text{True Score} + \text{Measurement Error},$$

or in abbreviated symbols,

$$X = T + E. \quad (1)$$

Equation 1 also has the corresponding equation

$$Var(X) = Var(T) + Var(E), \quad (2)$$

where the variance of the observed scores is equal to the variance of the true scores plus the variance of the measurement error, if the assumption that the true scores and errors are uncorrelated is met.

Measurement error (E) prevents one from being able to observe a subject's true score directly, and may be introduced by several factors. For example, measurement error may be introduced by imprecision, inaccuracy, or poor

scaling of the items within an instrument (i.e., issues of internal consistency); instability of the measuring instrument in measuring the same subject over time (i.e., issues of test-retest reliability); and instability of the measuring instrument when measurements are made between coders (i.e., issues of IRR). Each of these issues may adversely affect reliability, and the latter of these issues is the focus of the current paper.

IRR analysis aims to determine how much of the variance in the observed scores is due to variance in the true scores after the variance due to measurement error between coders has been removed (Novick, 1966), such that

$$\begin{aligned} \text{Reliability} &= \frac{Var(T)}{Var(X)} = \frac{Var(X) - Var(E)}{Var(X)} \\ &= \frac{Var(T)}{Var(T) + Var(E)} \end{aligned} . \quad (3)$$

For example, an IRR estimate of 0.80 would indicate that 80% of the observed variance is due to true score variance or similarity in ratings between coders, and 20% is due to error variance or differences in ratings between coders.

Because true scores (*T*) and measurement errors (*E*) cannot be directly accessed, the IRR of an instrument cannot be directly computed. Instead, true scores can be estimated by quantifying the covariance among sets of observed scores (*X*) provided by different coders for the same set of subjects, where it is assumed that the shared variance between ratings approximates the value of $Var(T)$ and the unshared variance between ratings approximates $Var(E)$, which allows reliability to be estimated in accordance with equation 3.

IRR analysis is distinct from validity analysis, which assesses how closely an instrument measures an actual construct rather than how well coders provide similar ratings. Instruments may have varying levels of validity regardless of the IRR of the instrument. For example, an instrument may have good IRR but poor validity if coders' scores are highly similar and have a large shared variance but the instrument does not properly represent the construct it is intended to measure.

How are studies designed to assess IRR?

Before a study utilizing behavioral observations is conducted, several design-related considerations must be decided *a priori* that impact how IRR will be assessed. These design issues are introduced here, and their impact on computation and interpretation are discussed more thoroughly in the computation sections below.

First, it must be decided whether a coding study is designed such that all subjects in a study are rated by multiple coders, or if a subset of subjects are rated by multiple coders with the remainder coded by single coders. The contrast between these two options is depicted in the left and right columns of Table 1. In general, rating all subjects is acceptable at the theoretical level for most study designs. However, in studies where providing ratings is costly and/or time-intensive, selecting a subset of subjects for IRR analysis may be more practical because it requires fewer overall ratings to be made, and the IRR for the subset of subjects may be used to generalize to the full sample.

Second, it must be decided whether the subjects that are rated by multiple coders will be rated by the same set of coders (fully crossed design) or whether different subjects are rated by different subsets of coders. The contrast between these two options is depicted in the upper and lower rows of Table 1. Although fully crossed designs can require a higher overall number of ratings to be made, they allow for systematic bias between coders to be assessed and controlled for in an IRR estimate, which can improve overall IRR estimates. For example, ICCs may underestimate the true reliability for some designs that are not fully crossed, and researchers may need to use alternative statistics that are not well distributed in statistical software packages to assess IRR in some studies that are not fully crossed (Putka, Le, McCloy, & Diaz, 2008).

Third, the psychometric properties of the coding system used in a study should be examined for possible areas that could strain IRR estimates. Naturally, rating scales already shown to have poor IRR are likely to produce low IRR estimates in subsequent studies. However, even if a rating system has been shown to have good IRR, restriction of range can potentially occur when a rating system is applied to new populations, which can substantially lower IRR estimates. Restriction of range often lowers IRR estimates because the $Var(T)$ component of equation 3 is reduced, producing a lower IRR estimate even if $Var(E)$ does not change. For example, consider two hypothetical studies where coders rate therapists' levels of empathy on a well-validated 1 to 5 Likert-type scale where 1 represents very low empathy and 5 represents very high empathy. The first study recruits therapists from a community clinic and results in a set of ratings that are normally distributed across the five points of the scale, and IRR for empathy ratings is good. The second study uses the same coders and coding system as the first study and recruits therapists from a university clinic who are highly trained at delivering therapy in an empathetic manner, and results in a set of ratings that are restricted to mostly 4's and 5's on the scale, and IRR for empathy ratings is low. IRR is likely to have been reduced due to restriction of range where $Var(T)$ was reduced in the second study even though $Var(E)$ may have been similar between studies because the same coders and

