Article

An Evaluation of Interrater Reliability Measures on Binary Tasks Using *d-Prime*

Applied Psychological Measurement
2017, Vol. 41(4) 264–276

DOI: 10.1177/0146621616684584
journals.sagepub.com/home/apm
SAGE

Malcolm J. Grant[1], Cathryn M. Button[1], and Brent Snook[1]

Abstract

Many indices of interrater agreement on binary tasks have been proposed to assess reliability, but none has escaped criticism. In a series of Monte Carlo simulations, five such indices were evaluated using *d-prime*, an unbiased indicator of raters' ability to distinguish between the true presence or absence of the characteristic being judged. *Phi* and, to a lesser extent, *Kappa* coefficients performed best across variations in characteristic prevalence, and raters' expertise and bias. Correlations with *d-prime* for *Percentage Agreement*, Scott's *Pi*, and Gwet's AC_1 were markedly lower. In situations where two raters make a series of binary judgments, the findings suggest that researchers should choose *Phi* or *Kappa* to assess interrater agreement as the superiority of these indices was least influenced by variations in the decision environment and characteristics of the decision makers.

Keywords

reliability, interrater agreement, *Phi* correlation, *Kappa*, *Percentage Agreement*, research methods

Reliability of measurement is a fundamental aspect of sound research practice. One of the most common indicators of reliability involves having two people examine a set of cases and independently decide, for each case, if a characteristic exists. For example, in a study where deception cues are coded, a researcher might have two people decide on the presence or absence of a cue (e.g., gaze aversion) for a series of truth tellers and liars (e.g., Vrij, 2008). At issue is the way agreement should be measured and what the different agreement measures might reveal about the raters' ability to differentiate *true* presence from absence for each of the cues in question. All research that uses human judges to rate some characteristic requires an acceptable level of interrater agreement to demonstrate that the data, and any conclusions drawn from that data, are trustworthy. Although there is general consensus on this point, there is less agreement about which measure of reliability is best. The goal of the current research was to move toward a resolution by assessing the performance of several reliability measures against an evaluative standard in a series of Monte Carlo simulations. The standard the authors chose was *d-prime*, a statistic borrowed from signal detection theory (Swets, 1964), that has long been used as an unbiased indicator of raters' ability to distinguish between the true presence or absence of the

[1]Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada

Corresponding Author:
Malcolm J. Grant, Department of Psychology, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada A1B 3X9.
Email: mjgrant@mun.ca

signal or characteristic being judged. The authors reasoned that reliability measures that correlated most positively with the average *d-prime* scores of the raters should be preferred over measures that showed weaker correlations.

Measures of Interrater Agreement

There are many ways to measure interrater agreement on binary tasks, but all are based on the same four cell frequencies in a 2 × 2 table of possible outcomes (Cohen, 1960; Gwet, 2002; Holley & Guilford, 1964; Hunt, 1986; Krippendorff, 2011; Scott, 1955; Xu & Lorber, 2014; Zhao, Liu, & Deng, 2013). Zhao et al. (2013) provided a useful description of 22 indices of interrater reliability, and a discussion of the statistical assumptions on which each of the indices is based. They demonstrated that the measures differ primarily in how and to what extent the overall frequency of agreements is adjusted to take into account the number of agreements that would be expected by chance. As there is considerable overlap among the 22 indices, the authors chose five measures of interrater agreement based on (a) the frequency of use in research and (b) the relative distinctiveness in terms of how the index corrects for chance. A brief description of each of the measures follows, along with a rationale for its inclusion.[1]

Percentage Agreement

The percentage of cases on which the two raters agree is the simplest measure, and it seems to be the most commonly used. Despite its intuitive appeal, this measure has been criticized for being too liberal as it fails to account for agreements that may occur by chance (Zhao et al., 2013) and thus may overestimate the degree of reliability in the judgments (Hunt, 1986).

Cohen's *Kappa*

Kappa (Cohen, 1960) is reported widely in the literature and is considered by Zhao and colleagues (2013) to be the most common of the chance-adjusted indices and a conservative estimate of reliability. The number of agreements expected by chance is calculated using each rater's separate set of marginal totals to estimate the chance probabilities of agreement in the 2 × 2 table of possible outcomes. *Kappa* has been criticized, however, for applying too drastic a correction for chance (one that can range from 0 to 1), especially in cases where the characteristic being judged is either very common or very rare in the environment (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Gwet, 2002).

Scott's *Pi*

The *Pi* (Scott, 1955) statistic is also chance-adjusted and, according to Zhao et al. (2013), is second only to *Kappa* in popularity. It is distinct from *Kappa*, however, in that the *average* of raters' marginal totals (rather than each set of totals separately) is used to calculate the probabilities of chance agreement. The assumption appears to be that differences between raters can be ignored because both sets of marginal totals are estimates of the same quantity, namely, *Prevalence*. In contrast, by treating the raters' marginal totals separately, *Kappa* allows for the possibility that raters may differ in their totals, perhaps because of differences in their expertise or bias.

