BRIEF REPORT

Interrater Agreement Statistics With Skewed Data: Evaluation of Alternatives to Cohen's Kappa

Shu Xu and Michael F. Lorber New York University

Objective: In this study, we aimed to evaluate interrater agreement statistics (IRAS) for use in research on low base rate clinical diagnoses or observed behaviors. Establishing and reporting sufficient interrater agreement is essential in such studies. Yet the most commonly applied agreement statistic, Cohen's κ , has a well known sensitivity to base rates that results in a substantial penalization of interrater agreement when behaviors or diagnoses are very uncommon, a prevalent and frustrating concern in such studies. **Method:** We performed Monte Carlo simulations to evaluate the performance of 5 of κ 's alternatives (Van Eerdewegh's V, Yule's V, Holley and Guilford's V, Scott's V, and Gwet's V, alongside V itself. The simulations investigated the robustness of these IRAS to conditions that are common in clinical research, with varying levels of behavior or diagnosis base rate, rater bias, observed interrater agreement, and sample size. **Results:** When the base rate was 0.5, each IRAS provided similar estimates, particularly with unbiased raters. V0 was the least sensitive of the IRAS to base rates. **Conclusions:** The results encourage the use of the V0 statistic for its consistent performance across the simulation conditions. We recommend separately reporting the rates of agreement on the presence and absence of a behavior or diagnosis alongside V0 as an index of chance corrected overall agreement.

Keywords: behavior observation, diagnosis, interrater agreement, skew, low base rate

Supplemental materials: http://dx.doi.org/10.1037/a0037489.supp

Behavior observation and clinical diagnosis are cornerstone research methods in clinical psychology and related disciplines (Bakeman & Gottman, 1997; Kazdin, 2003). In behavior observation, a rater typically views, listens to, or reads transcripts of behavior and makes judgments about the quantity being measured (e.g., whether a person displayed anger). In clinical diagnosis, a rater typically judges responses to a diagnostic interview (e.g., reports of symptoms, their durations, and consequences to functioning) to determine the presence or absence of a given disorder (e.g., major depressive disorder).

Human judgment is at once a significant strength as well as a weakness in behavior observation and clinical diagnosis. People are capable of integrating a complex set of cues to arrive at psychologically informed judgments. At the same time, these judgments are imperfect; they are influenced by the characteristics of the raters themselves (e.g., experiences, conscientiousness) as well as random error. Even when raters use the same set of rules, they often reach different conclusions (Lorber, 2006). Careful training can reduce but not eliminate differences among raters. Accordingly, establishing and

servation and clinical diagnosis studies. Researchers must be able to show that their judgments of behavior and diagnoses are sufficiently reproducible among different raters to be taken as scientifically objective observations. Furthermore, obtaining substantial interrater agreement is a critical step in establishing the reliability and validity of rating data (Bakeman & Gottman, 1997), although it does in and of itself guarantee their psychometric quality; see Mitchell's (1979) cogent discussion of this topic.

The most common interrater agreement statistic (IRAS) is Co-

reporting sufficient interrater agreement is essential in behavior ob-

The most common interrater agreement statistic (IRAS) is Cohen's κ (Cohen, 1960), which has become so commonplace that most authors no longer cite a supporting publication when using it. As popular as it is, κ has a well-known and pronounced sensitivity to distributional skew (e.g., Cicchetti & Feinstein, 1990; Uebersax, 1987; Vach, 2005). That is, if the behavior or diagnosis being rated happens at either a very high or a very low rate (i.e., the majority of ratings fall into only one of the response categories), κ is often very low, even with high observed interrater agreement. To illustrate, if 10% of study participants exhibit a given disorder and two diagnostic raters agree 90% of the time on the presence versus absence of these diagnoses, with balanced disagreements, κ is .44, which is well below the .6 threshold considered good by many researchers (Cicchetti, 1994). Cohen' k is even more sensitive to skew with lower base rate behaviors or diagnoses (e.g., prevalence of schizophrenia in an epidemiological study). Thus, Cohen's κ might not be an optimal option for clinical researchers because many of the most interesting phenomena in clinical psychology do

This article was published Online First August 4, 2014.

Shu Xu and Michael F. Lorber, Family Translational Research Group, New York University.

Correspondence concerning this article should be addressed to Shu Xu, Family Translational Research Group, 423 East 23rd Street, 2S- Library, New York University, New York, NY 10010. E-mail: sx5@nyu.edu

not occur very often (e.g., diagnoses of schizophrenia, physical aggression). In sum, even with very high observed interrater agreement, it is nearly impossible to achieve substantial κ when very uncommon or common behaviors or diagnoses are being examined. These issues have led some to conclude that "kappa is too responsive to base rates to be useful as an index of interrater agreement" (Uebersax, 1987, p. 141).

