*Article*

# A Ratio Test of Interrater Agreement With High Specificity

Educational and Psychological Measurement
2015, Vol. 75(6) 979–1001

DOI: 10.1177/0013164415574086
epm.sagepub.com
SAGE

**Denis Cousineau[1] and Louis Laurencelle[2]**

**Abstract**

Existing tests of interrater agreements have high statistical power; however, they lack specificity. If the ratings of the two raters do not show agreement but are not random, the current tests, some of which are based on Cohen's kappa, will often reject the null hypothesis, leading to the wrong conclusion that agreement is present. A new test of interrater agreement, applicable to nominal or ordinal categories, is presented. The test statistic can be expressed as a ratio (labeled $Q_A$, ranging from 0 to infinity) or as a proportion (labeled $P_A$, ranging from 0 to 1). This test weighs information supporting agreement with information supporting disagreement. This new test's effectiveness (power and specificity) is compared with five other tests of interrater agreement in a series of Monte Carlo simulations. The new test, although slightly less powerful than the other tests reviewed, is the only one sensitive to agreement only. We also introduce confidence intervals on the proportion of agreement.

**Keywords**

interrater, agreement test, kappa

## Introduction

Quantifying interrater agreement is useful in contexts where two raters must judge into what categories a series of observations shall be classified. If agreement between raters is perfect, all judgments will be identical, whereas if both raters use completely

[1]Université d'Ottawa, Ottawa, Ontario, Canada
[2]Université du Québec à Trois-Rivières, Trois-Rivières, Quebec, Canada

**Corresponding Author:**
Denis Cousineau, École de psychologie, Université d'Ottawa, 136 Jean-Jacques Lussier, Ottawa, Canada K1N 6N5.
Email: denis.cousineau@uottawa.ca

**Table 1.** Example Data of Ratings Performed by Two Raters (Psychiatrist 1 and Psychiatrist 2) Appraising the Severity of Depression of 100 Patients ($N$ = 100) Using Three Categories of Depression ($k$ = 3).

| | | *Psychiatrist* 2 Category of depression | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | Row sums |
| *Psychiatrist 1* | 1 | 81 | 1 | 1 | 83 |
| Category of depression | 2 | 1 | 3 | 5 | 9 |
| | 3 | 1 | 5 | 2 | 8 |
| Column sums | | 83 | 9 | 8 | $N$ = 100 |

different criteria, agreement will occur only by chance. Agreement is often assessed using Cohen's kappa (Cohen, 1960), for which a value greater than 0.40 is commonly considered a moderate agreement (Landis & Koch, 1977). Recently, Kraemer, Periyakoil, and Noda (2002) explained that a Cohen's kappa for more than two categories is equal to a weighted average of individual kappas. The individual kappa $\kappa_j$ indicates the agreement of the two raters with respect to category $j$ only. Hence, if one category shows a very strong agreement but the others do not, the mean kappa may nevertheless be high.

As an example, consider the following situation in which two psychiatrists examine $N$ = 100 patients suffering from depression in order to appraise the category of depression (this example is inspired from von Eye, Schauerhuber, & Mair, 2006). The psychiatrists used $k$ = 3 categories of severity. Raters' categorizations are summarized in Table 1 using a $k \times k$ judgment matrix in which cell $\{i, j\}$ contains the number of observations that were classified as instances of the $i$th category by the first rater and of the $j$th category by the second one. Perfect agreement would result in the main diagonal summing to $N$, and off-diagonal cells containing zero observations.

As seen, both raters share the belief that most patients belong to the first category of severity. This high prevalence does not explain the large number of concordant judgments in that category, as chance would only predict approximately 70 agreement ratings for the first category. Hence, there is good agreement regarding Category 1. On the other hand, the two psychiatrists rarely agree on Categories 2 and 3. For example, Psychiatrist 2 found a total of 8 cases belonging to the third category of severity; of those, only 2 cases are also put in the third category by Psychiatrist 1, a figure below what chance would predict. Instead, the results seem to indicate that what looks like a Level 2 depression to Rater 1 is a Level 3 depression to Rater 2 and vice versa. This suggests disagreement between the two raters on these two categories.

**Table 2.** An Example Where Agreement for Three Categories Accompanies Disagreement for Two Categories.

| | | *Rater* 2 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Row sums $o_{i,*}$ |
| *Rater 1* | 1 | 8 | 2 | 1 | 2 | 4 | 17 |
| | 2 | 4 | 11 | 5 | 5 | 2 | 27 |
| | 3 | 2 | 1 | 5 | 12 | 7 | 27 |
| | 4 | 1 | 4 | 15 | 7 | 3 | 30 |
| | 5 | 4 | 6 | 2 | 4 | 10 | 26 |
| Column sums | $o_{*,j}$ | 19 | 24 | 28 | 30 | 26 | $N = 127$ |

*Note.* Contrary to Table 1, there is no significant difference in prevalence between the categories.

Overall, 86% of the cases are judged identically by the two raters (the cases found in the main diagonal). Should we conclude that, overall, the two raters are in good agreement? Cohen's kappa suggests a moderate but significant agreement ($\kappa = 0.528, z = 3.43, p < .001$), despite the fact that agreement is missing for two out of three categories.

Similarly, consider the results of Table 2. In this example, 127 cases were classified by two raters in one of 5 categories. Again, agreement is not clear. The raters seem to agree well regarding Categories 2 and 5. They also agree very well on Category 1 (nearly half of the ratings in this category are agreements). However, they agree little regarding Categories 3 and 4. Inspection of the data in Table 2, *outside* the diagonal, shows that the two raters have opposite interpretations regarding Categories 3 and 4: cases that are instances of Category 3 for Rater 1 are instances of Category 4 for the other. The two raters are not responding randomly, *but they are not agreeing*. Hence, whereas the rate of agreement is moderately good (32.3%), it does not mean that the two raters agree a little on *all* categories. The results instead suggest that the raters agree well on *a few* categories. Overall agreement is missing. Yet Cohen's kappa is weak but still significantly different from zero ($\kappa = 0.148, z = 3.30, p < .001$).

These two examples show the need for a statistic of agreement that can distinguish situations with much agreement for a few categories from situations with some agreement for most categories.

The purpose of this article is to present two new measures of interrater agreement, which we call $Q_{\mathrm{A}}$ and $P_A$. The first is more convenient as it uses the Fisher $F$ tables for critical values, whereas the second is possibly more intuitive, as it is a proportion between 0 and 1. In the next section, we describe these measures, which can be used on nominal or ordinal classifications. Next, we examine the reliability of these measures along with their confidence intervals. Finally, we assess the statistical power but, more important, the specificity of this approach, that is, the ability of a test based

on $Q_A$ (or $P_A$) to make the difference between mixtures of agreeing and disagreeing category ratings from solely agreeing ratings. The approach developed is akin to analysis of variance, as it partitions variances found in each cell (not just those in the main diagonal) as supporting agreement or supporting disagreement.

In the remainder of this article, the judgments are structured in the form of a square $k \times k$ table of observed frequencies. Observed frequencies in cell $\{i, j\}$ will be noted $o_{i,j}$. As per the $\chi^2$ test on contingency table, we note $e_{i,j}$ the expected frequency in cell $\{i, j\}$, estimated from the marginal frequencies with $e_{i,j} = o_{i,*} \times o_{*,j}/N$, where $o_{i,*}$ is the total of the $i$th column and $o_{*,j}$ is the total on the $j$th row.