Table 1.*Designs for assigning coders to subjects IRR studies.*

		All subjects rated by multiple coders				Subset of subjects rated by multiple coders		
		Coder A	Coder B	Coder C		Coder A	Coder B	Coder C
Design fully crossed	Subject 1	X	X	X	Subject 1	X	X	X
	Subject 2	X	X	X	Subject 2	X		
	Subject 3	X	X	X	Subject 3	X	X	X
	Subject 4	X	X	X	Subject 4		X	
		Coder A	Coder B	Coder C		Coder A	Coder B	Coder C
Design not fully crossed	Subject 1		X	X	Subject 1	X	X	
	Subject 2	X		X	Subject 2	X		
	Subject 3		X	X	Subject 3		X	X
	Subject 4	X	X		Subject 4		X	

Note: "X" indicates that the ratings were provided by a given coder to the corresponding subject.

coding system were used. In cases where restricted range is likely, it is worth considering whether the scale should be modified, for example by expanding it into a 1 to 9 Likert-type scale, adjusting the anchoring points, or omitting the scale altogether. These decisions are best made before a study begins, and pilot testing may be helpful for assessing the suitability of new or modified scales.

Fourth, in studies using trained coders, it may often be necessary to conduct a considerable amount of training with practice subjects before subjects from the real study are coded. In these cases it is common to specify an *a priori* level of IRR that must be achieved before subjects from the real study are rated and to report this in the final study write-up. Commonly, the qualitative ratings for different IRR statistics can be used to assign these cutoff points; for example, a researcher may require all IRR estimates to be at least in the "good" range before coders can rate the real subjects in a study.

What are common mistakes that people make in assessing and reporting IRR?

Most general courses in statistics and experimental design devote little or no time to the study of IRR, which, combined with the lack of published comprehensive guidelines for assessing and reporting IRR, may result in several commonly-made mistakes in behavioral research. Several of these mistakes are briefly described below.

Using percentages of agreement. Despite being definitively rejected as an adequate measure of IRR (Cohen, 1960; Krippendorff, 1980), many researchers continue to report the percentage that coders agree in their ratings as an index of coder agreement. For categorical data, this may be expressed as the number of agreements in observations divided by the total number of observations. For ordinal, interval, or ratio data where close-but-not-perfect agreement may be acceptable, percentages of agreement are sometimes expressed as the percentage of ratings that are in agreement within a particular interval. Perhaps the biggest criticism of percentages of agreement is that they do not correct for agreements that would be expected by chance and therefore overestimate the level of agreement. For example, if coders were to randomly rate 50% of subjects as "depressed" and 50% as "not depressed" without regard to the subject's actual characteristics, the expected percentage of agreement would be 50% even though all overlapping ratings were due to chance. If coders randomly rated 10% of subjects as depressed and 90% as not depressed, the expected percentage of agreement would be 82% even though this seemingly high level of agreement is still due entirely to chance.

Not reporting which statistic or variant was used in an IRR analysis. Many studies fail to report which statistic was used to compute IRR (e.g., Cohen's kappa, Fleiss's kappa, ICCs) or which variant of that statistic was computed (e.g., Siegel & Castellan's 1988 variant of Cohen's kappa, two-way consistency average-measures ICC). Reporting both the statistic and its computational variant are crucial because there are many statistics for computing IRR and different

Table 2. Agreement matrix for nominal variable.

		Coder A		
		Absent	Present	*Total*
Coder B	Absent	42	13	55
	Present	8	37	45
	Total	50	50	100

variants can substantially influence the interpretation of IRR estimates. Reference manuals for statistical software packages typically will provide references for the variants of IRR statistics that are used for computations, and some software packages allow users to select which variant they wish to compute.

Not using the correct statistic for the study design. Many factors must be considered in the selection of the most appropriate statistical test, such as the metric in which a variable was coded (e.g., nominal vs. ordinal, interval, or ratio), the design of the study (e.g., whether all subjects vs. a subset of subjects are rated by multiple coders), and the intended purpose of the IRR estimate (e.g., to estimate the reliability of individual coders' ratings vs. the reliability of the mean ratings from multiple coders). Researchers should be careful to assess the appropriateness of a statistic for their study design and look for alternative options that may be more suitable for their study. Appropriate statistics for various study designs are discussed in more depth in the computation sections below.

Not performing IRR analyses on variables in their final transformed form. It is often more appropriate to report IRR estimates for variables in the form that they will be used for model testing rather their raw form. For example, if a researcher counts the frequency of certain behaviors then square-root transforms these for use in subsequent hypothesis testing, assessing IRR for the transformed variables, rather than the raw behavior counts, more accurately indicates the relative level of measurement error that is present in the final hypothesis testing. In situations where IRR estimates are high for a variable in its raw form but low for the variable in its final form (or *vice versa*), both IRR estimates may be reported to demonstrate that coders reliably rated subjects, despite the IRR for the final variable being low and possibly containing too much measurement error for further analysis.