Although the base rate sensitivity of κ is a very well-known concern, there are as of yet no widely used alternatives— κ has continued to be the most common measure of interrater agreement. This state of affairs is a difficult one for clinical researchers. Should one continue to use κ despite its sensitivity to skew because it is the standard metric, one that will putatively be viewed as more acceptable by readers, peer reviewers, and editors? Are there any alternatives that are less sensitive to skew?

Types of Interrater Agreement Statistics

In this study, we examined the performance of κ alongside several different alternatives for categorical judgments, focusing on the dichotomous case (i.e., diagnosis or behavior present vs. absent). These alternative statistics are Van Eerdewegh's V (Spitznagel & Helzer, 1985), Yule's Y (1912; Spitznagel & Helzer, 1985), Holley and Guilford's (1964) G, Scott's (1955) π , and Gwet's (2002) AC_1 . Two independent raters are considered for simplicity in the present investigation. Let us assume that the two raters must independently endorse the presence or absence of a behavior (e.g., anger display) or diagnosis (e.g., depression), per the Table 1 interrater agreement matrix. The response categories are mutually exclusive. Additionally, each behavior or diagnosis is rated independently.

The interrater agreement coefficients, Cohen's κ , Holley and Gilford's G, Scott's π , and Gwet's AC_1 , can be described in a general formulation as a single quantity $\frac{p_o-p_c}{1-p_c}$, where p_o denotes the observed agreement, which is the probability that two raters both endorsed the presence or absence of a behavior or diagnosis, and p_c denotes the probability that the two raters agreed by chance. The interrater agreement estimate is the ratio of the difference between obtained and chance agreement to the maximum non-chance agreement. Thus, the quantity reflected by these four IRAS can roughly be interpreted as the extent to which interrater agreement exceeds chance. A value close to 1 indicates near-perfect agreement, whereas a value close to 0 does not necessarily mean that agreement is poor but rather that agreement is no greater than would be expected by chance. Negative values denote less than chance agreement. These IRAS differ only in how chance agree-

Table 1
Interrater Agreement Matrix

	Rater 1			
Rater 2	Behavior/diagnosis present	Behavior/diagnosis absent		
Behavior/diagnosis present Behavior/diagnosis absent	а с	b d		

Note. Sometimes referred to as a *confusion matrix* (Bakeman & Gottman, 1997); the letters a, b, c, and d are the individual cell counts; thus the total count, N, is the sum of a, b, c, and d.

ment is calculated. The mathematical formula of each IRAS is presented in Table 2.

Cohen (1960) introduced κ as an alternative to such categorical association statistics as the χ^2 , noting that association statistics are increased by greater than chance disagreements as much as agreements. Thus, κ was formulated to exclusively reflect chance corrected agreement rather than degree of association. Cohen's κ provides a correction for agreement by chance based on the obtained distributions of two raters rather than the hypothetical random behavior of raters under a predetermined set of conditions. The calculation of chance agreement for κ depends on the obtained marginal frequencies (i.e., the row and column totals in Table 1). Asymmetries between the two row marginals and between the two column marginals increase the estimate of chance agreement. This feature explains κ 's base rate sensitivity, as low base rates produce unbalanced marginals; more of the ratings fall into the behavior or diagnosis absent than present category.

Holley and Gilford's G is equivalent to several other IRAS that have been proposed over the years (Krippendorff, 2011). In contrast to κ , the calculation of chance agreement in G does not depend on the obtained frequencies but is defined a priori. The chance agreement calculation for G assumes that random ratings would be equally distributed among the coding categories. Thus chance agreement is computed as the inverse of the number of response categories (e.g., .50 in the dichotomous case).

Scott's π is a special case of Krippendorff's (2011) more general IRAS, $\alpha(Scott, 1955)$. Like that of κ , the calculation of chance agreement in π is based on the obtained distributions of the two raters. In a second similarity to κ , π considers any deviation from a symmetrical distribution to be a factor that increases the probability of chance agreement. Thus, it can be expected to be sensitive to skew.

Gwet (2008) recommended AC_1 as a less base rate sensitive alternative to κ and many other IRAS. He argued that the other IRAS' corrections for chance agreement might be questionable because chance agreement should be taken into account only when a rater is not certain about which response category a behavior or diagnosis should be put into. Thus, the p_c of AC_1 is defined as chance agreement only under the circumstance that two raters agree; however, at least one of them has performed a random rating. This differs from the other IRAS, which consider the entire matrix of agreements and disagreements in their calculations of chance agreement.