Pearson's chi-square test of independence has been used on occasion to ascertain the significance of the agreement. However, this test only examines the null hypothesis ($H_0$) of by-chance co-occurrences in any $\{i,j\}$ cell. It can be seen as an omnibus test in the sense that any type of nonrandom structure can lead to the rejection of the null hypothesis. Therefore, as already noted by Cohen (1960), a significant $\chi^2$ statistic calculated from a judgment matrix indicates the presence of any type of category coupling between the two raters, whether they be in agreement (in the main diagonal) or disagreeing (anywhere else in the matrix).

The reader can find a brief description of five alternative tests of interrater agreements in Appendix A. These tests will be compared to the present approach in the subsequent section.

## Measures of the Global Structure of the Judgment Matrix

The new approach is a nonparametric one based on the same assumptions as Cohen's kappa (Brennan & Prediger, 1981). As will be seen, it is sensitive because it uses information outside the main diagonal as well as in the main diagonal. The crucial observation is that high agreement rates in the diagonal cells necessarily imply a shortage of cases outside the diagonal, and conversely. Hence, the whole matrix, not just the diagonal, is informative as to whether there is agreement or not.

The present approach partitions the matrix based on whether the observed cell frequencies deviates from chance (judged by whether the observed count is different from its expected value) and whether such ratings are supportive of agreement or not.

More formally, under $H_0$ (no agreement), the observed frequencies $o_{i,j}$ in the $k^2$ cells of the judgment matrix should fluctuate near their expected values $e_{i,j}$. Hence, about half of the cells should have a frequency count above their respective $e_{i,j}$, and the other half, below.[1] As in the $\chi^2$ test, the difference between observed and expected frequencies is standardized with $z_{i,j} = (o_{i,j} - e_{i,j})/\sqrt{e_{i,j}}$, which under $H_0$ and for reasonably large counts is normally distributed.

The first statistic, $Q_A$, is a ratio of cells favorable to agreement against cells unfavorable to agreement, the former being those that (i) on the main diagonal have counts higher than expected and (ii) off the main diagonal have counts smaller than expected, whereas the latter are those that (iii) on the main diagonal have counts

smaller than expected and (iv) off the main diagonal have counts larger than expected. Hence, $Q_{\mathrm{A}}$ is computed as

$$\begin{aligned} Q_A &= \frac{\text{Sum of } z^2 \text{ over cells supporting agreement}}{\text{Sum of } z^2 \text{ over cells supporting disagreement}} \\ &= \frac{(\mathrm{i})+(\mathrm{ii})}{(\mathrm{iii})+(\mathrm{iv})} \\ &= \frac{\sum_{i=1}^{k}(z_{i,i}^{+})^2 + \sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}(z_{i,j}^{-})^2}{\sum_{i=1}^{k}(z_{i,i}^{-})^2 + \sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}(z_{i,j}^{+})^2} \end{aligned} \tag{1}$$

where the terms in the sums (the components) will be added depending on the sign of the difference between the observed frequency and the expected frequency using

$$z_{i,j}^{s} = \begin{cases} (o_{i,j} - e_{i,j})/\sqrt{e_{i,j}} & \text{if } Sign(o_{i,j} - e_{i,j}) = s \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$Q_A$ values range from 0 to $+\infty$, and under $H_0$ its central value is about 1. On average under, $H_0$ the numerator should contain entries from half of the matrix and is therefore on average a sum of $k^2/2$ squared $z$ scores. Its distribution is thus close to the $\chi^2$ distribution with $(k-1)^2/2$ degrees of freedom (*df*). The same is true for the denominator. As a first approximation, the distribution of $Q_A$ is therefore a Fisher $F$ ratio distribution with both *df* equal to $(k-1)^2/2$. Hence, critical values for $Q_A$ can be read from the right tail of the Fisher $F$ ratio distribution.

The second statistic, $P_A$, is the proportion of variance supporting agreement onto total variance. It can be obtained from the $z$ scores or from $Q_A$ with

$$\begin{aligned} P_A &= \frac{(\mathrm{i})+(\mathrm{ii})}{(\mathrm{i})+(\mathrm{ii})+(\mathrm{iii})+(\mathrm{iv})} \\ &= \frac{Q_A}{1+Q_A} \end{aligned}$$

The statistic $P_A$ is akin to the eta squared ($\eta^2$) statistic and ranges from 0 to 1. When the ratings are random, its central value is ½, that is, 50% of the variance in the ratings suggest agreement and the other 50% of the variance suggest disagreement. Hence, a value of ½ supports neither interpretations and indicates random ratings. As found in the literature (e.g., Forbes, Evans, Hastings, & Peacock, 2010), a ratio of the form $\mathbf{X}/(1+\mathbf{X})$, where $\mathbf{X}$ follows a Fisher $F$ ratio distribution has a standard beta distribution with parameters $a$ and $b$ dependent on the *df* of the $F$ ratio distribution, here $a = b = (k-1)^2/4$.

$P_A$ and $Q_A$ are totally interchangeable. To detect whether the ratings deviate from random, one option is to use $Q_A$. The null hypothesis is $H_0$: $Q_A = 1$, and the decision

**Table 3.** Interpretation of the Statistics $P_A$.

| $P_A$ | Agreement is . . . |
|---|---|
| 0.40 . . . 0.60 | weak or absent |
| 0.60 . . . 0.70 | fair |
| 0.70 . . . 0.80 | moderate |
| 0.80 . . . 0.90 | strong |
| $>$ 0.90 | outstanding |

rule is to reject $H_0$ if the observed $Q_A$ exceeds a critical value read on an $F$ table with $(k-1)^2/2, (k-1)^2/2$ degrees of freedom. If $P_A$ is used, the null hypothesis is $H_0$: $P_A$ = ½ and the decision rule is to reject $H_0$ if the observed $P_A$ exceeds a critical value read from a Beta distribution with parameters $a$ and $b$ equal to $(k-1)^2/4$. Note that, in principle, these tests could be left tailed, testing for significant disagreement between the raters, but results to be shown will indicate that they should not be used for that purpose. In between significant agreement and significant disagreement, there is a zone in which the ratings are inconclusive either way, as is the case for random ratings.

Under the null hypothesis, approximately half of the cells, $k^2/2,$ should contribute to the numerator of these statistics and the other half to the denominator. If $P_A$ is larger than ½ ($Q_A$ larger than 1), it suggests that more than one half of the cells contribute to the numerator and less than one half to the denominator. Hence, the distribution of $P_A$ for an observed result different from ½ has Beta distribution with degrees of freedom changed to $a = P_A(k-1)^2/2$ and $b = (1-P_A)(k-1)^2/2$. From this observation, confidence intervals can be obtained with

$$CI_{1-\alpha} \text{ of } P_A = [B_{a,b}(\alpha/2), B_{a,b}(1-\alpha/2)]$$

where $B_{a,b}$ denotes the quantile function of the Beta distribution with its parameters. The following section will verify this assertion.

Appendix B shows how $Q_A$ and $P_A$ can be computed with the SPSS statistical package. It also shows how to get the $p$ value of the test and confidence intervals for $P_A$ for any level $1-\alpha$.Table 3 give some indications on how to interpret $P_A$.

## An Illustration With Computation of Formulas

We illustrate the computation of the $Q_A$ statistic with an example involving two raters having to examine and codify 200 observations into a system of $k$ = 5 categories, labelled 1 to 5. The judgment matrix and the marginal sums $o_{i,*}$ and $o_{*,j}$ are shown in Table 4, top part; the second part of Table 4 presents the expected theoretical frequencies $e_{i,j} = o_{i,*} \times o_{*,j}/N$; finally, the standardized differences $z_{i,j} = (o_{i,j} - e_{i,j})/\sqrt{e_{i,j}}$ are shown in the third part of Table 4.