Not interpreting the effect of IRR on power and pertinent study questions. Finally, many researchers neglect to interpret the effect of IRR estimates on questions of interest to their study. For example, if it is important to show that coders can independently reach similar conclusions about the subjects they observe, it can be helpful to provide qualitative interpretations of IRR estimates by comparing them to previously-observed IRR estimates from similar instruments or providing qualitative ratings based on pre-established cutoff points for good, acceptable, and unacceptable IRR.

Implications of IRR estimates on statistical power should be commented on if the variables observed in the study are subject to subsequent hypothesis testing. Low IRR indicates that the observed ratings contain a large amount of measurement error, which adds noise to the signal a researcher wishes to detect in their hypothesis tests. Low IRR may increase the probability of type-II errors, as the increase in noise may suppress the researcher's ability to detect a relationship that actually exists, and thus lead to false conclusions about the hypotheses under study.

Possible reasons for low IRR should be discussed, e.g., IRR may be low due to restricted range, poor psychometric properties of a scale, poorly trained coders, difficulty in observing or quantifying the construct of interest, or other reasons. Decisions about dropping or retaining variables with low IRR from analyses should be discussed, and alternative models may need to be proposed if variables are dropped.

Computing IRR

Kappa for Nominal Variables

Cohen's (1960) kappa and related kappa variants are commonly used for assessing IRR for nominal (i.e., categorical) variables. Different variants of kappa allow for IRR to be assessed in fully-crossed and non-fully crossed designs.

Mathematical foundations. Kappa statistics measure the observed level of agreement between coders for a set of nominal ratings and corrects for agreement that would be expected by chance, providing a standardized index of IRR that can be generalized across studies. The degree of observed agreement is determined by cross-tabulating ratings for two coders, and the agreement expected by chance is determined by the marginal frequencies of each coder's ratings. Kappa is computed based on the equation

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \quad (4)$$

where *P(a)* denotes the observed percentage of agreement, and *P(e)* denotes the probability of expected agreement due to chance. To illustrate the derivation of *P(a)* and *P(e)*, Table 2 provides hypothetical data from two coders who make one of two response options for each subject (e.g., the presence or absence of depression). For the data in Table 2, *P(a)* is equal to the observed percentage of agreement, indicated by the sum of the diagonal values divided by the total number of subjects, (42+37)/100 = .79. To compute *P(e)*, we note from

the marginal means of Table 2 that Coder A rated depression as present 50/100 times and Coder B rated depression as present 45/100 times. The probability of obtaining agreement about the presence of depression if ratings were assigned randomly between coders would be 0.50 × 0.45 = 0.225, and the probability of obtaining chance agreement about the absence of depression would be (1-0.50) × (1-0.45) = 0.275. The total probability of any chance agreement would then be 0.225 + 0.275 = 0.50, and κ = (0.79 - 0.50)/(1 - 0.50) = 0.58.

Possible values for kappa statistics range from -1 to 1, with 1 indicating perfect agreement, 0 indicating completely random agreement, and -1 indicating "perfect" disagreement. Landis and Koch (1977) provide guidelines for interpreting kappa values, with values from 0.0 to 0.2 indicating slight agreement, 0.21 to 0.40 indicating fair agreement, 0.41 to 0.60 indicating moderate agreement, 0.61 to 0.80 indicating substantial agreement, and 0.81 to 1.0 indicating almost perfect or perfect agreement. However, the use of these qualitative cutoffs is debated, and Krippendorff (1980) provides a more conservative interpretation suggesting that conclusions should be discounted for variables with values less than 0.67, conclusions tentatively be made for values between 0.67 and 0.80, and definite conclusions be made for values above 0.80. In practice, however, kappa coefficients below Krippendorff's conservative cutoff values are often retained in research studies, and Krippendorff offers these cutoffs based on his own work in content analysis while recognizing that acceptable IRR estimates will vary depending on the study methods and the research question.

Common kappa variants for 2 coders. Cohen's original (1960) kappa is subject to biases in some instances and is only suitable for fully-crossed designs with exactly two coders. As a result, several variants of kappa have been developed that accommodate different datasets. The chosen kappa variant substantially influences the estimation and interpretation of IRR coefficients, and it is important that researchers select the appropriate statistic based on their design and data and report it accordingly. Including full mathematical expositions of these variants is beyond the scope of the present article but they are available in the references provided.

Two well-documented effects can substantially cause Cohen's kappa to misrepresent the IRR of a measure (Di Eugenio & Glass, 2004, Gwet, 2002), and two kappa variants have been developed to accommodate these effects. The first effect appears when the marginal distributions of observed ratings fall under one category of ratings at a much higher rate over another, called the prevalence problem, which typically causes kappa estimates to be unrepresentatively low. Prevalence problems may exist within a set of ratings due to the nature of the coding system used in a study, the tendency for coders to identify one or more categories of behavior codes more often than others, or due to truly unequal frequencies of events occurring within the population under study. The second effect appears when the marginal distributions of specific ratings are substantially different between coders, called the bias problem, which typically causes kappa estimates to be unrepresentatively high. Di Eugenio and Glass (2004) show how two variants of Cohen's (1960) kappa (Byrt, Bishop, & Carlin, 1993; Siegel & Castellan, 1988, pp. 284-291) may be selected based on problems of prevalence and bias in the marginal distributions. Specifically, Siegel and Castellan's kappa obtains accurate IRR estimates in the presence of bias, whereas Cohen's and Byrt et al's kappa estimates are inflated by bias and therefore not preferred when bias is present. Alternatively, Byrt et al.'s formula for kappa corrects for prevalence, whereas Cohen's and Siegel and Castellan's kappa estimates are unrepresentatively low when prevalence effects are present and may not be preferred if substantial prevalence problems are present. No single kappa variant corrects for both bias and prevalence, and therefore multiple kappa variants may need to be reported to account for each of the different distributional problems that are present within a sample.