Gwet's AC_1

AC_1 (Gwet, 2002) is also chance-adjusted and uses average rather than individual marginal totals in the calculation of chance probabilities. Gwet, however, makes two assumptions: (a) a rater who is responding randomly would choose each alternative equally often and therefore would agree with the other rater, who may or may not be doing the same thing, no more than 50% of the time; (b) the more the average marginal totals deviate from 50%, the less likely it is that random responding is occurring, and therefore a less stringent correction for chance is needed. The chance correction for AC_1 is relatively moderate (i.e., ranging from 0 to a maximum of .5). Relative to *Kappa,* AC_1 is considered a liberal estimate of reliability (Zhao et al., 2013).

Phi Coefficient

The *Phi* coefficient is simply a Pearson's *r* when both variables are binary in form. This measure was included because it is familiar to researchers, readily calculated, easily interpreted, and reported widely in the literature. A strong positive value for *Phi* indicates that the raters are consistent in their judgments, presumably because they are responding to the same property in the cases being judged.

A Model of Binary Judgments

Although Zhao et al. (2013) distinguished between honest (i.e., objective) and dishonest (i.e., random) coding, most comparisons among the interrater indices have focused on statistical assumptions rather than psychological ones. In reality, characteristics of the judge and the environment, and the interaction among them, will influence decisions. What is needed is a model of binary judgments that incorporates the attributes of raters and the environment (e.g., Hogarth, 1987). As with any decision-making model, there are many factors that could potentially affect how people make decisions. As a starting point, the simplest model should reflect the central elements of everyday decision making. For instance, raters' decisions are likely influenced by (a) the prevalence of the characteristic in the environment, (b) the expertise of the raters in detecting the characteristic, (c) the extent and direction of bias in their judgments, and (d) fluctuating levels of attention to the task (see Goldstein & Hogarth, 1997, for a comprehensive review of the psychology of judgments and decision making).

Prevalence refers to the percentage of cases in the environment for which the characteristic is present. *Expertise* refers to a rater's ability to distinguish cases where a characteristic is present from cases where it is absent. This ability may also be thought of in terms of task difficulty, the clarity of the coding protocol, and the rater's level of training. If two raters are both high in expertise, agreement between them will be relatively high because both will be responding to the same objective reality. Lower levels of expertise for one or both raters will be associated with lower levels of agreement between them. *Bias* refers to a rater's tendency to use one of the classifications (i.e., present or absent) more often than the other, regardless of objective reality. Bias may sometimes be influenced by prevalence where, for example, a person who assumes that a particular characteristic is uncommon in the environment makes few decisions that it is present. Bias is conceptually distinct from expertise. Two raters may often agree if they are biased in the same direction even though neither has much expertise. Conversely, although less likely, they may disagree if they are biased in opposite directions even though each has considerable expertise. *Attention* refers to the moment-to-moment fluctuations in a rater's level of concentration. Lapses in attention will increase coding errors and lower agreement between two

raters. Unlike expertise and bias, which are stable characteristics of a rater, attention is a random situational factor that is unrelated to either bias or expertise.

For contextual purposes, consider our earlier example of deception research where two raters judge whether or not a deception cue is displayed by each of several targets (see Strömwall, Hartwig, & Granhag, 2006). In this task, there will be variations in the prevalence of the cues displayed by the targets (e.g., rare vs. common), the expertise of the people conducting the task (e.g., student assistants vs. experts in the area), the rater's expectations about which deception cues will likely be found (e.g., one researcher may think gaze aversion is ubiquitous whereas another researcher may think it is rare), and the factors that influence the attention of the raters (e.g., boredom, fatigue). All of these factors associated with the decision environment and raters can impact the reliability of judgments.

Evaluating Interrater Reliability Measures

In discussions of interrater agreement, authors (e.g., Xu & Lorber, 2014) have noted that comparative evaluations of measures are difficult because of the absence of a ''gold standard.'' That is to say, assessments of what constitutes a good measure of reliability have not examined whether a particular measure corresponds to the rater's ability to distinguish between actual absence or presence of a characteristic. Instead, evaluations have often relied upon beliefs about the validity of a measure's statistical assumptions. Such assumptions may be valid in some research situations but highly questionable in others. A more important focus involves identifying agreement measures that best reflect the ability of two raters to distinguish between the presence or absence of a characteristic. This reasoning led us to choose the computer simulation method where the presence or absence of a characteristic could be controlled.

In evaluating the ability to distinguish between the presence and absence of some characteristic across a series of cases, several measures have been used (cf. Alberg, Park, Hager, Brock, & Diener-West, 2004; Yaniv, Yates, & Keith-Smith, 1991). Common measures include the *Hit Rate* (the probability of detecting a characteristic when it is present) and (inversely) the *False-Alarm Rate* (the probability of mistakenly judging a characteristic to be present when it is absent). Both of these will be influenced by *Expertise* because of their connection to objective reality, but they will also be related to *Bias*. A rater who frequently judges a characteristic to be present (i.e., has a positive bias) will have not only a high *Hit Rate* but also a relatively high *False-Alarm Rate*. A rater who has a negative bias will tend to score low on both of these indices.