Van Eerdewegh's V and Yule's Y, described and recommended by Spitznagel and Helzer (1985), are the final two statistics considered herein. They do not use the general formulation and there is no correction for chance agreement. They are association statistics that are closely related to the odds ratio, which is known to be base rate insensitive. But V and Y are designed to assume the range of -1 to 1 like the other IRAS. Spitznagel and Helzer suggested that both V and Y are less sensitive to base rate than is

In this study, we conducted Monte Carlo simulations to provide empirical guidance to researchers in selecting among IRAS. We focus on V, Y, G, and AC_1 as each has been described as an alternative to κ that is less sensitive to base rate (Feng, 2013; Gwet, 2008; Spitznagel & Helzer, 1985); π is included for comparison as it is relatively common in literature.

Table 2
Formulas for Calculating Interrater Agreement Statistics

Statistic	Formula	p_c
Cohen's к		$\frac{(a+b)(a+c) + (c+d)(b+d)}{N^2}$ 0.5a
Holley & Guilford's G		N^2 0.5°
Scott's π	$\frac{p_o - p_c}{1 - p_c}$	$\left(\frac{(a+b)+(a+c)}{2N}\right)^2 + \left(\frac{(c+d)+(b+d)}{2N}\right)^2$
Gwet's AC_1		$2P_r(1 - P_r)$ where $P_r = \frac{(a+b) + (a+c)}{2N}$
Van Eerdewegh's V	$\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{(a+c)(b+d)}}$	
Yule's Y	$\frac{\sqrt{(a+c)(b+d)}}{\sqrt{ad} - \sqrt{bc}}$ $\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	
Cicchetti & Feinstein's p_{pos}	$\frac{2a}{(a+b)+(a+c)}$	
Cicchetti & Feinstein's p_{neg}	$\frac{2d}{(b+d)+(c+d)}$	

Note. The general interrater agreement formula applies to κ , G, π , and AC_I only. p_o = probability of observed agreement; p_c = probability of chance agreement. Lettered quantities correspond to those found in Table 1. ^a The chance agreement (p_c) of G is the inverse of number of response categories. In a two-rater-two-category case, the value of p_c is 0.5.

Prior Monte Carlo Evaluations of Kappa and Its Alternatives

Two prior Monte Carlo studies evaluated the performance of Cohen's k and its alternatives for quantifying interrater agreement despite the critical role of IRAS plays in clinical research. In Gwet (2008), various IRAS were compared against what the author described as "true interrater reliability" (p. 36), which he defined as the level of consistency between raters given that the chance agreements of the two raters are known. Simulation factors considered in his study were base rate (95%), sample size (20, 60, 80, and 100), and the chance of random rating. Gwet found that AC_1 was superior to κ , G, and π in that the difference between the estimated AC_1 and the true reliability was smallest in comparison to its alternatives. More recently, Feng (2013) conducted a Monte Carlo simulation to evaluate which among various IRAS were most and least affected by base rate. The results showed that κ , π , and AC_1 were impacted by base rate; however, a statistic that is equivalent to G (Maxwell's R.E; Gwet, 2008) was not. In Feng's study, one of the raters was presumed to be the gold standard. It is also important to point out that rater bias was not considered as a simulation condition.

Present Investigation

In previous simulation studies, Gwet's (2008) and Feng's (2013) results disagreed on which IRAS has the most desirable performance with skewed data, AC_1 or G, respectively. Moreover, neither simulation study evaluated the performance of V and Y, alternative statistics that are purportedly less sensitive to skew than is κ (Spitznagel &

Helzer, 1985). Additionally, Gwet's (2008) simulation did not vary base rate. Last, neither study modeled the impact of rater bias. Clinical researchers are very often faced with raters who have clear leanings in one direction or the other (e.g., a rater consistently fails to detect sarcasm; a rater overweighs mood disturbances compared with neurovegetative symptoms when diagnosing depression). Such leanings are an additional source of skew in observational data that may impact the performance of IRAS.

To address the above limitations, we conducted quantitative evaluations of κ and its best practice alternatives, V, Y, G, π , and AC_1 in the case of two independent raters making dichotomous judgments on the presence or absence of a behavior or diagnosis (see Table 1). A Monte Carlo experiment was designed to evaluate the performance of each statistic in a combination of conditions that are commonly encountered in clinical research, to yield results maximally applicable to such research:

Observed agreement: Clinical researchers are most often faced with high levels of overall or raw observed agreement (e.g., commonly in the 80%–90% range) because they have carefully trained their raters.

Rater bias: In clinical research, sometimes raters' disagreements are randomly assorted, with equally probable overcoding (false positives) and undercoding (false negatives; i.e., the counts in

 $^{^1}$ An anonymous reviewer suggested the inclusion of the rescaled κ statistic (κ'). We briefly describe and analyze κ' in the online supplemental material.

cells b and c in Table 1 are equal), yet other times raters have a clear bias in one direction or the other (i.e., the counts in cells b and c in Table 1 are different).

