An Unbiased Estimate of Global Interrater Agreement

Educational and Psychological
Measurement
2017, Vol. 77(5) 721–742
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164416654740
journals.sagepub.com/home/epm

Denis Cousineau¹ and Louis Laurencelle²

Abstract

Assessing global interrater agreement is difficult as most published indices are affected by the presence of mixtures of agreements and disagreements. A previously proposed method was shown to be specifically sensitive to global agreement, excluding mixtures, but also negatively biased. Here, we propose two alternatives in an attempt to find what makes such methods so specific. The first method, R_B , is found to be unbiased while at the same time rejecting mixtures, is detecting agreement with good power and is little affected by unequal category prevalence as soon as there are more than two categories.

Keywords

statistical test, interrater agreement, category prevalence

Introduction

Estimating the extent to which two raters are agreeing on their assessment of cases to be classified in one of k categories is a difficult problem. One reason is that it is difficult to have a model of correct discriminating judgments and thus, “above-chance” performance is difficult to define. Krippendorff (2011, p. 97) provides three different conceptions of chance agreement whereas Feng (2013, p. 2965) notes two different conceptions of disagreement. Also, as noted by Krippendorff (2011), should a measure at the lower end of the scale reflect random agreement or maximum

¹University of Ottawa, Ottawa, Ontario, Canada

²Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

Corresponding Author:

Denis Cousineau, School of Psychology, University of Ottawa, 136 Jean-Jacques Lussier, Ottawa, Ontario, K1N 6N5, Canada.

Email: denis.cousineau@uottawa.ca

disagreement? Because what characteristics of the ratings should be maximized by a measure of interrater agreement is often undisclosed, many measures of interrater agreement have been proposed (see Zhao, Liu, & Deng, 2013, and Feng, 2013, for reviews) but there is no universally accepted criterion by which to decide whether some of them should be favored or discarded (Lombard, Snyder-Duch, & Bracken, 2004). As stated by Xu and Lorber (2014), “the precise meaning of the quantities measured by each [inter-rater agreement] statistic is difficult to determine” (p. 1224), which explains the contradictory recommendations found in the literature.

Recently, some have argued that an adequate measure should maximize both sensitivity to deviations from random agreements and specificity against nonrandom judgments that are not agreements (e.g., Cicchetti & Feinstein, 1990; Feng, 2013). To that end, Cousineau and Laurencelle (2015) have proposed a measure of interrater agreement with high sensitivity and specificity. This measure is based on the agreement matrix and it contrasts cells that support agreements against cells that contradict agreements. Its formula is

$$Q_A = \frac{\sum_{i=1}^k (z_{ii}^+)^2 + \sum_{i=1}^k \sum_{j=1, j \neq i}^k (z_{ij}^-)^2}{\sum_{i=1}^k \sum_{j=1, j \neq i}^k (z_{ij}^+)^2 + \sum_{i=1}^k (z_{ii}^-)^2}, \quad (1)$$

where quantities z_{ii}^+ are based on cells in the main diagonal that are higher than expected by chance (the “+” sign) and z_{ij}^- are based on cells outside the main diagonal that are lower than expected by chance. This measure goes from 0 to infinity with 1 representing a perfect balance between the two alternatives (agreement and disagreement), that is, chance agreement only.

This measure was turned into a more intuitive measure,

$$P_A = \frac{Q_A}{1 + Q_A}. \quad (2)$$

P_A runs from 0 to 1 with $\frac{1}{2}$ indicating random agreement (see a subsequent section for more information). Scores close to 0 can be interpreted as the raters showing significant disagreement and scores close to 1 can be interpreted as the raters showing significant agreement.

This measure was shown to be sensitive and specific using Monte Carlo simulations. Specificity was assessed by introducing mismatching ratings in the raters for a few categories. The results showed that P_A was rejecting the null hypothesis of no agreement ($H_0: P_A = \frac{1}{2}$) no more than expected from Type I error rate. However, all the other measures of interrater agreement tested were concluding significant agreement on a large proportion of simulations, sometimes as high as 30%. Thus, a major conclusion of Cousineau and Laurencelle (2015) was that many measures of interrater agreement (such as Cohen’s κ) can detect if there is agreement on at least some categories whereas only P_A can detect if there is agreement on all categories simultaneously. This is why we suggest calling it a measure of *global* interrater agreement.

Unfortunately, it was found that P_A is biased downward. This bias was seen by the fact that in the absence of agreement or disagreement, it returns on average a score smaller than $\frac{1}{2}$. The bias is stronger for smaller samples and for more numerous categories. Because of this downward bias, P_A can be safely used to test for agreement ("Is P_A significantly larger than $\frac{1}{2}$?"") but not to test for disagreement ("Is P_A significantly smaller than $\frac{1}{2}$?""). In unreported explorations, we tried to find a bias-correction formula to apply to P_A but could not find any. Thus, we set our efforts to locate an alternative to P_A , one that should be unbiased while preserving its sensitivity and specificity.

In that endeavor, a central question is: To what operation does P_A owe its qualities? We identify four key components in P_A . First, all the cells of the agreement matrix are used to compute Q_A and consequently, P_A . Cohen's κ , in contrast, only uses the main diagonal cells and the marginal sums, ignoring information outside the main diagonal. The new alternatives explored hereafter will keep this feature of P_A .