Cohen (1968) provides an alternative weighted kappa that allows researchers to differentially penalize disagreements based on the magnitude of the disagreement. Cohen's weighted kappa is typically used for categorical data with an ordinal structure, such as in a rating system that categorizes high, medium, or low presence of a particular attribute. In this case a subject being rated as high by one coder and low by another should result in a lower IRR estimate than when a subject is rated as high by one coder and medium by another. Norman and Streiner (2008) show that using a weighted kappa with quadratic weights for ordinal scales is identical to a two-way mixed, single-measures, consistency ICC, and the two may be substituted interchangeably. This interchangeability poses a specific advantage when three or more coders are used in a study, since ICCs can accommodate three or more coders whereas weighted kappa can only accommodate two coders (Norman & Streiner, 2008).

Common kappa-like variants for 3 or more coders. The mathematical foundations of kappa provided by Cohen (1960) make this statistic only suitable for two coders, therefore IRR statistics for nominal data with three or more coders are typically formalized as extensions of Scott's (1955) Pi statistic (e.g., Fleiss's 1971) or are computed using the arithmetic mean of kappa or *P(e)* (e.g., Light 1971; Davies

Table 3. Hypothetical nominal depression ratings for kappa example.

Subject	Dep_Rater1	Dep_Rater2	Dep_Rater3
1	1	0	1
2	0	0	0
3	1	1	1
4	0	0	0
5	0	0	0
6	1	1	2
7	0	1	1
8	0	2	0
9	1	0	1
10	0	0	0
11	2	2	2
12	2	2	2

& Fleiss, 1982).

Fleiss (1971) provides formulas for a kappa-like coefficient that is suitable for studies where any constant number of *m* coders is randomly sampled from a larger population of coders, with each subject rated by a different sample of *m* coders. For example, this may be appropriate in a study where psychiatric patients are assigned as having (or not having) a major depression diagnosis by several health professionals, where each patient is diagnosed by *m* health professionals randomly sampled from a larger population. Gross (1986) provides formulas for a statistic similar to Fleiss's kappa for studies with similar designs when the number of coders in the study is large relative to the number of subjects. In accordance with the assumption that a new sample of coders is selected for each subject, Fleiss's coefficient is inappropriate for studies with fully-crossed designs.

For fully-crossed designs with three or more coders, Light (1971) suggests computing kappa for all coder pairs then using the arithmetic mean of these estimates to provide an overall index of agreement. Davies and Fleiss (1982) propose a similar solution that uses the average *P(e)* between all coder pairs to compute a kappa-like statistic for multiple coders. Both Light's and Davies and Fleiss's solutions are unavailable in most statistical packages; however, Light's solution can easily be implemented by computing kappa for all coder pairs using statistical software then manually computing the arithmetic mean.

A summary of the kappa and kappa-like statistical variants discussed here is outlined in Table 7.

Computational example. A brief example for computing kappa with SPSS and the R *concord* package (Lemon & Fellows, 2007) are provided based on the hypothetical nominal ratings of depression in Table 3, where "2" indicates current major depression, "1" indicates a history of major depression but no current diagnosis, and "0" indicates no history of or current major depression. Although not discussed here, the R *irr* package (Gamer, Lemon, Fellows, & Singh, 2010) includes functions for computing weighted Cohen's (1968) kappa, Fleiss's (1971) kappa, and Light's (1971) average kappa computed from Siegel & Castellan's variant of kappa, and the user is referred to the *irr* reference manual for more information (Gamer et al., 2010).

SPSS and R both require data to be structured with separate variables for each coder for each variable of interest, as shown for the depression variable in Table 3. If additional variables were rated by each coder, then each variable would have additional columns for each coder (e.g., Rater1_Anxiety, Rater2_Anxiety, etc.), and kappa must be computed separately for each variable. Datasets that are formatted with ratings from different coders listed in one column may be reformatted by using the *VARSTOCASES* command in SPSS (see tutorial provided by Lacroix & Giguère, 2006) or the *reshape* function in R.

A researcher should specify which kappa variant should be computed based on the marginal distributions of the observed ratings and the study design. The researcher may consider reporting Byrt et al.'s (1983) prevalence-adjusted kappa or Siegel & Castellan's (1988) bias-adjusted kappa if prevalence or bias problems are strong (Di Eugenio & Glass, 2004). Each of these kappa variants is available in the R *concord* package; however, SPSS only computes Siegel & Castellan's kappa (Yaffee, 2003).