**Table 4.** Cross-Classification Data From Two Simulated Raters, Indicating (Top) the Number of Ratings $o_{i,j}$ in Each *i, j* Cell, (Middle) Its Corresponding Theoretical Value $e_{i,j}$, and (Bottom) the Standardized Differences $z_{i,j}$ ($N = 200$).

Observed frequencies $o_{i,j}$

| | | *Rater* 2 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Row sums $o_{i,*}$ |
| *Rater 1* | 1 | 7 | 5 | 2 | 1 | 3 | 18 |
| | 2 | 5 | 13 | 10 | 7 | 8 | 43 |
| | 3 | 11 | 4 | 15 | 6 | 9 | 45 |
| | 4 | 8 | 11 | 7 | 9 | 6 | 41 |
| | 5 | 11 | 5 | 15 | 6 | 16 | 53 |
| Column sums | $o_{*j}$ | 42 | 38 | 49 | 29 | 42 | $N = 200$ |

Expected frequencies $e_{i,j}$

| | | *Rater* 2 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Row sums $e_{i,*}$ |
| *Rater 1* | 1 | 3.78 | 3.42 | 4.41 | 2.61 | 3.78 | 18 |
| | 2 | 9.03 | 8.17 | 10.53 | 6.24 | 9.03 | 43 |
| | 3 | 9.45 | 8.55 | 11.02 | 6.53 | 9.45 | 45 |
| | 4 | 8.61 | 7.79 | 10.05 | 5.95 | 8.61 | 41 |
| | 5 | 11.13 | 10.07 | 12.98 | 7.68 | 11.13 | 53 |
| Column sums | $e_{*,j}$ | 42 | 38 | 49 | 29 | 42 | $N = 200$ |

Standardized differences $z_{i,j}$

| | | *Rater* 2 | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| *Rater 1* | 1 | +1.656 | +0.854 | −1.148 | −0.997 | −0.401 |
| | 2 | −1.341 | +1.690 | −0.165 | +0.306 | −0.343 |
| | 3 | +0.504 | −1.556 | +1.197 | −0.206 | −0.146 |
| | 4 | −0.208 | +1.150 | −0.961 | +1.253 | −0.889 |
| | 5 | −0.039 | −1.598 | +0.559 | −0.608 | +1.460 |

In this example, the component (i), $\sum_{i=1}^{k}\left(z_{i,i}^{+}\right)^2$, includes all 5 diagonal cells, every cell being positive, and is $1.656^2 + 1.690^2 + 1.197^2 + 1.253^2 + 1.460^2 \approx 10.73$. Its counterpart, (iii), $\sum_{i=1}^{k}\left(z_{i,i}^{-}\right)^2$, is therefore zero. For the off-diagonal cells, the sum of negative-signed components (ii) is $\sum_{i=1}^{k}\sum_{j\neq i}\left(z_{i,j}^{-}\right)^2 = (-1.148)^2 +$

**Table 5.** Results of the Five Tests of Agreements and the $\chi^2$ Test on the Examples of Tables 1, 2, and 4.

| | Table 1 | | Table 2 | | Table 4 | |
|---|---|---|---|---|---|---|
| | Test result | *p* | Test result | *p* | Test result | *p* |
| *r* | 0.860 | | 0.323 | | 0.300 | |
| *k* | 0.528 | | 0.148 | | 0.125 | |
| Tests on kappa | | | | | | |
| $z_{k1}$ | 3.43 | <.01 | 3.30 | <.01 | 3.54 | <.001 |
| $z_{k2} = z_{k3}$ | 6.86 | <.01 | 3.31 | <.01 | 3.58 | <.001 |
| Tests on the diagonal elements | | | | | | |
| $z_{\Sigma 1}$* | 9.12 | <.001 | 3.10 | <.02 | 3.16 | <.001 |
| $z_{\Sigma 2}$ | 3.23 | <.002 | 3.42 | <.01 | 3.25 | <.001 |
| Tests on the structure of the matrix | | | | | | |
| $\chi^2$ | 8.20 | n.s. | 57.6 | <0.001 | 25.0 | n.s. |
| $Q_A$ | 0.62 | n.s. | 2.50 | n.s. | 8.23 | <.01 |
| Exact *p* value | .616 | | .108 | | .004 | |
| $P_A$ | 0.38 | | 0.714 | | 0.892 | |
| 95% CI of $P_A$ | [.01, .94] | | [0.38, 0.95] | | [0.62, 0.99] | |

*Note.* *Denotes the test assuming the indifference principle. n.s. = nonsignificant at the .05 level.

$(-0.997)^2 + \ldots + (-0.608)^2 \approx 11.58$, and the sum of positive-signed components (iv) is $\sum_{i=1}^{k} \sum_{j \neq i} \left(z_{i,j}^{+}\right)^2 = (0.854)^2 + (0.306)^2 + \ldots + (0.559)^2 \approx 2.71$. $Q_{\mathrm{A}}$ is therefore

$$Q_A = \frac{10.73 + 11.58}{0 + 2.71} \approx 8.23$$

This value, compared to a $F$ critical value at $\alpha = .05$, $F(8, 8) = 3.438$, suggests that the raters are in agreement, $Q_{\mathrm{A}}(8, 8) = 8.23$, $p < .05$.

Note that the sum of all the components of $Q_{\mathrm{A}}$, $10.73 + 11.58 + 0.00 + 2.71$, equals the usual chi-square test, 25.03. With its $(k-1)^2 = 16$ degrees of freedom, the critical value being 26.30, we would infer that the data pattern observed in Table 4 does not exhibit any structure and may be ascribed to a pure chance mechanism, $\chi^2(16) = 26.30$, $p > .05$. The chi-square test does not detect agreement in Table 4 because, not being specifically suited for agreement detection, it lacks power and specificity (as our Monte Carlo experiments will confirm).

The second statistic, $P_A$, is obtained with $8.23/(1 + 8.23) = 0.892$, well above the reference value of ½, suggesting a strong agreement. To get confidence intervals, computer software must be used as critical values for the Beta distribution are not given in statistics textbooks. We obtain a 95% confidence interval of [0.62, 0.99], not including ½.

For this example, five statistical tests (the four from Appendix A and $Q_A$) make the same decision, rejecting the null hypothesis even at the 0.01 level. However, the same does not happen for the previous examples of Tables 1 and 2 where $Q_{\mathrm{A}}$ alone