In signal detection theory (Stanislaw & Todorov, 1999; Swets, 1964), an attempt is made to separate a rater's ability from his or her response bias by defining a measure that reflects the *difference* between *Hit* and *False-Alarm Rates*. The probabilities of hits and false alarms are converted to their *Z*-score equivalents using the inverse cumulative normal distribution, and the difference between them is called *d-prime*:

$$d\text{-}prime = Z(Hit\ Rate) - Z(False\text{-}Alarm\ Rate).$$

An example showing the steps involved in calculating *d-prime* is given in the appendix.

Defined in this way, *d-prime* reflects a rater's ability to distinguish between true presence and absence. The authors believe it is this ability of raters that should be of central importance to both researchers and practitioners. If an agreement measure has a high correlation with *d-prime*, it indicates that the measure is capturing the ability of two raters to distinguish between the actual presence and absence of a characteristic being judged. By contrast, if a measure has

a low correlation with *d-prime*, it indicates that the measure provides little information about the raters' performance on the coding task.

The *d-prime* measure has the advantage that it assesses a rater's ability but, unlike other measures that attempt to do so, this measure is largely unaffected by a rater's level of *Bias*. The calculation of *d-prime*, however, requires that the researcher know when a signal is actually present or absent. This is typically the case in signal detection and recognition memory studies where this measure is often used, but it is not the case in many other studies where researchers must rely on the judgments of the raters to help them decide what is objectively the case.

The Current Study

The goal of the current study was to determine which of five measures of interrater agreement (i.e., *Percentage Agreement*, Cohen's *Kappa*, Scott's *Pi*, Gwet's AC_1, and the *Phi* coefficient) would correlate most strongly with *d-prime*. A strong correlation would indicate that the measure could serve as a proxy for *d-prime* when *d-prime* cannot be calculated directly (i.e., when the actual presence or absence of a characteristic is not known). To be clear, the authors were interested in identifying the measure of interrater agreement that best reflects raters' ability to distinguish real presence from absence of the characteristic in question. To accomplish this goal, the authors conducted several Monte Carlo simulations using the aforementioned model of binary judgments to generate data in a variety of situations that researchers might encounter.[2]

Method

A Model of Binary Judgments

The authors assumed that each rater's judgments are a function of (a) the true score for the case (i.e., whether the characteristic is present or absent), (b) the rater's *Expertise* or level of training in detecting the characteristic, (c) the rater's *Bias* toward thinking that the characteristic will be present or absent, and (d) the rater's *Attention* level. *Expertise* and *Bias* were assumed to be stable characteristics of raters that influence their judgment of each case. *Attention* level was assumed to be a factor that varies randomly over cases. In all simulations, the number of cases was set at 100 and the number of iterations within each simulation was set at 50,000.

True score (T). Each case was assigned a score of 1 (characteristic present) or −1 (characteristic absent). The percentage of cases where the characteristic was present (i.e., the *Prevalence* of the characteristic) was set at intervals of 10 in nine separate simulations (i.e., 10%, 20%, . . . , 90%).

Expertise (E). A randomly generated *Expertise* score was assigned to each rater. Across the 50,000 iterations of each simulation, these scores were normally distributed with a mean of zero and a standard deviation of one. Over the 100 cases within each iteration, these scores were constant (i.e., each rater's *Expertise* remained the same over cases).

Bias (B). A randomly generated bias score was assigned to each rater. Like the *Expertise* scores, across iterations of the simulation, these scores were normally distributed with a mean of zero and a standard deviation of one. Scores within each iteration were held constant for each rater.

Attention (A). A randomly generated level of *Attention* was assigned to each rater and for each case. These scores were normally distributed with a mean of zero and a standard deviation of one. *Attention* scores varied both within and across iterations of the simulation.

In the situation being modeled, the rater's task is to discriminate between cases that have true scores of 1 and those that have true scores of −1. A judgment score (J) was calculated for each rater and for each case as follows:

$$J_{ij} = \begin{cases} T_j + E_i + B_i + A_{ij} & \text{when } T = 1 \\ T_j - E_i + B_i - A_{ij} & \text{when } T = -1 \end{cases}$$

where $i = 1$ or 2 is the rater number and $j = 1, 2, \ldots, 100$ is the case number.

When the judgment score for a case was greater than zero, the rater's decision was *Present*, otherwise the decision was *Absent*. Thus, positive *Expertise* and *Attention* scores made judgments more accurate because judgment scores became more positive when the characteristic was present and more negative when the characteristic was absent. Negative *Expertise* and *Attention* scores made judgments less accurate because judgment scores became less positive when the characteristic was present and less negative when the characteristic was absent.

In contrast to *Expertise* and *Attention* scores, *Bias* scores simply increased or decreased the judgment scores regardless of the true value of the case. Raters with positive *Bias* were more likely to decide that the characteristic was present. Raters with negative *Bias* were more likely to decide that the characteristic was absent.

The model just described was applied to a situation where two raters judged the presence or absence of a characteristic across 100 cases. The characteristic was present in 10%, 20%, 30%, . . . , 90% of the cases. Each of these nine situations was simulated 50,000 times.