The marginal distributions for the data in Table 3 do not suggest strong prevalence or bias problems; therefore, Cohen's kappa can provide a sufficient IRR estimate for each coder pair. Since three coders are used, the researcher will likely wish to compute a single kappa-like statistic that summarizes IRR across all coders by computing the mean of kappa for all coder-pairs (Light, 1971). Syntax for computing kappa for two coders in SPSS and the R *concord* package are provided in Table 4, and the syntax may be modified to calculate kappa for all coder pairs when three or more coders are present. Both procedures provide point estimates and significance tests for the null hypothesis that $\kappa = 0$. In practice, only point estimates are typically reported, as significance test are expected to indicate that kappa is greater than 0 for studies that use trained coders (Davies & Fleiss, 1982).

The resulting estimate of Cohen's kappa averaged across coder pairs is 0.68 (coder pair kappa estimates = 0.62 [coders 1 and 2], 0.61 [coders 2 and 3], and 0.80 [coders 1 and 3]),

Table 4. Syntax for computing kappa in SPSS and R

SPSS Syntax

```
CROSSTABS
  /TABLES=Dep_Rater1 BY Dep_Rater2        'select the two variables to compute kappa
  /FORMAT=AVALUE TABLES
  /STATISTICS=KAPPA
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

R Syntax