Second, in Q_A , the numerator regroups cells that are congruent with agreement (i.e., z_{ij}^+ in the main diagonal and z_{ij}^- outside the main diagonal; the z_{ij} are computed with $(o_{ij} - e_{ij})/\sqrt{e_{ij}}$ where o_{ij} is the observed number of cases that are classified as members of category i by Rater 1 and as members of category j by Rater 2, and e_{ij} , the expected count pertaining to the null hypothesis of independence between the row variable and the column variable, is obtained as usual from the product of the marginal frequencies). The denominator of Q_A , on the other hand, regroups cells that are incompatible with agreement (the z_{ii}^- in the main diagonal and the z_{ij}^+ outside the diagonal). Thus, Q_A is based on a ratio of two opposite hypotheses.

Third, the differences $(o_{ij} - e_{ij})$ are standardized using a division by $\sqrt{e_{ii}}$. This division brings with it a biasing effect. Indeed, positive signed differences are, on average, divided by smaller e_{ij} values, thus producing higher z^2 components, whereas negative-signed ones produce lower z^2 . In the $k \times k$ classification matrix, there should be, under significant agreement, k positive-difference cells and $k(k - 1)$ negative-difference cells; the asymmetries between the number of negative and positive cells and between the amplitude of the negative and positive cells result in the observed bias.

Fourth, analogously to the chi-square test, the signs of the z scores are removed by squaring them, producing the cell components $z_{ij}^2 = (o_{ij} - e_{ij})^2 / e_{ij}$. This is probably the least useful operation, as the components of Q_A are being selected to have the same signs (e.g., all z_{ij}^- are negative).

In sum, Q_A is a sufficient statistic; it is based on a ratio; it uses standardized differences; and signs are removed by squaring them. Are the last two operations important? We could for example avoid standardization and remove signs using the absolute value function instead of squares. Similarly, instead of a ratio, we could use an additive contrast (a difference between the supports given to the two hypotheses).

In what follows, we examine two alternatives to Q_A in the hope to find a new measure which preserves the high specificity and sensitivity qualities but which is also unbiased. The benefits of an unbiased estimate are that (1) it should be more powerful

than Q_A (not being dragged down away from rejection) and (2) it will be possible to test significance of agreements but also significance of disagreement, random judgments being in-between these extremes.

Here, we present two alternative measures of interrater agreement, that we call R_B and C_C . Both P_A and R_B are based on ratios whereas C_C is based on a purely additive contrast. The two new measures are simpler in the sense that they do not require standardized differences (i.e., no division) and squares are not used to remove signs. One keeps the ratio form; the other replaces it with an additive contrast. Variances of these measures are proposed so that statistical tests of significance can be performed. Afterward, we compare these measures in term of statistical power and sensitivity to a well-known test of interrater agreement, the Cohen's κ . This test was representative of the many tests compared in Cousineau and Laurencelle (2015), except P_A , which was the only sensitive test. We also examine confidence intervals of these measures and their aptitude to be used to test significant disagreement as well. Finally, we explore whether these measures are affected by unequal prevalence between the categories.

A Ratio Without Square and Without Standardization

The first alternative explored herein, which we call Q_B from which R_B will be derived, is similar to Q_A .¹ It uses the values d_{ij} obtained as

$$d_{ij}^s = \begin{cases} o_{ij} - e_{ij} & \text{if } \text{sign}(o_{ij} - e_{ij}) = s \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

To obtain d_{ij} , no square and no standardization (no division by $\sqrt{e_{ij}}$) is used. Recall that $e_{ij} = (o_{.j} \times o_{i.})/N$ where $o_{.j}$ and $o_{i.}$ are the observed marginal sums and N is the total number of cases classified by the two raters. The ratio is given by

$$Q_B = \frac{\sum_{i=1}^k d_{ii}^+ - \sum_{i=1, i \neq j}^k \sum_{j=1}^k d_{ij}^-}{\sum_{i=1, i \neq j}^k \sum_{j=1}^k d_{ij}^+ - \sum_{i=1}^k d_{ii}^-}. \quad (4)$$

which is turned into simpler measures with

$$P_B = \frac{Q_B}{1 + Q_B} \quad (5a)$$

or

$$R_B = 2P_B - 1 \quad (6a)$$

R_B is akin to a correlation with values bounded between -1 and $+1$, and for which 0 indicates no significant agreement of disagreement. This is the measure that will be used hereafter.

Note that if we let $A = \sum_{i=1}^k d_{ii}^+ - \sum_{i=1, i \neq j}^k \sum_{j=1}^k d_{ij}^-$ and $D = \sum_{i=1, i \neq j}^k \sum_{j=1}^k d_{ij}^+ - \sum_{i=1}^k d_{ii}^-$ for simplicity, then P_B is equivalently obtained from

$$P_B = \frac{A}{A + D} = \frac{A}{T} \quad (5b)$$

and R_B from

$$R_B = \frac{A - D}{A + D}, \quad (6b)$$

where T denotes $A + D$, the sum of all the cell differences without signs.