When both raters use each of the two categories exactly 50% of the time, the values of *Kappa, Pi,* AC_1, and *Phi* are identical. When the marginal totals differ from 50% but are the same for the two raters (e.g., a 70/30 split), *Kappa, Pi*, and *Phi* are identical but, in general, are different from AC_1. Comparisons among these measures will therefore be most informative when raters' marginal totals differ from 50% and when the marginal totals are discrepant from each other (e.g., a 90/10 split for Rater 1 and a 60/40 split for Rater 2). Because marginal totals will be especially sensitive to *Prevalence*, as well as to raters' *Expertise* and *Bias*, the inclusion of these factors in the model should generate data sets where the different strengths of the measures of agreement between raters will be most apparent.

On each of the 50,000 iterations of the nine simulations, the five measures of interrater agreement described earlier were recorded. A small percentage of iterations yielded data where one of the agreement measures could not be calculated (e.g., where one or both raters' judgments were constant). In these instances, the measure was coded as missing. The following three statistics were also recorded related to the model: (a) the average *Expertise* of the two raters, (b) the average *Bias* of the two raters (toward deciding that a characteristic was present), and (c) the average *d-prime* score for the two raters.

Results

The means and standard deviations for the model parameters as a function of *Prevalence* are shown in Table 1. The means and standard deviations for the agreement measures as a function of *Prevalence* are shown in Table 2.

Correlations among the five agreement measures and the average of the two raters' *d-prime* scores across levels of *Prevalence* are shown in Figure 1. As can be seen, the *Phi* and *Kappa* measures performed the best and were the most stable as the *d-prime* correlations remained above .70 across all levels of *Prevalence*. The remaining three measures, however, followed an inverted-U pattern; the *d-prime* correlations were highest when the characteristic occurred about

Table 1. Means and Standard Deviations for Model Parameters as a Function of the Prevalence of the Characteristic Being Coded (N = 50,000).

	Prevalence of characteristic								
Model parameters	10%	20%	30%	40%	50%	60%	70%	80%	90%
Expertise									
M	−0.002	−0.000	0.003	−0.002	0.002	0.002	−0.001	−0.002	0.001
SD	0.715	0.717	0.716	0.712	0.714	0.716	0.713	0.717	0.715
Bias									
M	0.002	0.005	−0.004	0.003	−0.000	−0.002	0.003	0.004	−0.001
SD	0.715	0.718	0.714	0.719	0.720	0.717	0.714	0.715	0.714
d-prime									
M	1.591	1.676	1.718	1.720	1.733	1.727	1.712	1.678	1.595
SD	1.150	1.197	1.212	1.214	1.221	1.219	1.211	1.198	1.153

Note. All statistics are based on averages of two raters.

Table 2. Means and Standard Deviations for Agreement Measures as a Function of the Prevalence of the Characteristic Being Coded.

	Prevalence of characteristic								
Agreement measures	10%	20%	30%	40%	50%	60%	70%	80%	90%
Percentage Agreement									
M	59.450	59.390	59.600	59.440	59.500	59.500	59.500	59.490	59.620
SD	23.681	21.582	20.051	19.017	18.762	19.064	20.107	21.689	23.620
Kappa									
M	0.138	0.169	0.185	0.188	0.189	0.188	0.183	0.169	0.140
SD	0.238	0.277	0.304	0.319	0.328	0.322	0.305	0.279	0.240
Pi									
M	−0.024	0.036	0.074	0.088	0.093	0.089	0.071	0.037	−0.021
SD	0.378	0.388	0.392	0.392	0.394	0.394	0.394	0.390	0.378
AC_1									
M	0.275	0.262	0.257	0.248	0.248	0.250	0.255	0.264	0.279
SD	0.498	0.452	0.412	0.381	0.371	0.381	0.413	0.454	0.496
Phi									
M	0.151	0.182	0.201	0.206	0.208	0.207	0.199	0.182	0.153
SD	0.278	0.313	0.331	0.341	0.347	0.343	0.333	0.315	0.278

50% of the time and were much lower when the *Prevalence* of the characteristic was either very high or very low.[3] The *Pi* statistic performed nearly as well as *Phi* and *Kappa* when *Prevalence* was around 50%, but its performance declined sharply at the extremes. Of particular importance, *Percentage Agreement*—arguably the most widely reported measure of reliability—performed very poorly across all levels of *Prevalence*. The same was also true of AC_1, which was the poorest performer across all levels of *Prevalence*. The results indicate a general superiority for the *Phi* and *Kappa* measures but one that is most pronounced at the extreme levels of *Prevalence*. It is important to note that although the values of *Phi* and *Kappa* do vary across *Prevalence* (see Table 2), their ability to predict *d-prime* does not.[4]