```
library(concord)                    #Load the concord library (must already be installed)
print(table(myRatings[,1]))         #Examine marginal distributions of coder 1 for bias and
                                    #prevalence problems
print(table(myRatings [,2]))        #Examine marginal distributions of coder 2
print(cohen.kappa(myRatings[,1:2])) #compute kappa estimate
```

Note: R syntax assumes that data are in a matrix or data frame called "myRatings." SPSS syntax will compute Siegel and Castellan's (1988) kappa only. R syntax will compute kappa statistics based on Cohen (1960), Siegel and Castellan (1988), and Byrt et al. (1993).

indicating substantial agreement according to Landis and Koch (1977). In SPSS, only Siegel and Castellan's kappa is provided, and kappa averaged across coder pairs is 0.56, indicating moderate agreement (Landis & Koch, 1977). According to Krippendorff's (1980) more conservative cutoffs, the Cohen's kappa estimate suggests that tentative conclusions about the fidelity of the coding may be made, whereas the Siegel & Castellan's kappa estimate suggests that such conclusions should be discarded. Reporting of these results should detail the specifics of the kappa variant that was chosen, provide a qualitative interpretation of the estimate, and describe any implications the estimate has on statistical power. For example, the results of this analysis may be reported as follows:

> *An IRR analysis was performed to assess the degree that coders consistently assigned categorical depression ratings to subjects in the study. The marginal distributions of depression ratings did not indicate prevalence or bias problems, suggesting that Cohen's (1960) kappa was an appropriate index of IRR (Di Eugenis & Glass, 2004). Kappa was computed for each coder pair then averaged to provide a single index of IRR (Light, 1971). The resulting kappa indicated substantial agreement, κ = 0.68 (Landis & Koch, 1977), and is in line with previously published IRR estimates obtained from coding similar constructs in previous studies. The IRR analysis suggested that coders had substantial agreement in depression ratings, although the variable of interest contained a modest amount of error variance due to differences in subjective ratings given by coders, and therefore statistical power for subsequent analyses may be modestly reduced, although the ratings were deemed as adequate for use in the hypothesis tests of the present study.*

ICCs for Ordinal, Interval, or Ratio Variables

The intra-class correlation (ICC) is one of the most commonly-used statistics for assessing IRR for ordinal, interval, and ratio variables. ICCs are suitable for studies with two or more coders, and may be used when all subjects in a study are rated by multiple coders, or when only a subset of subjects is rated by multiple coders and the rest are rated by one coder. ICCs are suitable for fully-crossed designs or when a new set of coders is randomly selected for each participant. Unlike Cohen's (1960) kappa, which quantifies IRR based on all-or-nothing agreement, ICCs incorporate the magnitude of the disagreement to compute IRR estimates, with larger-magnitude disagreements resulting in lower ICCs than smaller-magnitude disagreements.

Mathematical foundations. Different study designs necessitate the use of different ICC variants, but all ICC variants share the same underlying assumption that ratings from multiple coders for a set of subjects are composed of a true score component and measurement error component. This can be rewritten from equation 1 in the form

$$X_{i,j} = \mu + r_i + e_{i,j} \quad (5)$$

where $X_{i,j}$ is the rating provided to subject *i* by coder *j*, μ is the mean of the true score for variable *X*, r_i is the deviation of the true score from the mean for subject *i*, and $e_{i,j}$ is the measurement error. In fully-crossed designs, main effects between coders where one coder systematically provides higher ratings than another coder may also be modeled by revising equation 5 such that

$$X_{i,j} = \mu + r_i + c_j + rc_{i,j} + e_{i,j} \quad (6)$$

where c_i represents the degree that coder *j* systematically

deviates from the mean and $rc_{i,j}$ represents the interaction between subject deviation and coder deviation. The variances of the components in equations 5 and 6 are then used to compute ICCs, with different combinations of these components employed based on the design of the study.

Higher ICC values indicate greater IRR, with an ICC estimate of 1 indicating perfect agreement and 0 indicating only random agreement. Negative ICC estimates indicate systematic disagreement, and some ICCs may be less than -1 when there are three or more coders. Cicchetti (1994) provides commonly-cited cutoffs for qualitative ratings of agreement based on ICC values, with IRR being poor for ICC values less than .40, fair for values between .40 and .59, good for values between .60 and .74, and excellent for values between .75 and 1.0.

Common ICC variants. Different ICC variants must be chosen based on the nature of the study and the type of agreement the researcher wishes to capture. Four major factors determine which ICC variant is appropriate based on one's study design (McGraw & Wong, 1996; Shrout & Fleiss, 1979) and briefly reviewed here.

First, the researcher must specify a one-way or two-way model for the ICC, which is based on the way coders are selected for the study. If a different set of coders is randomly selected from a larger population of coders for each subject then the researcher must use a one-way model. This is called "one-way" because the new random sample of coders for each subject prevents the ICC from accounting for systematic deviations due to specific coders (c_j in equation 6) or two-way coder × subject interactions ($rc_{i,j}$ in equation 6). In fully crossed designs, a two-way model is appropriate.

Second, the researcher must specify whether good IRR should be characterized by absolute agreement or consistency in the ratings. If it is important for raters to provide scores that are similar in absolute value, then absolute agreement should be used, whereas if it's more important that raters provide scores that are similar in rank order, then consistency should be used. For example, consider one coder who provides generally low ratings (e.g., 1-5 on an 8-point Likert scale) and another coder who provides generally high ratings (e.g., 4-8 on the same scale). One would expect the absolute agreement of these ratings to be low, as there were large discrepancies in the actual values of the ratings; however, it is possible for the consistency of these ratings to be high if the rank orderings of these ratings were similar between the two coders.