Base rate: Clinical researchers further encounter varying levels of skew in their data, ranging from behaviors and diagnoses that happen rarely (e.g., physical aggression; psychiatric diagnoses in epidemiological samples) to those that occur quite often (e.g., problem solving in couples interactions; externalizing disorders in an outpatient child psychology clinic).

Sample size: The sample sizes in observational studies and clinical research are typically not large. We simulated data with sample sizes comparable to the ones in previous simulation studies and published clinical and observational studies.

To our knowledge, there is no way of determining which IRAS best captures "the truth." The differences among most of the statistics are primarily in their methods of calculating chance agreement, against which each statistic compares observed agreement (*V* and *Y* excepted). There is no gold standard to determine true chance agreement (Gisev, Bell, & Chen, 2013; Uebersax, 1987)—only assumptions that vary among the methods each statistic uses to calculate chance agreement. Put differently, one cannot determine which correction is most correct. Given this challenge, our criteria for evaluating the statistics' relative merits were pragmatic. We argue that an optimal IRAS should (a) provide some level of protection from chance agreement, (b) be minimally sensitive to distributional skew, and (c) maintain these qualities with biased and unbiased raters and at varying sample sizes.

Certainly, the above evaluative criteria are not the only ones by which IRAS may be judged.² However, we believe that our criteria are appropriate for the following reasons. First, protection against chance agreement is important because the laws of probability dictate that some percentage of agreements will occur even if raters' decisions are completely independent. Researchers wish to know that the rate of agreement between raters is well above chance—a reflection of their internalization of the coding rules, more than agreements generated by pure luck. Second, as we have discussed, IRAS that are highly sensitive to skew often make it difficult for clinical researchers to evaluate interrater agreement in low or high base rate situations. With skewed data and a low IRAS value despite high observed agreement, the researcher does not know if it is the statistic or the raters that are to blame. Third, one would want a statistic that works well across a wide variety of conditions, such as differing levels of observed agreement and rater bias. Protection against chance agreement and relative insensitivity to behavior or diagnostic base rates is just as important when, for example, raters agree 80% versus 90% of the time and when raters' disagreements are either systematic (biased) or random (unbiased). Accordingly, we applied our three evaluative criteria to each IRAS in multiple scenarios to provide empirical guidance for researchers in choosing from among the various statistics.

Method

Simulation Conditions

Six IRAS were considered in this Monte Carlo simulation study: Cohen's κ, Van Eerdewegh's *V*, Yule's *Y*, Holley and Gilford's *G*,

Scott's π , and Gwet's AC_1 . The mathematical formula for each coefficient is presented in Table 2. We assumed under each simulation condition that two raters must report the presence or absence of a target behavior or diagnosis, each behavior or diagnosis is rated independently, and the response categories are mutually exclusive. We further assumed that each rater was fallible (Bakeman, McArthur, Quera, & Robinson, 1997), with neither providing a gold standard or "true" rating against which the other rater could be compared. Even expert judges' ratings against which other judges' ratings may be compared (e.g., senior psychiatrist vs. psychiatry resident) are fallible in that they do not have unerring access to "the truth" (Spitznagel & Helzer, 1985).

Data simulation was based on four independent factors, each crossed with the others:

Observed agreement: 80% and 90%.

Agreed on base rate: That is, the base rate of the behavior/diagnosis only considering cases for which the raters agreed, with all disagreements ignored³: 5%, 10%, 20%, 30%, 40%, and 50%.

Rater bias: In the unbiased rating condition, the raters each had a 50% rate of endorsing the presence of a behavior or diagnosis when they disagreed, and in the biased rating condition, one rater had a 75% rate of endorsing the presence of a behavior or diagnosis and a 25% rate of endorsing its absence when they disagreed, with the second rater doing the opposite.

Sample size: 20, 40, 80, and 120.

In summary, we evaluated the performance of IRAS in $2 \times 6 \times 2 \times 4 = 96$ conditions. The number of replicated samples in each condition was 1,000, which is one of most common choices in Monte Carlo simulation studies (Burton, Altman, Royston, & Holder, 2006).

Data Generation and Interrater Agreement Estimation

Data simulation was conducted in Data Step using Statistical Analysis System (SAS) 9.3 software. Ratings were generated by (a) repeatedly simulating random values from a uniform distribution, and then (b) recoding them to a response category according to the

 $^{^2}$ As pointed out by an anonymous reviewer, for example, some researchers have argued that the base rate of κ is, in fact, a desirable feature of the statistic (Cicchetti & Feinstein, 1990; Karelitz & Budescu, 2013; Simon, 2006).