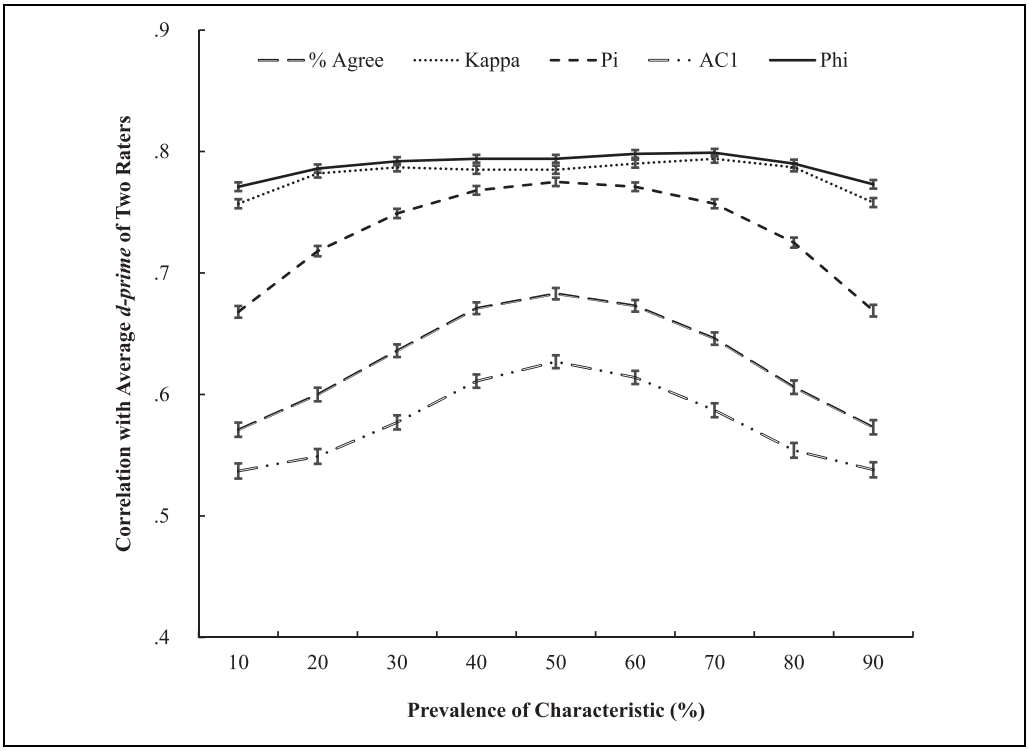

Figure 1. Correlations between measures of interrater agreement and average rater *d-prime* scores as a function of *Prevalence* of characteristic being rated.
Note. Error bars indicate 95% confidence intervals.

At two levels of *Prevalence* (50% and 90%), cases where the two raters were similar in their *Expertise* scores (both positive or both negative) and cases where they were different (one positive and the other negative) were examined. Those combinations were chosen because they are ones that are encountered frequently in everyday judgment environments, where one judge has more *Expertise* than the other or when both raters are similar in their level of *Expertise*. The correlations between measures of interrater agreement and the average of the two raters' *d-prime* scores for different combinations of rater *Expertise* are shown in Figure 2. As can be seen, *Phi* and *Kappa* maintained their superiority in predicting *d-prime* across all combinations of rater *Expertise*, and this was the case at both levels of *Prevalence*. Once again, the widely used measure of interrater reliability—*Percentage Agreement*—was a relatively poor indicator of raters' ability to distinguish between the presence and absence of the characteristic.

A similar procedure to that just described to examine differences in raters' biases was followed. Situations were compared where both raters were similar in bias (i.e., both negative or both positive) and ones where their biases differed. As with *Expertise*, these combinations were examined at two levels of characteristic *Prevalence* (50% and 90%). The correlations between measures of interrater agreement and average rater *d-prime* scores for different combinations of rater bias are shown in Figure 3. *Phi* and *Kappa* were superior to the other measures in almost all cases. The only exception occurred when *Prevalence* was 50% and the raters had opposite biases. In this case, all measures performed about equally well.