Third, the researcher must specify the unit of analysis that the ICC results apply to, that is, whether the ICC is meant to quantify the reliability of the ratings based on averages of ratings provided by several coders or based on ratings provided by a single coder. In studies where all subjects are coded by multiple raters and the average of their ratings is used for hypothesis testing, average-measures ICCs are appropriate. However, in studies where a subset of subjects is coded by multiple raters and the reliability of their ratings is meant to generalize to the subjects rated by one coder, a single-measures ICC must be used. Just as the average of multiple measurements tends to be more reliable than a single measurement, average-measures ICCs tend to be higher than single-measures ICCs. In cases where single-measures ICCs are low but average-measures ICCs are high, the researcher may report both ICCs to demonstrate this discrepancy (Shrout & Fleiss, 1979).

Fourth, the researcher should specify whether the coders selected for the study are considered to be random or fixed effects. If the coders in the study are randomly selected from a larger population and their ratings are meant to generalize to that population then the researcher may use a random effects model. These models are termed *random* because subjects and coders are both considered to be randomly selected. For example, this may be used in a study that assesses the degree to which randomly-selected psychologists give similar intelligence ratings to a set of subjects, with the intention of generalizing the results to a larger population of psychologists. If the researcher does not wish to generalize the coder ratings in a study to a larger population of coders or if the coders in a study are not randomly sampled, they may use a mixed effects model. These models are called *mixed* because the subjects are considered to be random but the coders are considered fixed. Note, however, that the ICC estimates for random and mixed models are identical, and the distinction between random and mixed is important for interpretation of the generalizability of the findings rather than for computation (McGraw & Wong, 1996).

ICCs use list-wise deletion for missing data, and therefore cannot accommodate datasets in fully-crossed designs with large amounts of missing data, and Krippendorff's alpha (Hayes & Krippendorff, 2007) may be more suitable when problems are posed by missing data in fully-crossed designs.

A summary of the ICC parameter options discussed here is outlined in Table 7.

Computational example. A brief example for computing ICCs with SPSS and the R *irr* package is provided based on the hypothetical 7-point empathy ratings in Table 5.

As with Cohen's kappa, SPSS and R both require data to be structured with separate variables for each coder for each variable of interest, as shown for one variable representing empathy ratings in Table 5. If multiple variables were rated for each subject, each variable for each coder would be listed

Table 5. Hypothetical ordinal empathy ratings for ICC example.

Subject	Emp_Rater1	Emp_Rater2	Emp_Rater3
1	6	5	6
2	5	5	5
3	6	6	7
4	2	1	3
5	3	3	3
6	2	1	1
7	6	5	5
8	7	6	6
9	5	5	4
10	4	3	5

Note: File structure is presented in spreadsheet format, where the first row must be converted to variable names when imported into SPSS or R.

in a new column in Table 5, and ICCs would be computed in separate analyses for each variable.

Both SPSS and the R *irr* package require users to specify a one-way or two-way model, absolute agreement or consistency type, and single- or average-measures units. The design of the hypothetical study informs the proper selection of ICC variants. Note that while SPSS, but not the R *irr* package, allows a user to specify random or mixed effect, the computation and results for random and mixed effects are identical. For this hypothetical study, all subjects were rated by all coders, which means the researcher should likely use a two-way model ICC because the design is fully crossed and an average-measures unit ICC because the researcher is likely interested in the reliability of the mean ratings provided by all coders. The researcher is interested in assessing the degree that coder ratings were consistent with one another such that higher ratings by one coder corresponded with higher ratings from another coder, but not in the degree that coders agreed in the absolute values of their ratings, warranting a consistency type ICC. Coders were not randomly selected and therefore the researcher is interested in knowing how well coders agreed on their ratings within the current study but not in generalizing these ratings to a larger population of coders, warranting a mixed model. The data presented in Table 5 are in their final form and will not be further transformed, and thus these are the variables on which an IRR analysis should be conducted.

Syntax for computing ICCs with SPSS and the R *irr* package are provided in Table 6. Both procedures provide point estimates, confidence intervals, degrees of freedom, and significance tests for the null hypothesis that $ICC = 0$. In practice, only point estimates are typically reported, although confidence intervals can provide additional useful information, particularly if ICCs are low or if the confidence interval is large due to a small sample size. Significance test results are not typically reported in IRR studies, as it is expected that IRR estimates will typically be greater than 0 for trained coders (Davies & Fleiss, 1982).

The resulting ICC is high, $ICC = 0.96$, indicating excellent IRR for empathy ratings. Based on a casual observation of the data in Table 5, this high ICC is not surprising given that the disagreements between coders appear to be small relative to the range of scores observed in the study, and there does not appear to be significant restriction of range or gross violations of normality. Reporting of these results should detail the specifics of the ICC variant that was chosen and provide a qualitative interpretation of the ICC

Table 6. Syntax for computing ICCs in SPSS and R.

SPSS Syntax