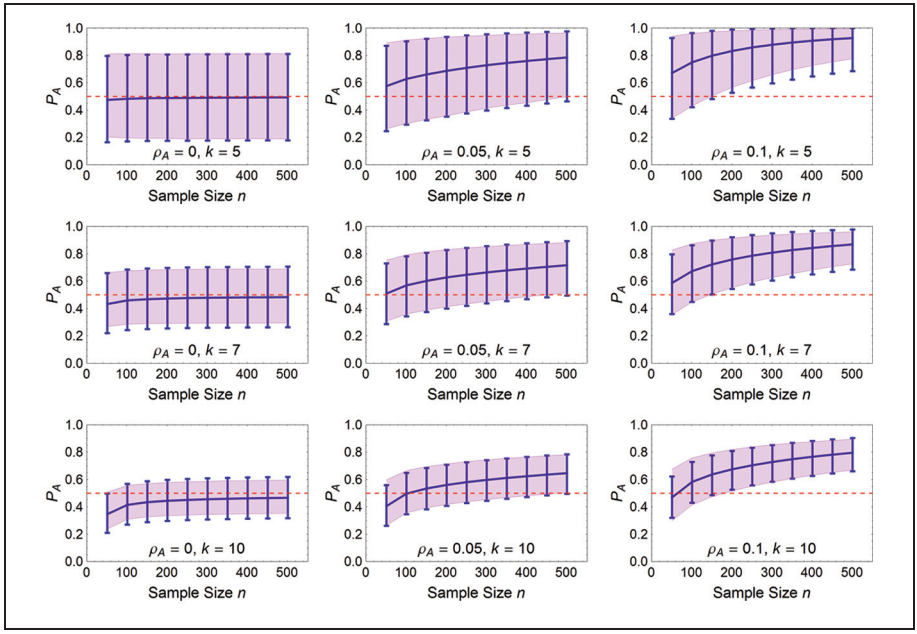


**Figure 1.** Mean $P_A$ (thick line), 95% confidence intervals (error bars), and spread where 95% of the simulated $P_A$ fell (gray area) as a function of sample size $N$, when $\rho_A$ is varied from 0 to 0.10 (columns) and $k$ is varied from 5 to 10 (rows). We see that $P_A$ is biased downward in the first column: with no effect, mean $P_A$ should be equal to zero for all $N$ and all $k$.

concludes that agreement is absent. Table 5 summarizes the results for the examples of Tables 1, 2, and 4.

## Reliability of the Statistics and Their Distributions

To examine the merit of the present approach, we explored the statistic $P_A$ using Monte Carlo simulations (the same results were obtained exploring $Q_A$). In particular, we examined if the scores' distributions correspond to their theoretical counterparts. To do so, we generated agreement matrices with various amount of true agreement and checked whether the theoretical 95% confidence intervals contained the results of 95% of the simulations (using the same methodology as in Harding, Tremblay, & Cousineau, 2014). We also checked 99% and 99.9% confidence intervals but the results were comparable and so we do not report the findings.

We manipulated the sample sizes ($N$, from 50 to 500 by increment of 50), the number of categories $k$ (from 4 to 25), and the true probability of an agreement, $\rho_A$. Details are given in Appendix C.

A subset of the results is shown in Figure 1. The first thing to note is that the theoretical confidence intervals (shown using error bars) match very closely the limits in which 95% of the simulated $P_A$ rest (shown using gray areas). The theoretical error

bars are the least accurate for small number of categories and strong effect size where they overestimate the spread of the results.

Less visible but more critical, the average $P_A$ (as well as the average $Q_A$, not shown) is biased downward, underestimating the strength of the agreement. This bias is visible in the left panels where there is no agreement ($\rho_A = 0$), more so for large numbers of categories (e.g., $k > 8$) and for small sample sizes. We could not find a simple way to undo the bias, which would be suitable to all $k$. Because of the downward bias, these measures are biased toward detecting disagreement and against detecting agreement. This is why these measures should not be used for detecting disagreement (performing a left tail test of $Q_A$ or $P_A$). The bias results in a conservative test (i.e., less powerful than it could be). However, as we will see next, this has little impact on its specificity.

## Sensitivity and Specificity of the Test

In order to evaluate the relative merits of the present test of agreement, we ran three more series of Monte Carlo experiments, comparing our ratio test to four other tests of agreement. In the first series, we explored statistical power by manipulating the true rate of agreement $\rho_A$ from 0 to 1. In all the simulations, we manipulated the number of categories $k$ (from 5 to 25), the total sample sizes $N$ (from 50 to 500), and the significance levels ($\alpha$, .05 and .01). The statistical tests to be compared (described in Appendix A) are the following:

$z_{\kappa 1}$, Cohen's simple $z$ test
$z_{\kappa 2}$, a $z$ test using Fleiss, Cohen, and Everitt's (1969) more accurate variance approximation
$z_{\Sigma 2}$, a sum-of-$z$ test not assuming the indifference principle
$Q_A$, the ratio of agreement test (equivalent to $P_A$)
$\chi^2$, the standard $\chi^2$ omnibus test of independence

We included the $\chi^2$ test even though it is not a test of agreement to provide a reference. We do not report results for $z_{\kappa 3}$ (described in Appendix A) as its results are undistinguishable from $z_{\kappa 2}$ nor do we report results from $z_{\Sigma 1}$ (for reasons explained in Appendix C).

Apart from being powerful, a good test should also be mostly, if not exclusively, sensitive to agreeing judgments, and not to any other type of consistent categorizations. For example, if observations classified as instances of category 2 by the first rater are systematically put in Category 3 by the second rater (as seen in Table 1), this is a consistent pairing, but not an agreement, and the tests *should not* reject $H_0$ based on situations of this type. The two raters, at least for these two categories, are consistent, but they nevertheless disagree. Specificity is that property of a statistical test in virtue of which it is most sensitive to data configurations relevant to its intended rejection condition but keeps relatively insensitive to other configurations. A test

having high specificity should not reject $H_0$ at a rate higher than the significance level $\alpha$ when the hypothesized condition of rejection is not present.

The purpose of the last two series of Monte Carlo experiments is to evaluate the specificity of the tests. As before, we manipulated the parameters $\rho_C$, $k$, and $N$. However, in the present context, the parameter $\rho_C$ represents the *rate of consistent categorizations* between raters. This rate is large when raters have consistent decision rules even if these rules are not in agreement. Overall, with large $\rho_C$, cases will be more frequent in cells not necessarily on the main diagonal.

In these two series of simulations, there is no overall agreement, but possibly accidental agreements for one or a few categories. Hence, the probability of rejecting $H_0$ should be small, otherwise the test is making too many Type I errors. Ideally, the probability of Type I errors should be equal to the significance level (5% in the subsequent figures) and fairly constant across rates of consistent categorizations.

To control precisely for the presence of accidental agreements, the two series present two conditions. In the *Random with possible coincident pairings* condition, Category ''$x$'' from Rater 1 is coupled to some Category ''$y$'' of Rater 2, where $y$ may be any category among 1 to $k$, including by chance the same as Category $x$ (an agreement for this category). In the *Random excluding coincident pairings* condition, there cannot be a single agreeing pair in the main diagonal.

A detailed description of the simulations' parameters and algorithms is given in Appendix C.

## Results

Because of the very large number of conditions explored, we only report illustrative results, shown in Figures 2 to 4, all using a significance level of .05; these results are typical of what was found in the other conditions and with a significance level of .01. Figure 2 presents the results for one set of parameter ($k$ = 5 and $N$ = 125), one panel per condition. In Figure 3, everything is the same as in Figure 2, except that the sample size is doubled ($N$ =250); in Figure 4, everything is the same as in Figure 2, except that the number of categories is doubled ($k$ = 10).

As seen, the results follow three patterns.

*Chi-Square Test.* As expected, all our results disqualify the $\chi^2$ test as a test of agreement. First, this test is the least powerful. In Figure 2, top panel, for example, the four other tests reach a power of 50% at $\rho_A \approx 0.07$ to 0.09, while chi-square needs to wait until $\rho_A \approx$ 0.15 to attain it. Second, the $\chi^2$ test is the least specific, rejecting $H_0$ even when there is not a single coincident pairing (*Random excluding coincident pairings* condition).