³ Agreed on base rate can be understood in terms of Table 1. It represents the base rate considering only cases for which the raters agree on the presence (cell a) and absence (cell d) of a behavior or diagnosis, given by a/(a+d). Ideally, one would simply model base rate as it is commonly understood, irrespective of coder agreement (i.e., the rate of the behavior/diagnosis across all cases). However, with two raters, the base rate in Rater 1's data [(a+c)/N] is potentially different than the base rate in Rater 2's data [(a+b)/N]. Rater 1 and 2 base rates are only the same if cells b and c are equal (i.e., the disagreements are balanced, indicating that there is no rater bias). We chose to model agreed on base rate to navigate around this issue, creating experimental conditions that applied equally to both raters. See Appendix A for the correspondence between agreed on base rate and the ordinary base rates of Raters 1 and 2 for each experimental condition.

probability of observing each response category. The population values under each simulation condition are provided in Appendix A. Following data generation, IRAS were computed in SAS Data Step. Sample codes and data are provided in Appendix B.

Results

The means of the different IRAS under each simulation condition are summarized in Figure 1; the full data set is available on request from Shu Xu. Separate panels are included for observed agreement (80% vs. 90%) crossed with rater bias (unbiased vs. biased). In each panel, the mean of each empirical IRAS was presented on the y-axis and the values of base rate on the x-axis. A dotted line was added as the reference line of population observed agreement (i.e., the value used for data simulation) to show the discrepancy between this population value and each estimated IRAS. The depicted data were for a sample size of 40. The results for samples sizes of 20, 80, and 120 are omitted because they were highly similar to the N=40 results (results are available on request from Shu Xu).

When the base rate was 0.5, each IRAS provided similar estimates, particularly with unbiased raters. At lower base rates, the statistics deviated from the population agreement rate to varying extents. Four

of the IRAS exhibited similar patterns of base rate sensitivity. As the base rate decreased, k deviated more from the observed agreement. This pattern was clearly present at both 80% and 90% observed agreement and for both unbiased and biased ratings. However, κ 's sensitivity to base rate appeared to be somewhat enhanced with lower agreement. The results for Scott's π were so similar to those for κ that it was impossible to plot clearly separate lines for the two statistics; for this reason, π was omitted from Figure 1. The performance of V was very similar to that of κ and π , with slightly less deviation from the population observed agreement but high sensitivity to base rate. At 80% observed agreement, Yule's Y exhibited similar base rate sensitivity to κ , π , and V. However, Y's base rate sensitivity was somewhat diminished at 90% observed agreement, particularly with biased raters. At the lowest base rate modeled (5% agreed on base rate), each of the statistics in this paragraph leveled a substantial mean penalty for chance agreement (M = 0.45 across each statistic and condition).

Values of AC_1 exhibited a pattern of sensitivity to base rate that ran opposite to the results obtained for κ , π , V, and Y, although to a lesser extent: AC_1 drew closer to the population observed agreement as the base rate decreased. This base rate sensitivity was greater in the 80% than in the 90% observed agreement condition. At a 5% agreed on base rate and averaged across the observed

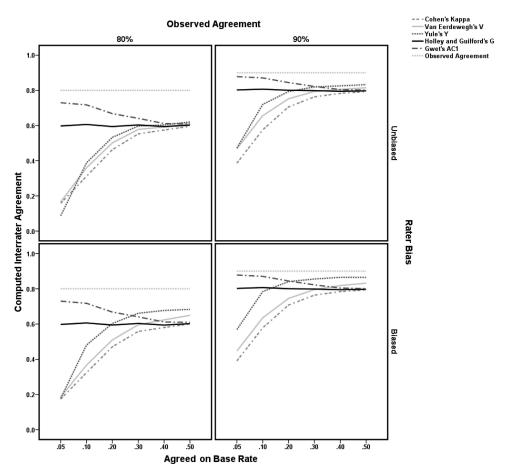


Figure 1. The performance of interrater agreement statistics s a function of behavior/diagnosis agreed on base rate, rater bias, and observed interrater agreement. Scott's π is not included as its results were virtually identical to those of Cohen's κ .

agreement and rater bias conditions, AC_I provided only a 0.04 mean penalty for chance agreement.

G was the least sensitive IRAS to base rates. This pattern held for both 80% and 90% observed agreement and for both biased and unbiased raters. At a 5% agreed on base rate and averaged across the observed agreement and rater bias conditions, G provided a 0.15 mean penalty for chance agreement.

We also assessed how much each IRAS is sensitive to rater bias. The estimates of G, AC_1 , and π were not affected by rater bias. Other IRAS (κ , V, and Y) were slightly sensitive to rater bias under all simulation conditions. Specifically, as the base rate increased, rater bias had less of an impact on κ , V, and Y. The values of κ and Y in the unbiased rater conditions were always lower than that in the biased rater conditions (the difference ranged from -0.0008 to -0.016 and from -0.032 to -0.097, for κ and Y, respectively). This pattern was not seen in V.