```
RELIABILITY
  /VARIABLES=Emp_Rater1 Emp_Rater2 Emp_Rater3
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /ICC=MODEL(RANDOM) TYPE(CONSISTENCY) CIN=95 TESTVAL=0.
```

R Syntax

```
library(irr)                    #Load the irr package (must already be installed)
hist(myRatings[,1]))            #Examine histogram for rater 1 for violations of normality
hist(myRatings[,2]))            #Examine histogram for rater 2
print(icc(myRatings, model="twoway", type="consistency", unit="average"))
                                #Specify the ICC model, type, and unit as appropriate.
                                #Use help(icc) for keywords
```

Note: R syntax assumes that data are in a matrix or data frame called "myRatings." In SPSS, model may be MIXED, RANDOM, or ONEWAY, type may be CONSISTENCY or ABSOLUTE. Single- and average-measures units will be included in SPSS output. In R, model may be twoway or oneway, type may be consistency or absolute, and unit may be average or single.

Table 7. Summary of IRR statistics for nominal variables.

Statistical Family	Variant	Uses	R command	SPSS command	Reference(s)
Kappa (two coders)	*Cohen's kappa*	No bias or prevalence correction	cohen.kappa(…)$kappa.c (*concord* package)	None*	Cohen (1960)
	Siegel & Castellan's kappa	Bias correction	cohen.kappa(…)$kappa.sc (*concord* package)	CROSSTABS \STATISTICS=KAPPA …	Siegel & Castellan (1988, pp. 284-291)
	Byrt et al's kappa	Prevalence correction	cohen.kappa(…)$kappa.bbc (*concord* package)	None*	Byrt et al. (1993)
	Cohen's weighted kappa	Disagreements differentially penalized (e.g., with ordinal variables)	kappa2(…, weight = c("equal", "squared')) (*irr* package)	None, but quadratic weighting is identical to a two-way mixed, single-measures, consistency ICC	Cohen (1968)
Kappa-like Pi-family statistics (three or more coders)	*Fleiss's kappa*	Raters randomly sampled for each subject	kappam.fleiss(…) (*irr* package)	None*	Fleiss (1971); Gross (1986)
	Light's kappa,	Average kappa across all rater pairs	kappam.light(…) (*irr* package)	None*, but two-rater kappa can be computed for each coder pair then averaged manually	Light (1971)
	Davies & Fleiss's kappa	Kappa-like coefficient across all rater pairs using average *P(e)*	None*	None*	Davies & Fleiss (1982)

Note: *Macros and syntax files may be available for computing statistical variants that are not natively available in SPSS or the R *concord* or *irr* packages. The reader is referred to the SPSSX Discussion hosted by the University of Georgia (http://spssx-discussion.1045642.n5.nabble.com) as one example repository of unverified user-created macros and syntax files created for SPSS.

estimate's implications on agreement and power. The results of this analysis may be reported as follows:

> *IRR was assessed using a two-way mixed, consistency, average-measures ICC (McGraw & Wong, 1996) to assess the degree that coders provided consistency in their ratings of empathy across subjects. The resulting ICC was in the excellent range, ICC = 0.96 (Cicchetti, 1994), indicating that coders had a high degree of agreement and suggesting that empathy was rated similarly across coders. The high ICC suggests that a minimal amount of measurement error was introduced by the independent coders, and therefore statistical power for subsequent analyses is not substantially reduced. Empathy ratings were therefore deemed to be suitable for use in the hypothesis tests of the present study.*

Conclusion

The previous sections provided details on the computation of two of the most common IRR statistics. These statistics were discussed here for tutorial purposes because of their common usage in behavioral research; however, alternative statistics not discussed here may pose specific advantages in some situations. For example, Krippendorff's alpha can be generalized across nominal, ordinal, interval, and ratio variable types and is more flexible with missing observations than kappa or ICCs, although it is less well-known and is not natively available in many statistical programs. The reader is referred to Hayes and Krippendorff (2007) for an introduction and tutorial on Krippendorff's alpha. For certain cases of non-fully crossed designs, Putka et al. (2007) provide an index of IRR that allow systematic deviations of specific coders to be removed from the error variance term, which in some cases may be superior to ICCs because ICCs cannot remove systematic coder deviations in non-fully crossed designs.

Many research designs require assessments of IRR to show the magnitude of agreement achieved between coders. Appropriate IRR statistics must be carefully selected by researchers to ensure their statistics fit with the design and goal of their study and that the statistics being used are appropriate based on the distributions of the observed ratings. Researchers should use validated IRR statistics when assessing IRR rather than using percentages of agreement or other indicators that do neither account for chance agreement nor provide information about statistical power. Thoroughly analyzing and reporting results of IRR analyses will more clearly convey one's results to the research community.

References

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology*, *46*(5), 423-429.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of

Table 8. Summary of ICC statistic parameters for ordinal, interval, or ratio variables.

Statistical Family	Parameter	Uses	R command	SPSS command	Reference(s)
Intraclass Correlation		Agreement for ordinal, interval, or ratio variables	icc(…) (*irr* package)	RELIABILITY …	
	Model	*One-way*: Raters randomly sampled for each subject *Two-way*: Same raters across subjects	model="oneway" or model="twoway"	/ICC= MODEL(ONEWAY), or /ICC= MODEL(MIXED) (for two-way mixed) or /ICC=MODEL(RANDOM) (for two-way random)	McGraw & Wong (1996); Shrout & Fleiss (1979)
	Type	*Absolute agreement*: IRR characterized by agreement in absolute value across raters *Consistency*: IRR characterized by correlation in scores across raters	type="agreement" or type="consistency"	/ICC= TYPE(ABSOLUTE) or /ICC= TYPE(CONSISTENCY)	McGraw & Wong (1996)
	Unit	*Average-measures*: All subjects in study rated by multiple raters *Single-measures*: Subset of subjects in study rated by multiple raters	unit="average" or unit="single"	Both unit types are provided in SPSS output	McGraw & Wong (1996); Shrout & Fleiss (1979)
	Effect	*Random*: Raters in study randomly sampled and generalize to population of raters *Mixed*: Raters in study not randomly sampled, do not generalize beyond study	None, but both effect parameters are computationally equivalent	/ICC= MODEL(MIXED) (for two-way mixed) or /ICC= MODEL(RANDOM) (for two-way random)	McGraw & Wong (1996)

thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220.

Davies, M., & Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, *38*(4), 1047-1051.

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, *30*(1), 95–101.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Gamer, M., Lemon, J., Fellows, I., & Sing, P. (2010). irr: Various coefficients of interrater reliability and agreement (Version 0.83) [software]. Available from http://CRAN.R-project.org/package=irr

Gross, S. T. (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics*, *42*(4), 883-893.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment*, *1*(6), 1–6.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications, Beverly Hills, CA.

Lacroix, G.L., & Giguère, G. (2006). Formatting data files for repeated-measures analyses in SPSS: using the Aggregate and Restructure procedures. *Tutorials in Quantitative Methods for Psychology, 2*(1), 20-26.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lemon, J. & Fellows, I. (2007). concord: Concordance and reliability (Version 1.4-9) [software]. Available from http://CRAN.R-project.org/package=concord

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365-377.

Lord, F. M. (1959). Statistical inferences about true scores. *Psychometrika 24*(1), 1-17.

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

Norman, G. R., & Streiner, D. L. (2008). *Biostatistics: The bare essentials*. BC Decker: Hamilton, Ontario.

Novick, M. R. (1966). The axioms and principle results of classical test theory. *Journal of Mathematical Psychology 3*, 1-18.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959-981.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321-325.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Yaffee, R. A. (2003). Common correlation and reliability analysis with SPSS for Windows. Retrieved July 21, 2011, from http://www.nyu.edu/its/statistics/Docs/correlate.html

Manuscript received November 15th, 2011.
Manuscript accepted January 17th, 2012.