*Cohen's ($z_{\kappa 1}$), Fleiss' ($z_{\kappa 2}$), and the Sum-of-z ($z_{\Sigma 2}$) Tests.* For the next three tests, Cohen's simple formula ($z_{\kappa 1}$), Fleiss et al.'s ($z_{\kappa 2}$), and $z_{\Sigma 2}$ behave roughly in the same manner, be it for the power (top panels of Figures 2 to 4) or the specificity results (middle and lower panels of Figures 2 to 4). It is interesting to note that Fleiss et al.'s intricate formula for the variance of $\kappa$, $var_2(\kappa)$, despite its good

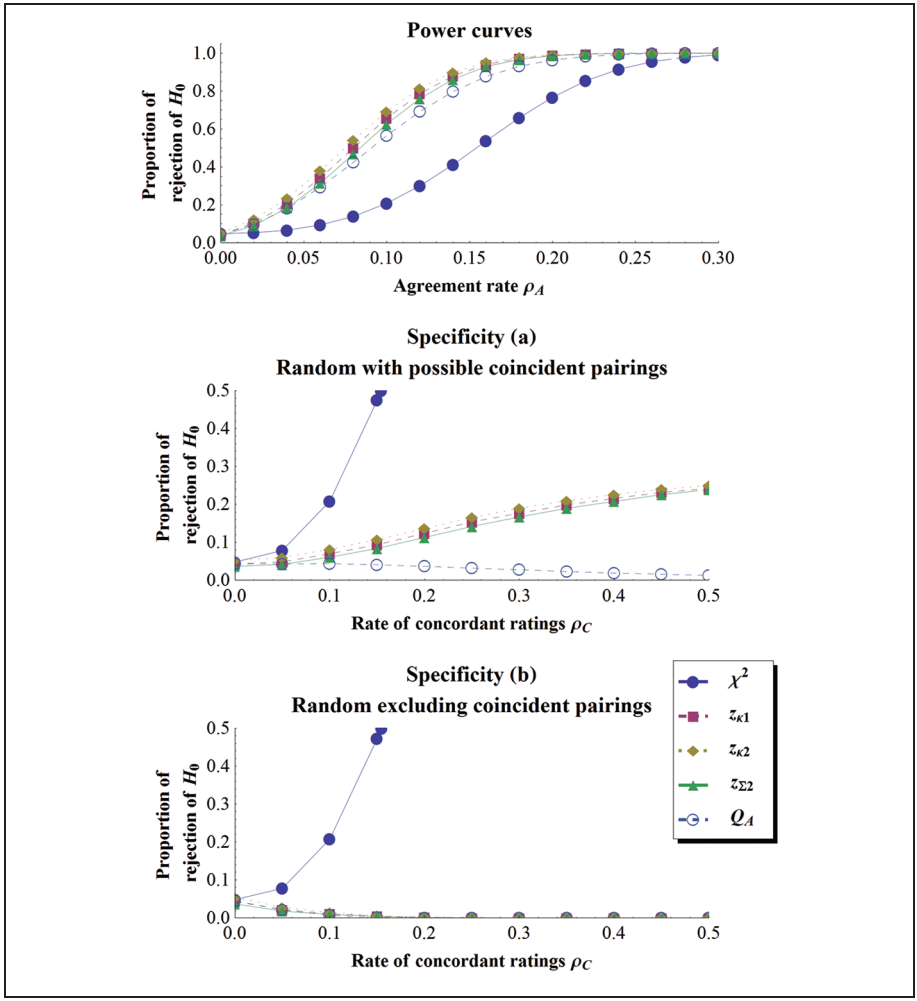


**Figure 2.** (Top panel) Power curve for 5 tests of interrater agreement as a function of the true agreement rate ($\rho_A$) when $k$ = 5 and $N$ = 125; (Middle and bottom panels) specificity for the same tests in the same condition as a function of the rate of consistent categorizations ($\rho_C$).

performances, does not seem worth the effort as compared to Cohen's much simpler one. The three tests display the best performance on power, surpassing the new $Q_A$ test, more so when the number of categories is large. However, they are also markedly *worse* than the $Q_A$ test regarding specificity (sometimes by as much as 40%). Hence, when the raters make disagreeing judgments in a consistent manner, those tests will spuriously signal agreement, an unwanted behavior. Indeed, under the condition of *Random with possible coincident pairings* (middle panels of Figures 2, 3, and 4), $z_{\kappa 1}$ (like its variant $z_{\kappa 2}$) and $z_{\Sigma 2}$ react positively to $\rho_C$, the rate of consistent

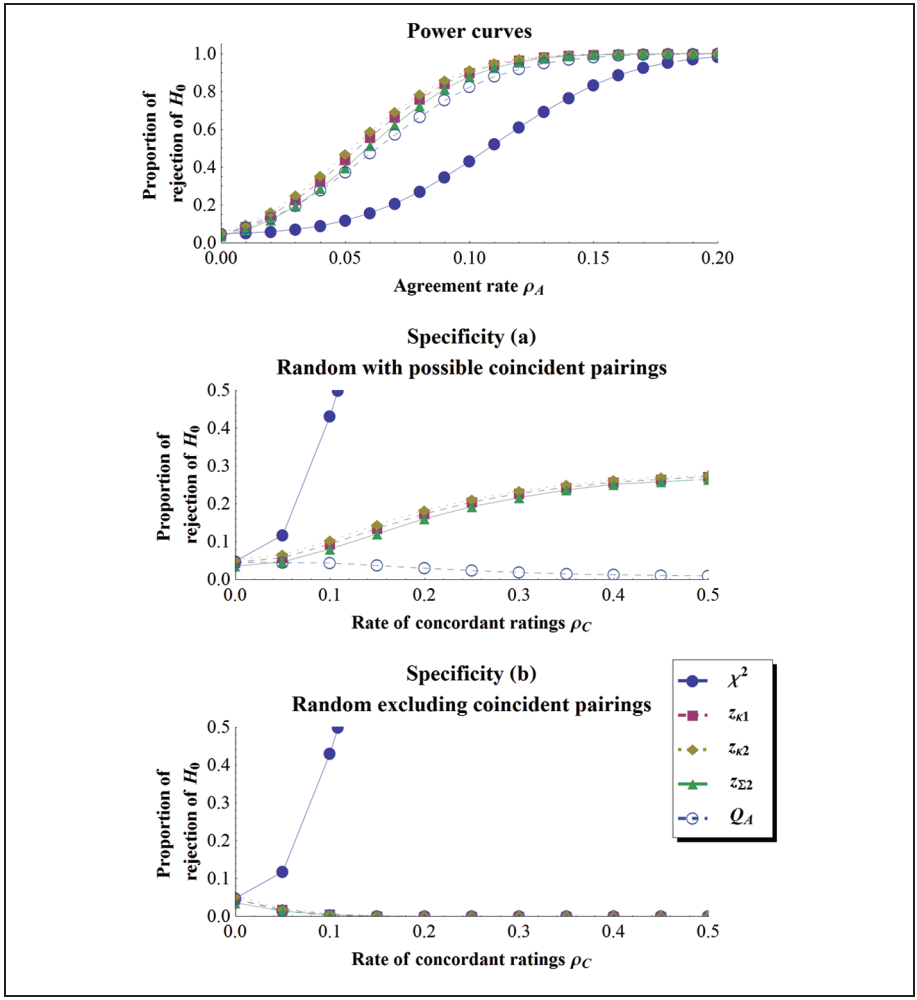


**Figure 3.** Power curve and specificity curves when $k$ = 5 and $N$ = 250 in the same format as Figure 2.

categorizations, eating away its specificity. Only in conditions where not a single coincident pair is allowed (bottom panels of Figures 2, 3, and 4) do the three tests perform almost identically to the $Q_A$ test.