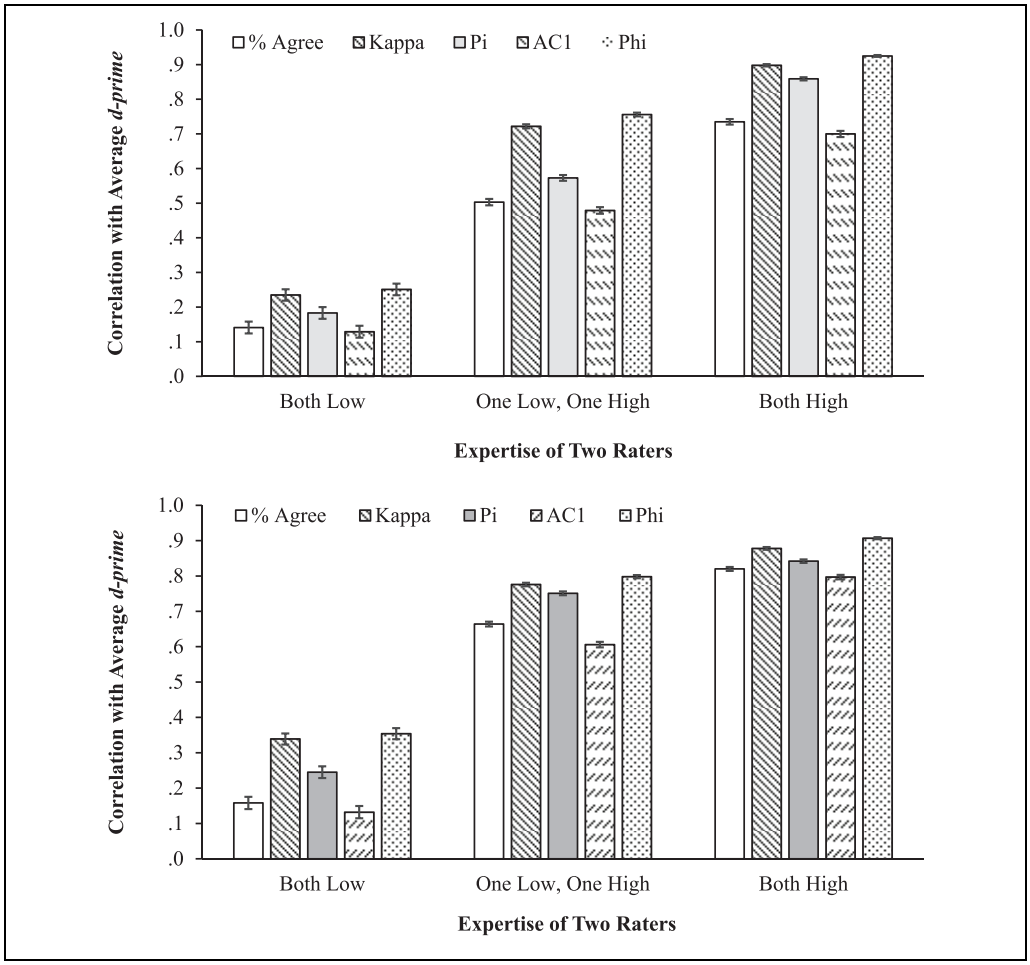

Figure 2. Correlations between measures of interrater agreement and average rater *d-prime* scores as a function of raters' *Expertise* (*Prevalence* = 50% in top panel, 90% in bottom panel).
Note. Error bars indicate 95% confidence intervals.

Discussion

Having reliable data is a precondition for any statistical test and any meaningful inference based on that test. For example, across judgment tasks that are common in forensic, clinical, medical, and academic settings, it is critical to establish interrater reliability because of the consequential nature of these judgments. All too often, judgment errors will increase the risk of misdiagnoses, miscarriages of justice, inadequate therapies, and the publication of scientific results that, in fact, reflect little more than random error. It is this danger that underlines the importance of obtaining judgments from more than one rater and using a measure of interrater agreement that provides the best indication of the reliability of those judgments. The contribution of the present article lies in the evidence it provides for the superiority of *Phi* and *Kappa*, relative to other measures of reliability that have sometimes been proposed.

As expected, the measures of interrater agreement that were examined were most clearly differentiated in situations where the characteristic being judged was either common or rare.