Discussion

We argued that an optimal IRAS should (a) protect against chance agreement, (b) be minimally sensitive to distributional skew (i.e., base rate), and (c) maintain these qualities with biased and unbiased raters and at varying sample sizes. Holley and Guilford's G was the statistic that best satisfied these criteria under all simulation conditions.

Gwet's AC_1 did the second best job in satisfying our criteria. Yet at lower base rates, AC_1 consistently had the highest values among its counterparts, which may yield overly optimistic estimates of interrater agreement. In other words, its relation to base rate is opposite to that of κ , albeit to a lesser extent. As base rate goes to the extreme low end, AC_1 reaches the asymptote of overall agreement, offering virtually no protection against chance agreement. Accordingly, AC_1 is difficult to recommend in clinical and behavior observational research settings for which low base rates are common.

Van Eerdewegh's V, Yule's Y, and Scott's π performed only slightly better than did κ relative to our criteria. Each of these three statistics was exceptionally sensitive to base rate. The sensitivities of V and Y were magnified at lower levels of interrater agreement and with rater bias. Thus, none of these three statistics offered substantial advantages over one another for conditions common in clinical research.

Limitations

The criteria that the IRAS were evaluated against are admittedly not the only ones that can be used for the task. We make no claim that they are the correct criteria, only that they provide a practically significant set of benchmarks for clinical research. By these benchmarks, G was the clearly superior metric. Ideally, Monte Carlo analyses like ours would be able to distinguish which of the IRAS was the most accurate or least biased. This is an elusive goal, however, as there does not seem to be a gold standard (Gisev et al., 2013; Uebersax, 1987). To restate the problem, we know of no method to determine which of the various corrections for chance interrater agreement is the "most correct."

There are two corollaries of the ambiguity in ascertaining which IRAS is the most accurate. First, the precise meanings of the quantities measured by each statistic are difficult to determine.

They each correct for chance agreement, but with no consensus on exactly what chance agreement is or how to compute it. Second, conventions of what magnitudes of each IRAS merit labels such as *poor*, *fair*, *good*, and *substantial* (e.g., Cicchetti, 1994) become quite arbitrary. Some authors have, in fact, argued that such guidelines are counterproductive (e.g., Gwet, 2012).

The present simulations were limited to two raters deciding on the presence versus absence of a behavior or diagnosis. Yet rating systems often have more than two categories and researchers often employ more than two raters. Generalizations of IRAS to accommodate such situations have been developed (e.g., Gwet, 2012; Krippendorff, 2011). However there may not yet be a multirater version of G, which proved to be the best performing statistic relative to our evaluative criteria. Moreover, we echo the caution expressed by Bakeman et al. (1997), who noted that IRAS can produce overly optimistic estimates of agreement with greater numbers of coding categories. Accordingly, we recommend the application of IRAS on a code-by-code basis and in pairs of coders when there are more than two.

A categorical approach in which the presence or absence of behaviors and diagnoses are measured is extremely common in clinical research. The evaluation of interrater agreement statistics for continuous outcomes may also be of interest in many circumstances. For example, the dimensional diagnosis of psychopathology has gained increasing momentum in recent years (Few et al., 2013). The IRAS we studied can be adapted to accommodate ordinal data in some cases (e.g., weighted k; Cohen, 1968), but they cannot be applied to truly continuous ratings. We warn against dichotomizing continuous measures, for example, categorizing continuous symptom severity ratings by splitting the sample at an arbitrary point on the scale of measurement. Dichotomization of continuous measures can yield misleading results (MacCallum, Zhang, Preacher, & Rucker, 2002). Other IRAS are available for ordinal and continuous data structures (Landis & Koch, 1975a, 1975b, 1977; McGraw & Wong, 1996; Whitehurst, 1984).

Conclusion and Practical Implications

Of the interrater agreement statistics, G appears to have the most balanced profile, leading us to endorse its use as an index of overall interrater agreement in clinical research. Of note for clinical and behavioral researchers, who are often faced with low base rate diagnoses and behaviors, G is insensitive to base rate.

We also note that although G offered the best performance relative to our criteria, it can only be considered an index of overall agreement. A high G does not necessarily imply that interrater agreement is acceptable for judging both the presence and the absence of a behavior or diagnosis. For this reason, we echo Cicchetti and Feinstein's (1990) recommendations to separately report the rates of agreement on the presence $(p_{\rm pos})$ and absence $(p_{\rm neg})$ of a behavior, alongside a chance corrected IRAS for overall agreement.⁴

In addition to the importance of including IRAS in research reports, IRAS are also valuable training tools. When training behavioral coders or clinical diagnosticians, IRAS can be used to measure raters' progress in learning a given set of decision criteria. Ordinarily this progress is judged relative to some sort of master

⁴ SAS code for the computation is presented in Appendix B.