The sampling distribution of this measure is not known. However, exploratory simulations varying the number of categories (k) and the number of cases to be classified (N) suggest that in the absence of real agreement, it is very closely normally distributed with mean 0 and standard deviation:

$$\sigma_B \approx \frac{7.389}{k(k-1)(k+1)}. \quad (6)$$

Note that $7.389 \approx e^2$, but in the absence of a formal demonstration, this may be just a coincidence. Consequently,

$$\text{Reject } H_0 \text{ if } \frac{R_B}{\sigma_B} > z(1-\alpha) \quad (7)$$

is a right-tail test of the null hypothesis

$$H_0 : R_B = 0, \quad (8)$$

that is, there is no significant agreement or disagreement in the data.

A confidence interval can likewise be obtained from

$$CI_\gamma = [R_B - z((1 - \gamma)/2)\sigma_B, R_B + z((1 + \gamma)/2)\sigma_B], \quad (9)$$

where γ is the confidence level, typically 0.95.

A Contrast Measure Without Squaring and Without Ratio

A ratio is a useful approach to contrast two alternatives. It is commonly used in the probability domain (such as for expressing odds ratio). As the two terms go in opposite directions (if the numerator of Q_A increases, the denominator of Q_A necessarily decreases), the ratio magnifies these two opposite trends in a single measure. A simpler alternative is to subtract the two trends. This approach is commonly used in conjunction with analyses of variance (ANOVAs) and contrasts. Thus, we propose a second variation based on an additive contrast:

$$C_C = \left(\left(\sum_{i=1}^k d_{ii}^+ - \sum_{i=1}^k \sum_{j=1, j \neq i}^k d_{ij}^- \right) - \left(\sum_{i=1}^k \sum_{j=1, j \neq i}^k d_{ij}^+ - \sum_{i=1}^k d_{ii}^- \right) \right) / N, \quad (10)$$

where $d_{ij}^{\pm}$ are given in Equation (3). This measure has bounds given by -1 and $+1$; zero represents an absence of agreement or disagreement. With the above shortcuts, $C_C = (A - D)/N = (T - 2D)/N = T \times R_B / N$.

This measure is composed of a difference between two terms from quasi normal distributions that have their signs made positive. Thus, each component, under the null hypothesis, is the convolution of a random number of half-normal distributions. Putting two half-normal distributions back to back results in a distribution very similar to a normal distribution (as shown in the appendix). The number of differences added being an auxiliary variable, we were unable to find an exact formula for the standard deviation of this mixture. Exploratory simulations suggest that the distribution of C_C has a mean of zero and a standard deviation given by

$$\sigma_C \approx \frac{2}{k} \sqrt{\frac{k-1}{N}}. \quad (11)$$

A formal test is therefore

$$\text{Reject } H_0 \text{ if } \frac{C_C}{\sigma_C} > z(1-\alpha) \quad (12)$$

for a right-tail test of the null hypothesis

$$H_0 : C_C = 0 \quad (13)$$

A confidence interval is likewise given by

$$CI_\gamma = [C_C - z((1-\gamma)/2)\sigma_C, C_C + z((1+\gamma)/2)\sigma_C]. \quad (14)$$

Three Examples

To clarify the computations, let us consider the following scenario where two raters classified 200 cases into one of five categories (hence, $k = 5$ and $N = 200$). Table 1, top part, summarizes their ratings in a classification matrix. As seen, the agreeing ratings, located in the main diagonal, are moderately frequent, totalizing 30% of all the cases whereas 20% is expected if all the ratings were random.

Table 1, middle part, shows the expected counts based on the marginal frequencies. This method takes into account prevalence that can potentially be (and frequently is) unequal across categories.

Table 1, lower part, shows the differences between observed and expected scores, d_{ij} . With this table, it is possible to compute the three measures of interrater agreement (P_A , R_B , and C_C). The first measure, P_A requires $z_{ij} = d_{ij} / \sqrt{e_{ij}}$. Table 2 (top part) gives the results for all three measures as well as the Cohen's κ measure. All four

Table 1. Example of a Classification Matrix Showing a Fair Agreement Between the Two Raters.

Observed frequencies o_{ij}						
Rater 2						Sums
Rater 1	7	5	2	1	3	18
	5	13	10	7	8	43
	11	4	1	6	9	45
	8	11	7	9	6	41
	11	5	15	6	16	53
Sums	42	3	49	29	42	200
Expected frequencies e_{ij}						
Rater 2						Sums
Rater 1	3.78	3.42	4.41	2.61	3.78	18
	9.03	8.17	10.5	6.24	9.03	43
	9.45	8.55	11.0	6.53	9.45	45
	8.61	7.79	10.0	5.95	8.61	41
	11.1	10.1	13.0	7.69	11.1	53
Sums	42	38	49	29	42	200
Difference frequencies d_{ij}						
Rater 2						Sums
Rater 1	+ 3.22	+ 1.58	− 2.41	− 1.61	− 0.78	0
	− 4.03	+ 4.83	− 0.535	+ 0.77	− 1.03	0
	+ 1.55	− 4.55	+ 3.98	− 0.52	− 0.45	0
	− 0.61	+ 3.21	− 3.05	+ 3.06	− 2.61	0
	− 0.13	− 5.07	+ 2.02	− 1.69	+ 4.87	0
Sums	0	0	0	0	0	0

Note. The top part of the table shows the raw data; the middle part, the expected data; and the bottom part, the difference between the top and middle parts.

tests return scores that are above random (greater than $\frac{1}{2}$ for P_A , greater than 0 for the other three measures). These scores are all significant at $p < .01$.