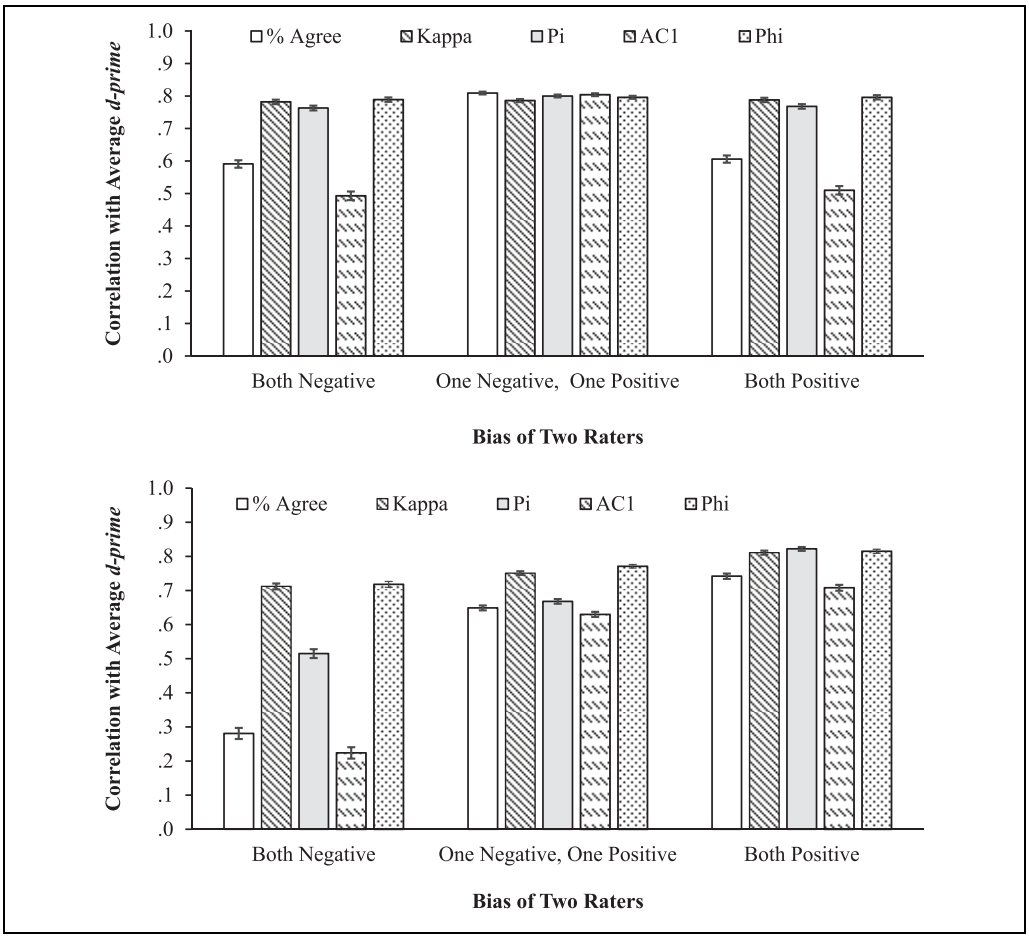

Figure 3. Correlations between measures of interrater agreement and average rater *d-prime* scores as a function of raters' *Bias* (*Prevalence* = 50% in top panel, 90% in bottom panel).
Note. Error bars indicate 95% confidence intervals.

Across all the variations in *Prevalence* that were simulated, *Phi* and *Kappa* were most closely related to *d-prime*. The authors also found that *Phi* and *Kappa* were as good as, and usually better than, the other measures regardless of whether raters were similar or different in their *Expertise* and *Bias*. The present findings suggest that in two-rater, binary-task situations like the ones simulated here, *Phi* or *Kappa* should be used to measure interrater agreement. On the contrary, researchers should not use *Percentage Agreement* or AC_1 to assess reliability.

It is a common and perplexing experience for researchers to calculate multiple measures of agreement, only to find that *Percentage Agreement* is quite high (e.g., above 80%) whereas *Phi* or *Kappa* is quite low (e.g., below .20). In such situations, *Percentage Agreement* is likely to be reported because it has intuitive appeal, is easily understood, and the necessity for chance correction is not obvious. The results clearly indicate that *Phi* and *Kappa* are superior measures of interrater reliability as they are correlated highly with *d-prime* and are less influenced by variations in *Expertise, Bias*, and *Prevalence* than are other measures. Thus, *Phi* or *Kappa* is to be preferred over *Percentage Agreement*.

The authors chose to simulate a situation where 100 cases are judged by two raters. Although this is a situation that is common in many research applications, there are certainly applications where fewer or more cases are judged. As the number of cases decreases, the utility of any agreement index will also decrease. Furthermore, with very few cases (e.g., $n = 10$), the calculation of interrater agreement would not likely be useful, and the correlations between the indices and *d-prime* would be less stable. With more than 100 cases, reliability of the measures would, of course, be higher, especially where the characteristics being rated are rare in the environment. With more than two raters or coding tasks with more than two alternatives, different measures such as a weighted *Kappa* (Cohen, 1968), intraclass correlation (Fleiss, 1975), or measures recently suggested by Cousineau and Laurencelle (2015) would need to be explored. The simulation procedure used in this study could be easily adapted to examine interrater agreement in these more data-extensive applications.

Another interesting direction for future research efforts might involve extending and empirically validating the model of binary judgments. The model that was used in this study was a simple one with many arbitrary features. The goal was to generate plausible data that would mimic real research data across a wide variety of applications. In this regard, the authors believe they succeeded but make no claim for the model's general psychological validity. Other researchers might consider and test several modifications of the model. In the model that was used, several simplifying assumptions were made. For example, the *Expertise, Bias*, and *Attention* factors were given equal weights but a model with differing weights might be more realistic in many situations. Similarly, levels of *Bias* of the raters were generated so as to be independent of the characteristic's prevalence, but in many situations it would be reasonable to expect some degree of positive correlation. Raters, for example, may come to their task with clear expectations, based on their experience, that the characteristic they are looking for will be relatively rare or common. In other cases, raters may come to the task with few expectations but may develop biases over time as they encounter cases where the characteristic is most often present or most often absent. Finally, attention levels may diminish over time, especially when a large number of cases have to be rated, and this possibility could also be incorporated in future simulations.

Researchers may also wish to use a similar simulation procedure to examine measures of interrater agreement other than the ones that were examined. As noted earlier, there are many candidates from which to choose, and it is possible that one or other of these may eventually prove to be superior to any of the ones considered in this study. Until this is demonstrated, however, the authors reiterate the recommendation that researchers choose either *Phi* or *Kappa* to assess interrater agreement on binary tasks.