ratings (e.g., consensus diagnoses of experienced clinicians or the behavioral observations of the trainer). In these cases, G will provide an overall sense of the chance corrected level of agreement. In addition, $p_{\rm pos}$ and $p_{\rm neg}$ can be used to separately probe raters' skill in deciding when a behavior or diagnosis is present versus absent. Low $p_{\rm pos}$ values indicate rater undersensitivity and low $p_{\rm neg}$ values indicate oversensitivity. These numbers can be used to provide specific feedback to uncover misunderstandings and clarify decision rules. Beyond the training phase, the combined use of G, $p_{\rm pos}$, and $p_{\rm neg}$ can be used in periodic checks of interrater agreement to quantify performance and prevent drift in the use of rating criteria.

IRAS can also be useful in applied clinical settings. Psychologists routinely make clinical diagnoses and judge the behavior of their clients. Those judgments carry great weight in affecting, for example, whether a diagnosis is assigned and which treatment is given. Inconsistency in clinical judgments can sometimes have very serious implications, as in the case of judgments of child abuse (Slep & Heyman, 2006), in which criminal prosecution and child removal from the home are potentially at stake. As described above, the combination of G, $p_{\rm pos}$, and $p_{\rm neg}$ can be used to help calibrate clinical judgments against an appropriate standard and, if possible, periodically monitor their consistency. These particular statistics are easily calculated, enhancing the feasibility of their use in clinical and research settings.

References

- Bakeman, R., & Gottman, J. M. (1997). Observing interaction: An introduction to sequential analysis. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/CBO9780511527685
- Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2, 357–370. doi:10.1037/1082-989X .2.4.357
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25, 4279–4292. doi:10.1002/sim.2673
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6, 284–290. doi:10.1037/1040-3590.6.4.284
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43, 551–558. doi:10.1016/0895-4356(90)90159-M
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37–46. doi:10.1177/001316446002000104
- Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220. doi:10.1037/h0026256
- Feng, G. C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, 47, 2959–2982. doi:10.1007/s11135-012-9745-9
- Few, L. R., Miller, J. D., Rothbaum, A. O., Meller, S., Maples, J., Terry, D. P., . . . MacKillop, J. (2013). Examination of the Section III DSM-5 diagnostic system for personality disorders in an outpatient clinical sample. Journal of Abnormal Psychology, 122, 1057–1069. doi:10.1037/a0034878
- Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy, 9, 330–338. doi: 10.1016/j.sapharm.2012.04.004

- Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Retrieved from http://www.agreestat.com/research_papers/kappa_statistic_is_not_satisfactory.pdf
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Sta*tistical Psychology, 61, 29–48. doi:10.1348/000711006X126600
- Gwet, K. L. (2012). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters (3rd ed.). Gaithersburg, MD: Advanced Analytics Press.
- Holley, J. W., & Guilford, J. P. (1964). A note on the G index of agreement. Educational and Psychological Measurement, 24, 749–753. doi:10.1177/001316446402400402
- Karelitz, T. M., & Budescu, D. V. (2013). The effect of the raters' marginal distributions on their matched agreement: A rescaling framework for interpreting kappa. *Multivariate Behavioral Research*, 48, 923–952. doi:10.1080/00273171.2013.830064
- Kazdin, A. E. (2003). Research design in clinical psychology (4th ed.). Boston, MA: Allyn & Bacon.
- Krippendorff, K. (2011). Agreement and information in the reliability of coding. Retrieved from http://repository.upenn.edu/asc_papers/278
- Landis, J. R., & Koch, G. G. (1975a). A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statistica Neerlandica*, 29, 101–123. doi:10.1111/j.1467-9574.1975
- Landis, J. R., & Koch, G. G. (1975b). A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). Statistica Neerlandica, 29, 151–161. doi:10.1111/j.1467-9574.1975 .tb00259.x
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33, 363–374. doi:10.2307/2529786
- Lorber, M. F. (2006). Can minimally trained observers provide valid global ratings? *Journal of Family Psychology*, 20, 335–338. doi:10.1037/0893-3200 20 2 335
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40. doi:10.1037/1082-989X.7.1.19
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46. doi:10.1037/1082-989X.1.1.30
- Mitchell, S. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, 86, 376–390. doi:10.1037/0033-2909.86.2.376
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325. doi:10.1086/ 266577
- Simon, P. (2006). Including omission mistakes in the calculation of Cohen's kappa and an analysis of the coefficient's paradox features. *Educational and Psychological Measurement*, 66, 765–777. doi:10.1177/0013164405285548
- Slep, A. M. S., & Heyman, R. E. (2006). Creating and field-testing child maltreatment definitions: Improving the reliability of substantiation determinations. *Child Maltreatment*, 11, 217–236. doi:10.1177/ 1077559506288878
- Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, 42, 725–728. doi:10.1001/archpsyc.1985.01790300093012
- Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101, 140– 146. doi:10.1037/0033-2909.101.1.140
- Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58, 655–661. doi: 10.1016/j.jclinepi.2004.02.021