*$Q_A$ Test.* As can be seen in the top panels of Figures 2 and 3, power curves of the $Q_A$ test follow well the power curves of the $z_{\kappa 1}$, $z_{\kappa 2}$, and $z_{\Sigma 2}$ tests more so if the number of categories is smaller. In the worst case presented ($k$ = 10, top panel of Figure 4), a power of 50% is attained by $Q_A$ at $\rho \approx 0.046$, and by Cohen's test at $\rho \approx 0.026$. This loss of power continues to increase as $k$ increases. On the other hand, the specificity

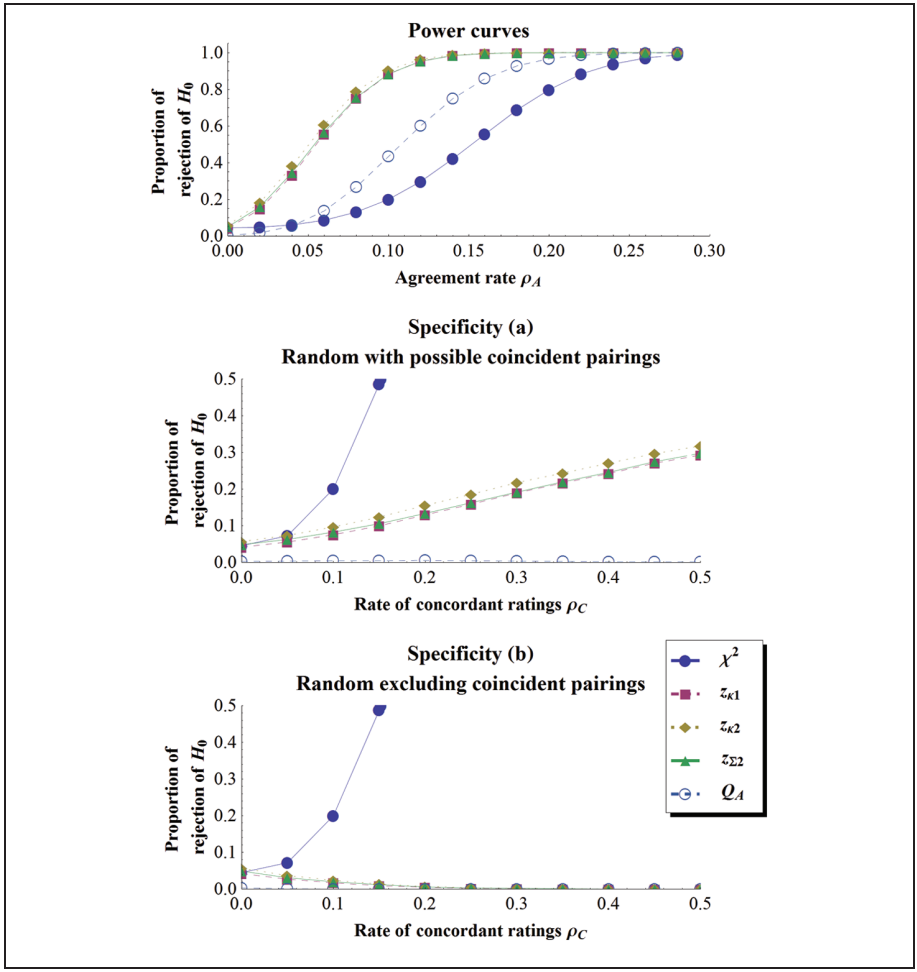


**Figure 4.** Power curve and specificity curves when $k$ = 10 and $N$ = 125 in the same format as Figure 2.

curves present a remarkable stability for $Q_A$: the significance rate in all cases starts at the alpha value (e.g., .05) or below and declines steadily as the rate of consistent categorizations increases.

The $Q_A$ test considers all circumstances in which the raters may agree and disagree. Hence, it rejects the null hypothesis only if the evidence gathered from the agreeing cells outweighs the evidence from the disagreeing cells. Consistent pairings outside the main diagonal is thus for $Q_A$ a strong cue against agreement.

In sum, the new test $Q_A$ was found to be just a little less powerful than the other tests when $k$ is small ($k < 8$). However, these tests proved to be far less specific. One possibility is that on a small percentage of the simulations, random pairings might

have occurred, triggering these tests to significance but not $Q_A$. Hence, it is not clear whether the new test is truly less powerful or is simply more selective.

## General Discussion

We examined the performance of five tests for determining whether two observers, each one classifying $N$ observations into a set of $k$ categories, agree or not above chance level. Monte Carlo simulations were used to compare the five tests with respect to their sensitivity to true agreement—a quality that translates into statistical power—and their specificity, that is, their insensitivity to anything but true agreement. As could be expected (see for instance Cohen, 1960), the standard $\chi^2$ statistic on contingency tables, although not insensitive to true agreement, ranks well behind the other tests for power as well as for specificity. The $\kappa$-to-$z$ tests (Cohen's $z_{\kappa 1}$, Fleiss et al.'s $z_{\kappa 2}$) and $z_{\Sigma 2}$ perform best on power, but they are impaired by their poor specificity, a problem that does not affect the ratio-based $Q_A$ test.

The statistics described here provide a nice set of tools to quantify agreement and assess its significance. Confidence intervals are also defined, allowing for easy comparisons between studies. The only limitation of the present statistics is that they are all biased downward (and the bias is important for $k \geq 10$ and small $N$). Future work should try to find the correction to this limitation. In the meanwhile, the present statistics should not be used to detect significant disagreement.

Our results show that a significant test on kappa can indicate that agreement is strongly present in a few categories or weakly present in all categories (or a continuum between these extremes). On the other hand, a significant $Q_A$ test only indicates that agreement is present in all the categories. The simulations showed that the $Q_A$ test is the only test that is sensitive to agreement within all categories, not just a few, as seen by comparing the markedly different results in the *Random excluding coincident pairings* with the *Random with possible coincident pairings* simulations.

What does distinguish the $\kappa$-to-$z$ and sum-of-$z$ statistics, on the one hand, and $Q_A$, on the other, and wherefrom does the latter earn its high specificity? The uniqueness of $Q_A$ comes from the fact that the whole matrix of agreement is used. Kappa-based and sum-of-$z$ tests only exploit the main diagonal. Hence, they implicitly assume that information outside it is uninformative, which is obviously false, as we discussed with the first two examples.

## Appendix A

### *Existing Tests of Interrater Agreement*

In what follows, we note the *rate of agreement* ($r$) as the proportion of observed frequencies in the diagonal cells, $r = \left(\sum_{i=1}^{k} o_{i,i}\right)/N$, and $E(r)$ the *expected rate of agreement* that would arise by pure chance, $E(r) = \left(\sum_{i=1}^{k} e_{i,i}\right)/N$, where again

$e_{i,j} = o_{i,*} \times o_{*,j}/N$ are computed as in the $\chi^2$ test of independence using the row sums $o_{i,*} = \sum_{j=1}^{k} o_{i,j}$ and the column sums $o_{*,j} = \sum_{i=1}^{k} o_{i,j}$.

## Two Tests Based on Cohen's Kappa

Cohen (1960) proposed a descriptive measure for interrater agreement, the kappa coefficient ($\kappa$). This coefficient indicates the relative rate of above-chance agreement between raters. It is therefore an alternative to $r$. This coefficient, defined as

$$\kappa = \frac{r - E(r)}{1 - E(r)} \tag{A.1}$$

should be near zero under the null hypothesis of only chance agreement.