Anticipating on the subsequent section, we also consider two more examples. Table 3 presents an instance where the raters are showing disagreement. As may be seen, the ratings contain some randomness; however, when one rater says “A” ($1 \leq A \leq k$), the second rater tends to avoid that response. Thus, the main diagonal contains fewer cases than expected by chance (and conversely, ratings are more numerous outside the main diagonal). A total of 17 cases only are located in the main diagonal, for a mere 8.5% proportion of agreement.

The middle part of Table 2 gives the results of the four tests of interrater agreement on these data using a left-sided test. All four confirm a significant lack of agreement (all $ps < .01$).

Table 2. Statistics Computed From the Ratings of Tables 1, 3, and 4.

Measure	Value	95% CI	Test statistic	p value
Data from Table 1 (right-sided tests)				
P_A	0.892	[0.168, 0.998]	8.223 ^a	.004
R_B	0.686	[0.198, 1.175]	2.754 ^b	.003
C_C	0.199	[0.089, 0.310]	3.527 ^b	<.001
κ	0.125	[0.045, 0.204]	3.078 ^b	.001
Data from Table 3 (left-sided tests)				
P_A	0.136	[0.005, 0.427]	0.158 ^a	.008
R_B	−0.685	[−1.173, −0.196]	−2.747 ^b	.003
C_C	−0.229	[−0.339, −0.118]	−4.046 ^b	<.001
κ	−0.143	[−0.191, −0.095]	−5.804 ^b	<.001
Data from Table 4 (right-sided tests)				
P_A	0.630	[0.295, 0.904]	1.704 ^a	.234
R_B	0.334	[−0.152, 0.824]	1.348 ^b	.089
C_C	0.245	[0.139, 0.361]	4.428 ^b	<.001
κ	0.156	[0.075, 0.237]	3.773 ^b	<.001

Note. CI = confidence interval. Test statistics is the measured statistics standardized; the p value is the probability of such measure or more extreme given the null hypothesis.
^aThis measure is to be interpreted as an *F* score. ^bThese measures are to be interpreted as *z* scores.

Table 3. Example of a Classification Matrix Showing Significant Disagreement.

		Observed frequencies o_{ij}				
		Rater 2				Sums
Rater 1	3	10	9	5	11	38
	11	4	10	10	5	40
	10	7	3	10	10	40
	16	3	11	2	10	42
	8	8	9	10	5	40
Sums	48	32	42	37	41	200

Finally, one last possibility is where partial agreement is mixed with partial disagreement: Table 4 shows an example. The raters agree well on Categories 1 and 5 but disagree on Categories 2, 3, and 4. This is a situation where the raters agree well on a few categories but where global agreement is weak. Yet proportion of agreeing judgments is 32.5% (65 cases out of 200), a rate higher than in the first example.

Table 2 (bottom part) gives the results of the four tests on those data using a (right-sided) test of agreement. As seen, two of the four tests do not reject the null hypothesis, P_A and R_B . They lead to the conclusion that global agreement is not present in this last example. Conversely, Cohen’s κ is lured by the presence of localized

Table 4. Example of a Classification Matrix Showing a Mix of Agreement and Disagreement and Thus Lacking Global Agreement.

		Observed frequencies o_{ij}				
		Rater 2				Sums
Rater 1	28	3	3	5	3	42
	2	8	24	6	4	47
	3	4	2	22	8	39
	4	21	4	8	6	43
	1	1	4	4	19	29
Sums	38	37	37	48	40	200

agreement, concluding to highly significant agreement ($\kappa = 0.156, p < .001$). This test is better seen as a test of *at least one agreeing category*.

This example shows that the Cohen's κ test lacks specificity: A mix of agreement and disagreement is deemed as agreement by this test. Anticipating on the results presented next, the second proposition, C_C , is also lacking specificity.

Comparison of the Three Tests to the Test Based on Cohen's κ

In this section, we present Monte Carlo simulations built around four scenarios (agreement of increasing strength, mixture of agreement and disagreement, disagreement, and agreement with unequal category prevalence). The results will be presented in three subsections. The first one reports sensitivity and specificity of the four tests using statistical power calculations. The second one assesses the adequacy of the proposed standard deviations and confidence intervals. The last one reports the impact of prevalence on four statistics of agreement. Before we begin, we provide some additional information on P_A and κ .