Appendix

Calculation of *d-prime*: An Example

Consider a case where an observer, over 100 trials, attempts to judge the presence or absence of a characteristic that is, in fact, present on 50 of those trials. On 32 of the 50 trials when the characteristic is present, the observer correctly reports that the characteristic is present, yielding a hit rate of 32 / 50 or .64. On nine of the 50 trials when the characteristic is absent, the observer mistakenly reports that the characteristic is present, yielding a *False-Alarm Rate* of 9 / 50 or .18. In a cumulative standard normal distribution, a probability value of .64 corresponds to a Z value of .35846 (i.e., 64% of Z scores in a normal distribution are less than or equal to .35846). Similarly, a probability value of .18 corresponds to a Z value of $-.91537$. The value of *d-prime* is simply the difference between these two Z values.

$$\begin{aligned} d\text{-}prime &= Z(Hit\ Rate) - Z(False\text{-}Alarm\ Rate) \\ &= 0.35846 - (-0.91537) \\ &= 1.27383. \end{aligned}$$

In cases where *Hit* or *False-Alarm Rates* equal 0 or 1, a slight adjustment to the rates is necessary before the *Z* conversion can be applied. The 1 / (2*N*) rule (see Macmillan & Kaplan, 1985; Stanislaw & Todorov, 1999) was used, whereby proportions of 0 are increased by 1 / (2*N*) and proportions of 1 are decreased by the same amount. In both cases, *N* refers to the number on which the proportion is based.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. A fuller discussion of their statistical properties and assumptions can be found in Zhao, Liu, and Deng (2013) and Feng (2013). A clear presentation of the formulae for calculating each measure can be found in Xu and Lorber (2014).
2. The simulation program, written in *BASIC*, can be obtained by emailing the first author.
3. Note that the *Prevalence* dimension is symmetrical. For example, the results for the simulations using a 90% *Prevalence* rate always closely mirrored the results for the 10% *Prevalence* rate.
4. At the suggestion of a reviewer, the possibility of nonlinearity in the relations between the agreement measures and the mean *d-prime* values for the two raters was examined. At two levels of *Prevalence* (50% and 90%) and in a separate analysis for each measure, the authors regressed the *d-prime* means on the agreement measure as a predictor. In each analysis, the authors recorded the proportion of variance in *d-prime* accounted for by the linear component of the relationship (R^2 *linear*) and the additional variance accounted for when the quadratic component was added (R^2 *increase*). The values for R^2 *linear* ranged from .290 to .631 with a mean of .495. The values for R^2 *increase* across all analyses ranged from .005 to .057 with a mean of .030. Not surprising in view of the very large *N*, the increases in R^2, although small, were all statistically significant. Most important, however, in all of the models that included both the linear and quadratic components, *Phi* and *Kappa* continued to be the strongest predictors of *d-prime* while *Percentage Agreement* and AC_1 were the weakest. Thus, the conclusions based on the zero-order correlations shown in Figure 1 were reaffirmed in the models that included the higher order component.

References

Alberg, A. J., Park, J. W., Hager, B. W., Brock, M. V., & Diener-West, M. (2004). The use of ''overall accuracy'' to evaluate the validity of screening or diagnostic tests. *Journal of General Internal Medicine*, *19*, 460-465. doi:10.1111/j.1525-1497.2004.30091.x

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). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-220. doi:10.1037/h0026256

Cousineau, D., & Laurencelle, L. (2015). A ratio test of interrater agreement with high specificity. *Educational and Psychological Measurement*, *75*, 979-1001. doi:10.1177/0013164415574086

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543-549. doi:10.1016/0895-4356(90)90158-l

Feng, G. C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality and Quantity*, *47*, 2959-2982. doi:10.1007/s11135-012-9745-9

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651-659. doi:10.2307/2529549

Goldstein, W. M., & Hogarth, R. M. (Eds.). (1997). *Judgment and decision making: Currents, connections, and controversies*. Cambridge, UK: Cambridge University Press.

Gwet, K. (2002). *Kappa statistic is not satisfactory for assessing the extent of agreement between raters*. Available from www.agreestat.com

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision* (2nd ed.). Chichester, UK: John Wiley.

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

Hunt, R. J. (1986). Percent agreement, Pearson's correlation, and Kappa as measures of inter-examiner reliability. *Journal of Dental Research*, *65*, 128-130. doi:10.1177/00220345860650020701

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, *5*, 93-112. doi:10.1080/19312458.2011.568376

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*, 185-199. doi: 10.1037//0033-2909.98.1.185

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321-325. doi:10.1086/266577

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *3*, 137-149. doi:10.3758/bf03207704

Strömwall, L. A., Hartwig, M., & Granhag, P. A. (2006). To act truthfully: Nonverbal behaviour and strategies during a police interrogation. *Psychology, Crime & Law*, *12*, 207-219. doi: 10.1080/10683160512331331328

Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers*. New York, NY: John Wiley.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). Chichester, UK: John Wiley.

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's Kappa. *Journal of Consulting and Clinical Psychology*, *82*, 1219-1227. doi: 10.1037/a0037489

Yaniv, I. J., Yates, J. F., & Keith-Smith, J. E. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611-617. doi:10.1037/0033-2909.110.3.611

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind inter-coder reliability indices. In C. T. Salmon (Ed.), *Communication yearbook* (pp. 419-480). New York, NY: Taylor & Francis.