Cohen (1960, Equation 9) also presented an approximation for the variance of $\kappa$ under the null hypothesis,

$$var_1(\kappa) \approx \frac{E(r)}{N(1 - E(r))} \tag{A.2}$$

from which a $\kappa$-to-$z$ transformation

$$z_{\kappa 1} = \kappa / \sqrt{var_1(\kappa)} \tag{A.3}$$

can be used for testing the null hypothesis that $\kappa = 0$. If we ignore the $\pm$½ continuity correction, this test is equivalent to a binomial test of proportion comparing the proportion $r$ against its hypothesized counterpart $E(r)$.[2]

Everitt (1968) derived the exact (but somewhat involved) formula for the variance of $\kappa$, whereas Fleiss et al. (1969, Equation 14) derived this more reliable approximation:

$$var_2(\kappa) \approx \frac{1}{N(1 - E(r))^2} \left( \sum_{i=1}^{k} p_{i,*} p_{*,j} (1 - p_{i,*} - p_{*,j})^2 + \sum_{i=1}^{k} \sum_{j \neq i} p_{i,*} p_{*,j} (p_{i,*} + p_{*,j})^2 - E(r)^2 \right) \tag{A.4}$$

in which $p_{i,*} = o_{i,*}/N$ and $p_{*,j} = o_{*,j}/N$ are the row and column observed proportions found in the judgment matrix. Using this second approximation, a $z$ test of the null hypothesis that population $\kappa$ equals zero can be proposed using the observed $\kappa$,

$$z_{\kappa 2} = \kappa / \sqrt{var_2(\kappa)} \tag{A.5}$$

This is the version of the kappa test implemented in SPSS using the CROSSTABS command.

Finally, if one is willing to compute the exact variance of observed $\kappa$ found by Everitt (1968), the exact test of the null hypothesis can be derived (let us call it $z_{\kappa 3} = \kappa / \sqrt{var_3(\kappa)}$). However, the gain in precision relative to $z_{\kappa 2}$ is immaterial.

**Table A.1.** Seven Test Formulas for Assessing Agreement From a Judgment Matrix.

| Name and formula | Null hypothesis | Reference |
|---|---|---|
| *k*-to-*z* tests | | |
| $z_{\kappa 1} = \kappa / \sqrt{var_1(\kappa)}$ | $H_0$: $\kappa = 0$ | Cohen (1960) |
| $z_{\kappa 2} = \kappa / \sqrt{var_2(\kappa)}$ | $H_0$: $\kappa = 0$ | Fleiss et al. (1969) |
| Sum-of-*z* tests | | |
| $z_{\Sigma 1} = \sqrt{\frac{N}{k}}(k \cdot r - 1)$ | $H_0 : \sum_{i=1}^{k} o_{i,i} = N/k$ | von Eye et al. (2006)[a] |
| $z_{\Sigma 2} = \sum_i z_{i,i}\sqrt{k}$ | $H_0 : \sum_{i=1}^{k} o_{i,i} = \sum_i e_{i,i}$ | This article |
| Tests on the structure of the judgment matrix | | |
| $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} z_{i,j}^2$ | $H_0 : o_{i,j} = o_{i,*} \cdot o_{*,j}/N$ | Pearson (1932) |
| $Q_A = \frac{\sum_{i=1}^{k}(z_{i,i}^{+})^2 + \sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}(z_{i,j}^{-})^2}{\sum_{i=1}^{k}(z_{i,i}^{-})^2 + \sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}(z_{i,j}^{+})^2}$ | $H_0 : Q_A = 1$ | This article |
| $P_A = \frac{Q_A}{1+Q_A}$ | $H_0 : P_A = 1/2$ | This article |

*Note.* $k$ is the number of categories and $N$ is the number of observations classified.
[a]This test, based on the indifference principle, is not included in the simulations reported.

## *Two Tests on the Diagonal Elements*

The above tests are based on a binomial distribution. They are then converted to a $z$ score using the normal approximation to the binomial. We now examine two more tests that are built using a sum-of-$z$ approach (sometimes called a *pooled test*). It is based on the idea that the sum of a number $k$ of independent $z$ scores is also a $z$ score with mean zero and variance $k$. The sum-of-$z$ test of agreement was first proposed by von Eye et al. (2006). It relies on the assumption that each entry in the main diagonal should contain only random agreements unaffected by the prevalence of the categories (the *indifference principle*). The test (referred to as Stouffer's test in von Eye et al., 2006; see Stouffer, Suchman, DeVinney, Star, & Williams, 1949) is given by

$$z_{\Sigma 1} = \frac{\sum_{i=1}^{k} z_{i,i}}{\sqrt{k}} \tag{A.6}$$

in which $z_{i,i} = \frac{o_{i,i} - N/k^2}{\sqrt{N/k^2}}$. This test simplifies to the following:

$$z_{\Sigma 1} = \sqrt{\frac{N}{k}}(k \cdot r - 1) \tag{A.7}$$

A generalization inspired by the above test is one in which the indifference principle is not assumed; the observed count in each cell of the diagonal $o_{i,i}$ is compared to its corresponding probable value $e_{i,i}$ under $H_0$. The formula is therefore

$$z_{\Sigma 2} = \frac{\sum_{i=1}^{k} z_{i,i}}{\sqrt{k}} \tag{A.8}$$

where $z_{i,i} = (o_{i,i} - e_{i,i})/\sqrt{e_{i,i}}$. Contrary to $z_{\Sigma 1}$, this formula cannot be further simplified.

Table A.1 summarizes all the tests reviewed in this appendix.

## Appendix B

### *Computing Agreement With SPSS*

Researchers can compute Cohen's kappa and the $z_{\kappa 2}$ test using SPSS. However, there are no computer packages that can compute the sum-of-$z$ tests or the $Q_A$ test. If the *Essentials for Python* extension to SPSS is installed (available free from the SPSS website), the following BEGIN PROGRAM END PROGRAM block will compute $Q_A$. The data editor must contain only the ratings and there must be as many lines as there are columns (i.e., the ratings must be in the form of an agreement matrix). To enter the commands, open a new ''syntax'' window and execute the following:

```
BEGIN PROGRAM python.
import spss
# Agreement test for SPSS
# D. Cousineau & L. Laurencelle, 2015. A ratio
# test of inter-rater agreement with high specificity

# get the observed frequencies o_{ij} from SPSS
dataCursor = spss.Cursor()
o = dataCursor.fetchall()
k = len(o[0])
dataCursor.close()

# get the marginal counts
totalrow = [sum(x) for x in o ]
totalcol = [sum(x) for x in zip(*o) ]
N = sum(totalrow)

# compute the expected frequencies e_{ij}
e = [[i*j/N for i in totalcol] for j in totalrow]

# compute the four terms of the test
z_ii_plus = sum(
  pow((o[i][i]-e[i][i]),2.0)/e[i][i]
```

```
  for i in range(k) if o[i][i] > e[i][i]
)
z_ii_minus= sum(
  pow((o[i][i]-e[i][i]),2.0)/e[i][i]
  for i in range(k) if o[i][i] < e[i][i]
)
z_ij_plus = sum(
  pow((o[i][j]-e[i][j]),2.0)/e[i][j]
  for i in range(k) for j in range(k) if (o[i][j] > e[i][j]) &
  (i!=j)
)
z_ij_minus= sum(
  pow((o[i][j]-e[i][j]),2.0)/e[i][j]
  for i in range(k) for j in range(k) if (o[i][j] < e[i][j]) &
  (i!=j)
)

# compute Q_A and P_A
Qa = (z_ii_plus + z_ij_minus) / ( z_ii_minus + z_ij_plus )
Pa = Qa / (1 + Qa)

# That's it! Just show the results.
spss.StartProcedure("Agreement test")
spss.TextBlock("Result","Q_A = "+str(Qa))
spss.TextBlock("Result","P_A = "+str(Pa))
spss.EndProcedure()

END PROGRAM.
```