The P_A Measure, Its Statistical Testing, and Confidence Intervals

The distribution of P_A given in Equation (2) is not normal; instead, it follows approximately a Beta distribution with parameters $\nu_1 = P_A(k - 1)^2/2$ and $\nu_2 = (1 - P_A)(k - 1)^2/2$. Under H_0 , it is approximately a Beta distribution with both parameters equal to $\nu = \frac{1}{4}(k - 1)^2$. Thus, a right-sided test of P_A is given by

$$\text{Reject } H_0 \text{ if } P_A > B_{\nu, \nu}(1 - \alpha), \tag{15}$$

where $B_{\nu, \nu}(1 - \alpha)$ is the $1 - \alpha$ quantile of the Beta distribution. Likewise, a confidence interval on P_A is obtained with

$$CI_\gamma = \left[B_{v_1, v_2} \left(\frac{1 - \gamma}{2} \right), B_{v_1, v_2} \left(\frac{1 + \gamma}{2} \right) \right]. \quad (16)$$

Cohen's κ

Cohen's κ is one of the most commonly used measures of interrater agreement. It is given by

$$\kappa = \frac{r - E(r)}{1 - E(r)}, \quad (17)$$

where r is the observed rate of agreement given by $r = \sum_{i=1}^k o_{ii}/N$ and $E(r)$ is the expected rate of agreement given by $E(r) = \sum_{i=1}^k e_{ii}/N$. This statistic has an approximate standard deviation given by (Cohen, 1960, equation 9)

$$\sigma_\kappa \approx \sqrt{\frac{E(r)}{N(1 - E(r))}}. \quad (18)$$

Everitt (1968) found the exact expression of σ_κ whereas Fleiss, Cohen, and Everitt (1969, equation 14) found a more reliable but still manageable approximation. Yet in Cousineau and Laurencelle (2015), the benefit of using better expressions for σ_κ was found to be only marginal in terms of statistical power. With Equations (17) and (18), a straightforward right-tail test of the null hypothesis of no agreement is

$$\text{Reject } H_0 \text{ if } \frac{\kappa}{\sigma_\kappa} > z(1 - \alpha). \quad (19)$$

Finally, its confidence interval is given by

$$CI_\gamma = [\kappa - z((1 - \gamma)/2)\sigma_\kappa, \kappa + z((1 + \gamma)/2)\sigma_\kappa]. \quad (20)$$

Simulation Procedures

To compare the four tests, we ran many conditions. In the first set of simulations, we assess the four tests' statistical power and Type I error rate. To that end, we generated classification matrices. In one typical condition, agreement occurs with probability ρ , otherwise raters' choices are random so that spurious agreement on any category could occur with probability $1/k$. Algorithm 1 describes precisely how the ratings were obtained (taken from Cousineau & Laurencelle, 2015).

Algorithm 1. Steps to generate judgments with agreement from N cases into k categories with probability of agreement given by ρ_A ($0 < \rho_A < 1$).

Step 1: Generate a random integer x in $[1, \dots, k]$; (this is categorization of first rater)

- Step 2: With probability ρ_A , make $y \leftarrow x$ (an agreement between the two raters)
 Otherwise, generate a random integer y in $[1, \dots, k]$;
 Step 3: Add one observation in cell $\{x, y\}$;
 Step 4: Repeat Step 1 to Step 3 N times to obtain N observations:

Across conditions, the parameter ρ was varied from 0 to 0.3 by steps of 0.05 (above 0.3, power is perfect). We also varied the total number of cases ($N = 50$ to 500 by steps of 50) and the number of categories ($k = 3, 5$, and 7). For one combination of ρ , N , and k , we generated a classification matrix then computed the four statistics of interrater agreement (P_A , R_B , C_C , and κ). The ensuing statistical tests were performed at a level α of 5% (we also used 1% and 0.1% but did not find any qualitative differences). This process was repeated 50,000 times for each ρ , N , and k condition and the proportion of rejection of H_0 was computed. In distinct simulations, the whole process was repeated except that 95% confidence intervals were computed.

In the second set of simulations, we examine specificity by introducing a certain number of disagreements. To that end, with probability ρ we assign to what seems an instance of Category “A” for one rater ($1 \leq A \leq k$) the Category “B” to the second rater ($1 \leq B \leq k$). The mapping of the A s to the B s is random across simulations but constant within a given simulation. It is possible by chance that A and B are the same category; however, it is very unlikely that both raters have the exact same category assignment for all k categories. Thus, except on very rare occasions, there is no global agreement in this second set of simulations. Algorithm 2 describes precisely how the simulated ratings were generated.

Algorithm 2. Steps to generate judgments where some or all categories of one rater may not be the same categories for the second rater, given k , N , and ρ_C the probability of a nonrandom judgment ($0 < \rho_C < 1$).

- Step 0: Initialize a set of random pairings $\{x; y(x)\}$ of integers 1 to k ;
 Step 1: Generate a random integer x in $\{1, \dots, k\}$;
 Step 2: With probability ρ_C , make $y \leftarrow y(x)$, Otherwise, generate a random integer y in $\{1, \dots, k\}$;
 Step 3: Add one observation in cell $\{x, y\}$;
 Step 4: Repeat Step 1 to Step 3 N times to obtain N observations

In the third set of simulations, we examine the behavior of the four tests when there is disagreement between the raters. To that end, in case of a disagreement, Rater 2 could choose any category *except* the one chosen by Rater 1. The probability of a disagreement was controlled by the parameter ρ . Algorithm 3 provides the procedure followed to generate a classification matrix containing disagreement.

Algorithm 3. Steps to generate judgments with disagreement from N cases into k categories with probability of agreement given by ρ_D ($0 < \rho_D < 1$).