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist*, *39*, 22–28. doi:10.1037/0003-066X.39 .1.22

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75, 579–652. doi: 10.2307/2340126

Appendix A
Population Parameters Used for Data Simulation

Condition	Observed agreement	Agreed on base rate	Biased rating	а	b	c	d	Rater 1 base rate	Rater 2 base rate
1	0.90	0.50	Biased	0.450	0.025	0.075	0.450	0.525	0.475
2	0.90	0.50	Unbiased	0.450	0.050	0.050	0.450	0.500	0.500
3	0.90	0.40	Biased	0.360	0.025	0.075	0.540	0.435	0.385
4	0.90	0.40	Unbiased	0.360	0.050	0.050	0.540	0.410	0.410
5	0.90	0.30	Biased	0.270	0.025	0.075	0.630	0.345	0.295
6	0.90	0.30	Unbiased	0.270	0.050	0.050	0.630	0.320	0.320
7	0.90	0.20	Biased	0.180	0.025	0.075	0.720	0.255	0.205
8	0.90	0.20	Unbiased	0.180	0.050	0.050	0.720	0.230	0.230
9	0.90	0.10	Biased	0.090	0.025	0.075	0.810	0.165	0.115
10	0.90	0.10	Unbiased	0.090	0.050	0.050	0.810	0.140	0.140
11	0.90	0.05	Biased	0.045	0.025	0.075	0.855	0.120	0.070
12	0.90	0.05	Unbiased	0.045	0.050	0.050	0.855	0.095	0.095
13	0.80	0.50	Biased	0.400	0.050	0.150	0.400	0.550	0.450
14	0.80	0.50	Unbiased	0.400	0.100	0.100	0.400	0.500	0.500
15	0.80	0.40	Biased	0.320	0.050	0.150	0.480	0.470	0.370
16	0.80	0.40	Unbiased	0.320	0.100	0.100	0.480	0.420	0.420
17	0.80	0.30	Biased	0.240	0.050	0.150	0.560	0.390	0.290
18	0.80	0.30	Unbiased	0.240	0.100	0.100	0.560	0.340	0.340
19	0.80	0.20	Biased	0.160	0.050	0.150	0.640	0.310	0.210
20	0.80	0.20	Unbiased	0.160	0.100	0.100	0.640	0.260	0.260
21	0.80	0.10	Biased	0.080	0.050	0.150	0.720	0.230	0.130
22	0.80	0.10	Unbiased	0.080	0.100	0.100	0.720	0.180	0.180
23	0.80	0.05	Biased	0.040	0.050	0.150	0.760	0.190	0.090
24	0.80	0.05	Unbiased	0.040	0.100	0.100	0.760	0.140	0.140

Note. The letters *a, b, c,* and *d* are the individual cell proportions corresponding to the interrater agreement matrix (see Table 1); these 24 conditions were simulated at each of four sample sizes (20, 40, 80, and 120) for a total of 96 conditions.

(Appendices continue)

Appendix B

SAS Code for Interrater Agreement Statistics

Sample data and SAS code are below.

```
*Input data set;
Data tempdatal;
input a b c d;
datalines;
19 2 2 17
Run;
*Sample SAS code for computing \kappa, \pi, AC_1,V, Y, G, p_{pos}, and p_{neg} statistics in a two-rater-
two-category case;
Data tempdata2;
 set tempdatal;
 N = a+b+c+d;
po = (a+d)/N;
Run:
Data tempdata3;
set tempdata2;
Kappa_pc = ((a+b) * (a+c)+(c+d) * (b+d))/(N*N);
AC1_pr = ((a+b)+(a+c))/(2*N);
AC1_pc = 2*AC1_pr * (1-AC1_pr);
Pi_pc = ((a+b+a+c)/(2*N))**2 + ((c+d+b+d)/(2*N))**2;
Kappa = (po-Kappa_pc)/(1-Kappa_pc);
AC1 = (po-AC1_pc)/(1-AC1_pc);
Pi = (po - Pi_pc)/(1-Pi_pc);
V = (sqrt(a*d) - sqrt(b*c)) / (sqrt(a+c) *sqrt(b+d));
Y = (sqrt(a*d) - sqrt(b*c)) / (sqrt(a*d) + sqrt(b*c));
G = ((a+d)/N-0.5)/(1-0.5);
Ppos = 2a/(a+b+a+c);
Pneg = 2d/(b + c + c + d);
RUN:
```

Received July 25, 2013
Revision received June 5, 2014
Accepted June 19, 2014