To compute significance of $Q_A$ and confidence interval on $P_A$, you may then run the following lines (there must be at least one line of data in the data editor), adjusting the first four lines to your results:

```
COMPUTE k = 3. /* number of categories */
COMPUTE n = 100. /* number of observations */
COMPUTE alpha = 0.05. /* for 95% confidence interval */
COMPUTE Qa = 0.6235. /* obtained from above script */

COMPUTE pvalue = 1 - cdf.F(Qa, 0.5*(k-1)**2, (1-0.5)*(k-
1)**2).
FORMAT pvalue (f8.6).
COMPUTE Pa = Qa/(1 + Qa).
COMPUTE PaCIlo = idf.beta( alpha/2, Pa*(k-1)**2/2, (1-
Pa)*(k-1)**2/2).
```

```
COMPUTE PaCIhi = idf.beta(1-alpha/2, Pa*(k-1)**2/2, (1-
Pa)*(k-1)**2/2).

EXECUTE.
```

The results will appear in the data editor in new columns.

## Appendix C

### *Simulation Procedures*

The following describes the different parameters and procedures of our simulation experiments for assessing confidence intervals, the statistical power, and the specificity of the various test procedures. First, the true agreement rate $\rho_A$ between the two simulated raters was varied from no agreement between the raters ($\rho_A = 0$) to perfect agreement ($\rho_A = 1$) by increment of 1/200. Second, we varied the number of categories ($k$) from 5 to 25. Finally, the total number of observations was manipulated from small ($N = 50$) to large ($N = 500$).

The following algorithm was used to generate one judgment matrix:

Step 1 : Generate a random integer $x$ in $[1 \ldots k]$; this is the categorization of the first rater;

Step 2 : With probability $\rho_A$, make $y \leftarrow x$ (an agreement between the two raters), Otherwise, generate a random integer $y$ *in* $[1 \ldots k]$;

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

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

Note that for such a scheme of data generation, the expected *observed* rate of agreement is given by $\rho_A + (1 - \rho_A)/k$ (i.e., larger than $\rho_A$) because there may be an agreement by chance when selecting the categorization of Rater 2 randomly.

A Monte Carlo simulation went as follows. First, generate one square $k \times k$ matrix (using the algorithm above). Second, compute all test statistics, each compared with its appropriate critical value. This process is repeated 200,000 times for each $\rho_A \times k \times N$ combination. The significance level $\alpha$ (.05 or .01) is also varied. The proportion of rejections is then computed.

To examine specificity, the parameter $\rho_C$ was used, this time with a different meaning, that is, as a ''concordance rate'' describing the rate of concordant but not necessarily agreeing judgments. This rate might apply to cells outside the main diagonal.

Two types of configurations were explored, each with its own set of Monte Carlo simulations: *Random with possible coincident pairings* and *random excluding coincident pairings*. In both scenarios, for each category $x$ the first rater is choosing, the category $y(x)$ is chosen by the second rater when they are classifying the observations in a concordant fashion. The two scenarios differ in whether they allow $y(x)$ to be

equal to $x$, that is, to be true agreement. In the *Random with possible coincident pairings* scenario, occasional agreeing pairings are possible by chance. For the *Random excluding coincident pairings* condition, occasional agreeing pairs were excluded to insure that there was not a single agreeing pairing. The following algorithm was used:

Step 0 : Initialize a *set* of random pairings $\{x; y(x)\}$ of integers 1 to $k$ ; in the *Random excluding coincident pairings* variant, iterate untill there are no coincident pairs, *i.e.*, no $x \neq y(x)$

Step 1 : Generate a random integer $x$ in $[1 \ldots k]$; this is the categorization of the first rater;

Step 2 : With probability $\rho_C$, make $y \leftarrow y(x)$, Otherwise, generate a random integer $y$ *in* $[1 \ldots k]$;

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

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

As previously, the algorithm was iterated 200,000 times, keeping track of the rejection rate of the six tests, for each combination of $\rho_C, k$, and $N$ for each of the two scenarios described above. The significance level $\alpha$ (.05 or .01) was also varied.

In unreported simulations, we also manipulated the prevalence of the categories using a ''slope'' parameter $\Delta$. For $\Delta = 1$, all $k$ categories were equally frequent in the population (i.e., satisfying the *indifference principle*). For $\Delta = 3$, some categories are slightly more frequent than others (similar to the data of Table 4). For $\Delta = 10$, there is an important discrepancy between category frequencies. The data in Table 1 display extreme differences in category frequencies that would roughly correspond to $\Delta = 30$. Formally, $\Delta$ is the ratio of the largest to the smallest category prevalence probability, $\Delta = p_k / p_1$. The probabilities for the intermediate categories were varied linearly with the constraint that $\sum p_i = 1$.

By varying $\Delta$, only one new result was apparent: The $z_{\Sigma 1}$ test, based on the *indifference principle* (the hypothesis of an equal, random, distribution of observations across categories), fares correctly on power as long as the observations are indeed evenly distributed across the categories ($\Delta$ = 1, in which case it superimposes on Cohen's $z_{\kappa 1}$ results). However, its power curve rises well above the alpha-level and unto unjustified heights, in conditions with uneven prevalence of categories (15% of Type I error rate for $\Delta$ = 3 and up to 50% for $\Delta$ =10). As a consequence, the $z_{\Sigma 1}$ does not respect the prescribed alpha level when there is no agreement (i.e., for $\rho_A$ = 0) and displays spurious power when $\rho_A$ is greater than zero and the prevalence of the categories are not uniform.

## Acknowledgments

We would like to thank Bradley Harding for his comments on an earlier version of this article.

## 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. To be precise, half of the cells should have a count above the expected *median* but the median of a binomial distribution is complex to evaluate (Kaas & Buhrman, 1980) and close to the mean as soon as the expected value exceeds five.
2. Let the number of observed agreement be $x = N \cdot r$, and the estimated probability of concordance be $E(r)$. The distribution of $x$ is approximately binomial, that is, $B(N, E(r))$, and its normal approximation, $z = (N \cdot r - N \cdot E(r))/\sqrt{N \cdot E(r) \cdot (1 - E(r))}$, is algebraically equivalent to Equation A.3.

## References

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687-699. doi:10.1177/001316448104100307

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46. doi:10.1177/001316446002000104

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

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

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). *Statistical distributions*. New York, NY: Wiley.

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.

Kaas, R., & Buhrman, J. M. (1980). Mean, median and mode in binomial distributions. *Statistica Neerlandica*, *34*, 13-18. doi:10.1111/j.1467-9574.1980.tb00681.x

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, *21*, 2109-2129. doi:10.1002/sim.1180

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174. doi:10.2307/2529310

Stouffer, S. A., Suchman, E. A., De Vinney, L. C., Star, S. A., & Williams, R. M.Jr. (1949). *The American soldier: Vol. I. Adjustment during army life*. Princeton, NJ: Princeton University Press.

von Eye, A., Schauerhuber, M., & Mair, P. (2006). *Significance tests for the measure of raw agreement*. Retrieved from http://interstat.statjournals.net