- Step 1: Generate a random integer x in $[1, \dots, k]$;

Step 2: With probability ρ_D , make $y \leftarrow$ random integer in $[1, \dots, k, \text{except } x]$ (dis-agreement) Otherwise, generate a random integer y in $[1, \dots, k]$;

Step 3: Add one observation in cell $\{x, y\}$;

Step 4: Repeat Step 1 to Step 3 N times to obtain N observations:

In the final set of simulations, we return to the study of agreement. However, this time the k categories can have unequal prevalence. To model prevalence, we assign weights to the categories: The first category always had a weight of 1; the last category had a weight of ω ; the in-between categories had weights linearly increasing between 1 and ω . When picking a choice at random for a rater, the choice was more frequently picked when its weight was larger. Algorithm 4 indicates the steps followed to generate a classification matrix in this last scenario. The weights ω used were $\{1, 3, 5, 10, 15, 20, \text{and } 25\}$.

Algorithm 4. Steps to generate judgments with agreement from N cases into k categories with probability of agreement given by ρ_A ($0 < \rho_A < 1$) and category imbalance given by ω ($\omega \geq 1$).

Step 0: Generate linear weights $W = \{1, \dots, \omega\}$ with steps $(\omega - 1) / (k - 1)$

Step 1: Generate a random integer x in $[1, \dots, k]$ weighted by W ;

Step 2: With probability ρ_A , make $y \leftarrow x$ Otherwise, generate a random integer y in $[1, \dots, k]$ weighted by W ;

Step 3: Add one observation in cell $\{x, y\}$;

Step 4: Repeat Step 1 to Step 3 N times to obtain N observations:

In all four sets of simulations, the same conditions were explored (ρ , N , and k). The following results only report some illustrative cases.

Results

Power and Specificity

We first report the power curves. These curves are in the right half of the plots of Figure 1 for three selected conditions $N = 125$ or 250 and $k = 5$ or 10 . As seen, all four measures have a Type I error rate (when true agreement rate is zero) very close to α , the most visible exception being P_A when $k = 10$ (bottom panel of Figure 1).

As ρ departs from zero, power rises steadily. Cohen's κ has a small power advantage over P_A and R_B but is below C_C . P_A is not the most powerful test and its power declines as k increases and as N diminishes. This is caused by its downward bias mentioned earlier. This bias is also the cause of a Type I error rate lower than α , visible in the bottom panel but present in all the simulations.

The left side of the plots of Figure 1 illustrates specificity. Increasing ρ (the minus sign is only used for graphing purposes) results in more concordant ratings but these concordant ratings are on the main diagonal only by chance so that there is no true agreement on the left side of Figure 1. Thus, rejection of H_0 should occur at a rate no

Figure 1. Proportion of rejection of H_0 as a function of ρ , where positive values indicate the strength of true agreement and negative values indicate the strength of concordance that may or may not be agreement. Top panel: $k = 5$, $N = 125$; middle panel: $k = 5$, $N = 250$; bottom panel: $k = 10$, $N = 125$.

higher than α . As seen, the only test where the Type I error rate is flat is R_B . P_A 's Type I error rate is dragged down to 0% because of its downward bias. As of κ and C_C , they have Type I error rates that are increasing with the importance of cells outside the main diagonal.

In light of the specificity results, we cannot conclude that κ and C_C are more powerful tests. Instead, we believe that some matrices generated randomly from Algorithm 1 turned out to be more representative of concordance without agreement

situations and were thus rejected (rightly) by P_A and R_B . The fact that both C_C and κ have almost the same lack of specificity and the same enhanced power curves strongly supports this hypothesis. Furthermore, the conditions where C_C has weaker specificity than κ are also the conditions where it surpasses κ in power (e.g., the bottom panel of Figure 1).

Bias and Confidence Intervals

In this subsection, we examine if the measures are biased or not. To that end, we illustrate in Figure 2 the distribution of the measures as a function of N (the horizontal axis) for three different ρ (rows of graphics). Each measure is shown in a different column. The full line shows the mean of the measures; the gray area shows where 95% of the results lie; the error bars shows the expected 95% confidence intervals (e.g., Equations 9, 14, 16, and 20; see Harding, Tremblay, & Cousineau, 2014, for similar plots).

Concentrating first on bias, we see that P_A has a small bias, visible for small N and $k = 5$ (top left panel of Figure 2). This confirms what was reported in Cousineau and Laurencelle (2015). This bias increases with k (not shown). Of the four measures explored, it is the only one biased. Thus, P_A is problematic (and as we will see next, this bias has a complex behavior, so there may be no way to correct for its presence).

Focusing on the other three measures, we see that C_C and κ confidence intervals capture near perfectly the scatter of the measures. The confidence intervals of R_B are valid only under H_0 ; as the effect size grows and as the sample size grows, the confidence intervals become overly conservative.

In Figure 3, we explored mixtures of agreement and disagreement. In these conditions, global agreement is zero and so the measures should be centered on $\frac{1}{2}$ for P_A and on 0 for the other three measures. Again, we see that P_A is biased downward but this time, bias increases with sample size. All the other measures are unbiased. However, we see that variances of C_C and κ are underestimated, more so with increasing concordance rate (lower panel) and with increasing k (not shown). This is the cause of the lack of specificity of these two measures.

Finally, in Figure 4, we varied the disagreement rate. Both P_A and R_B have exaggerated confidence intervals and to a lesser extent, the confidence intervals of C_C are too large as well. The confidence intervals of κ are quite accurate. We observe that when the true rate of agreement is positive or null, κ is an unbiased estimate of the parameter ρ_A used in the simulations. When disagreement is simulated, $(k - 1) \kappa$ is an unbiased estimate of ρ_D .

Summarizing the first two subsections under Results, we see that C_C has no specificity, as compared with R_B . Thus, this measure cannot be used to test for global interrater agreement. As a test of disagreement, it is too conservative (with too large error bars seen in Figure 4). Thus, this test has no attractive features and should not be used. It was however informative to realize that a ratio of two competing hypotheses is central to specificity.

Figure 2. Behavior of the four tests (columns) as a function of the sample size N (horizontal axis) and the amount of true agreement in the data (ρ_A , top line $\rho_A = 0.00$, middle line, $\rho_A = 0.05$, and bottom line, $\rho_A = 0.10$). The parameter k is constant at 5. The area denotes the zone where 95% of the observed scores were located whereas the error bars shows the 95% confidence intervals found from the formulas proposed in the text.

Figure 3. Behavior of the four tests (columns) as a function of the sample size N (horizontal axis) and the amount of concordance (not necessarily agreements) in the data (ρ_C , top line $\rho_C = 0.10$, middle line, $\rho_C = 0.20$, and bottom line, $\rho_C = 0.30$) in the same format as Figure 2. The parameter k is constant at 5.

Figure 4. Behavior of the four tests (columns) as a function of the sample size N (horizontal axis) and the amount of true disagreement in the data (ρ_D , top line $\rho_D = 0.15$, middle line, $\rho_D = 0.30$, and bottom line, $\rho_D = 0.45$) in the same format as Figure 2. The parameter k is constant at 5.

Cohen's κ is likewise not specific. However, for "pure" situation (not a mixture of agreement and disagreement), it is an excellent statistic. In situations with mixture of agreement and disagreement, it has underestimated error variance (and consequently, too short error bars) because it does not consider information outside the main diagonal. It assumes that the ratings are uniform outside the main diagonal, dismissing valuable information.

Sensitivity to Unequal Category Prevalence

Many authors noted that when category prevalence is quite unequal, it is possible to find a high proportion of agreement but a weak Cohen's κ , which was noted as a paradox. Some concluded that κ in this situation is biased downward by the presence of prevalence (e.g., Gwet, 2002). To see if a similar effect affects the statistic R_B , we ran simulations manipulating the importance of prevalence (from no difference in prevalence, $\omega = 1$, to a large difference, $\omega = 25$; intermediate levels were 2, 5, then by steps of 5). Because prevalence effects were mostly noted for the two-category classification situation ($k = 2$), we report results for $k = 2, 3$, and 5. We do not report C_C 's results anymore but report instead the observed rate of agreement r .

Figure 5 shows the results for two agreement rate parameters $\rho = 0.1$ and $\rho = 0.2$. As seen, the observed rate of agreement (r) increases with prevalence, *even though* internally the probability of an agreement is constant. The observed rate of agreement is so much affected by prevalence when $k = 2$ that the difference between the two true rates of agreement ($\rho = 0.1$ and $\rho = 0.2$) is impossible to see. On the other hand, κ is constant, reflecting accurately the true rates of agreement simulated. Thus, contrary to Gwet (2002), we conclude that κ is not biased downward; instead, it is r which is biased upward by the presence of prevalence.

As of R_B , we see that it is strongly affected by prevalence for $k = 2$, but not anymore for k greater than 2. This is caused by the fact that with high prevalence and small number of categories, the chance that some of the cells contain zero is important. In these cases, the result of Q_B is either 0 or ∞ , and R_B is either 0 or 1. For the condition $\omega = 25$, $\rho = 0.1$, $k = 2$, and $N = 125$, all the simulations returned $R_B = 0$ or 1.

Finally, P_A is affected by prevalence for all number of categories explored (2, 3, and 5).

General Discussion

The results show that the ratio is a key ingredient to compute global interrater agreement with sensitivity, as seen by the difference between C_C and R_B . Normalization should be avoided as it introduces uncontrollable biases. Finally, squaring is not useful to specificity as R_B is devoid of this operation.

R_B is the most reliable statistical test of global interrater agreement at this time. In particular, it is not making excessive Type I errors when presented with mixtures of agreement and disagreement. It is also unaffected by unequal prevalence when the number of categories exceeds 2. However, its confidence intervals are larger than

Figure 5. Mean estimates ± 1 standard deviation for four measures of interrater agreement (columns) as a function of imbalance factor (ω), the importance of the differences in prevalence and three number of categories (rows). Dark line is for the true rate of agreement $\rho_A = 0.1$ and the dark line for $\rho_A = 0.2$. The dashed line is the position of the null hypothesis.

needed when the data depart from H_0 , although this defect does not seem to hinder its power. Hence, a better expression of its variance could be envisioned to replace Equation (6). The κ measure, on the other end, has accurate confidence intervals. However, this index is very sensitive when a mixture of agreement and disagreements coexists. Therefore, this measure should only be used after global interrater agreement has been confirmed with R_B . Cohen's κ is an excellent descriptive statistic of positive agreement but is not recommended as a statistical test of global interrater agreement.

In sum, an appropriate sequence for analyzing interrater agreement would go as follows: First, perform a test on the agreement matrix using R_B to assess if there is global agreement. If so, then Cohen's κ may be used as an effect size measurement. If it is not the case, test the κ coefficient to see if it is significantly greater than 0. If such is the case, then one may infer that there is no *global* agreement but only *local* agreement for at least one category. In that case, the diagonal cell with the largest z score points to a category where there is significant agreement. It is not possible to tell anything regarding the remaining categories as there is no method at this time to decompose an R_B measure (unlike ANOVAs which can be decomposed using post hoc tests; the calculation of chi-square z components as in Equation (1) might lead to a solution but further work is required).

Note that this sequence is also valid if you found a nonsignificant R_B and a significant disagreement in κ (κ significantly less than 0). In this eventuality, there is no *global* interrater agreement (or disagreement), but there is *local* disagreement. Hence, the cell in the main diagonal with the most negative z score is one category with significant disagreement. Again, for the remaining categories, it is not possible to say anything at this time.

Take for example the data of Table 4. The R_B index is marginally significant which indicates an absence of global agreement (or a mild one; $R_B = 1.35$, $p = .0879$). The index κ , on the other hand, is significantly greater than 0 ($\kappa = 3.88$, $p < .0001$). The data indicate at least one category with significant agreement, and the most positive z score in the main diagonal is with respect to category 1 ($z = 2.509$). The most negative z score in the main diagonal is with regard to Category 3 ($z = -0.723$). This category is either showing a lack of agreement between the raters or disagreement.

Appendix

The Distribution of the Difference of the Absolute Value of Two Independent Variates Following a Normal Distribution

Let $\mathbf{X}$ and $\mathbf{Y}$ be independent normally distributed variates with mean zero and common standard deviation σ . Then, the distribution of the difference between $|\mathbf{X}|$ and $|\mathbf{Y}|$ is symmetrical with mean zero. The distribution function $F_{|\mathbf{X}|-|\mathbf{Y}|}(s) = Pr(|\mathbf{X}| - |\mathbf{Y}| \leq s) = Pr(\{\mathbf{X}, \mathbf{Y}\} \in S(s))$ where $S(s)$ is an area defined as in Figure

A.1, dependent on the sign of s . Because $\{\mathbf{X}, \mathbf{Y}\}$ is a bivariate normal distribution symmetrical about zero, the two areas of Figure A.1 can be rotated so that the sides are parallel to the axes and distant $s/\sqrt{2}$ from them. Also note that the areas are based on two regions of the same size. Hence,

$$F_{|X|-|Y|}(s) = \begin{cases} 2Pr(\mathbf{X} \leq \frac{s}{\sqrt{2}} \ \& \ \mathbf{Y} \leq \frac{s}{\sqrt{2}}) = 2(\Phi(\frac{s}{\sigma\sqrt{2}}))^2 & \text{if } s \leq 0 \\ 1 - 2Pr(\mathbf{X} \leq \frac{-s}{\sqrt{2}} \ \& \ \mathbf{Y} \leq \frac{-s}{\sqrt{2}}) = 1 - 2(\Phi(\frac{-s}{\sigma\sqrt{2}}))^2 & \text{if } s \geq 0 \end{cases}$$

from which we derive $f_{|X|-|Y|}(s)$ as usual,

$$f_{|X|-|Y|}(s) = \begin{cases} \frac{2\sqrt{2}}{\sigma} \Phi(\frac{s}{\sigma\sqrt{2}}) \phi(\frac{-s}{\sigma\sqrt{2}}) & \text{if } s \leq 0 \\ \frac{2\sqrt{2}}{\sigma} \Phi(\frac{-s}{\sigma\sqrt{2}}) \phi(\frac{s}{\sigma\sqrt{2}}) & \text{if } s \geq 0 \end{cases}$$

where Φ is the cumulative standard normal distribution and ϕ is the density of the standard normal distribution.

The mean of this distribution is zero, its variance is $2(1 - \frac{2}{\pi})\sigma^2$, its skew is 0, and its kurtosis excess is $\frac{\pi(3\pi-8)}{\pi-2} \approx 3.4346$. The 2.5% and 97.5% quantiles are -2.02 and 2.02, compared with -1.96 and 1.96 for the standard normal distribution. Thus, this distribution is little different from the normal distribution when used for statistical testing.

Figure A.1. Plot of $S(s)$ defined as the area where $|X| - |Y|$ is smaller or equal to s , based on the sign of s .

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.

Note

1. Q stands for quotient, which is French for ratio. We avoided R for ratio because this letter is often associated with correlation-like measures (between -1 and $+1$).

References

- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43, 551-558.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46. doi:10.1177/001316446002000104
- 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
- Everitt, B. S. (1968). Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 21, 97-103. doi:10.1111/j.2044-8317.1968.tb00400.x
- Feng, G. C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Qualitative Quantitative*, 47, 2959-2982. doi:10.1007/s11135-01209745-9
- Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327. doi:10.1037/h0028106
- Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment*, 1, 1-5.
- Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *Quantitative Methods for Psychology*, 10, 107-123.
- Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5, 93-112. doi:10.1080/19312458.2011.568376
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004) A call for standardization in content analysis reliability. *Human Communication Research*, 30, 434-437. doi: 10.1111/j.1468-2958.2004.tb00739.x
- 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
- 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: Routledge.