Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2010, Article ID 480364, 10 pages
doi:10.1155/2010/480364

Research Article

Peirce's i and Cohen's κ for 2×2 Measures of Rater Reliability

Beau Abar and Eric Loken

Departement of Human Development and Family Studies, The Pennsylvania State University, University Park, PA 16802, USA

Correspondence should be addressed to Beau Abar, beau.abar@gmail.com

Received 11 November 2009; Revised 27 April 2010; Accepted 16 May 2010

Academic Editor: Junbin B. Gao

This study examined a historical mixture model approach to the evaluation of ratings made in "gold standard" and two-rater 2×2 contingency tables. Peirce's i and the derived i average were discussed in relation to a widely used index of reliability in the behavioral sciences, Cohen's κ. Sample size, population base rate of occurrence, the true "science of the method", and guessing rates were manipulated across simulations. In "gold standard" situations, Peirce's i tended to recover the true reliability of ratings as well as better than κ. In two-rater situations, i_{ave} tended to recover the true reliability as well as better than κ in most situations. The empirical utility and potential theoretical benefits of mixture model methods in estimating reliability are discussed, as are the associations between the i statistics and other modern mixture model approaches.

1. Introduction

In 1884, Peirce proposed an index of association, i, for a 2×2 contingency table. Peirce's index went beyond simple percent agreement, as a set of predictions can show substantial agreement with observed reality in situations when the predicted event rarely occurs. For example, by consistently predicting that a tornado will *not* occur [1, 2], a meteorologist could almost always be correct. Peirce derived a way to quantify what he called the "science of the method" [1] with his coefficient that predates even Pearson's correlation coefficient of association [2]. To understand Peirce's coefficient, suppose that a 2×2 contingency table with predictions and outcomes is constructed as follows:

$$\begin{array}{cc|c|c} & & \multicolumn{2}{c}{\text{Observed}} \\ & & \multicolumn{2}{c}{\text{Outcome}} \\ & & \text{Yes} & \text{No} \\ \text{Rater} & \text{Yes} & a & b \\ \cline{3-4} \text{Prediction} & \text{No} & c & d \end{array} \tag{1.1}$$

Peirce then defined i as

$$i = \frac{ad - bc}{(a+c)(b+d)}. \tag{1.2}$$

In this derivation, agreement between the prediction and the outcome is considered a combination of events due to the science of the method employed to arrive at the prediction and chance agreement. In Peirce's mixture, the science of the method refers to the predictions of a hypothetical "infallible observer," who correctly makes predictions (i.e., according to science). The chance component is produced by a hypothetical completely "ignorant observer" [1] whose random predictions are sometimes correct and sometimes incorrect.

Although Peirce's i has rarely been mentioned in the behavioral sciences [2, 3], it was "rediscovered and renamed" in the meteorological literature three times in the 20th century as the Hanssen-Kuipers discriminant, the Kuipers performance index, and the true skill statistic [4]. It currently serves as a popular measure of the precision and utility of a weather forecasting system [5]. The model was also proposed by Martín Andrés and Luna del Castillo [6] in the context of multiple-choice tests.

Peirce's innovation anticipated a recent trend of using mixture models to estimate reliability [7–10]. Although Peirce was concerned mostly with prediction, his insights are relevant to other cross-classified tables, including agreements between pairs of raters. Both Schuster and Smith [10] andAickin [7] derived models for viewing rating data as a mixture of agreements for cause and agreements due to chance guessing. Their formulations are generalizable to situations with more than two raters and more than two categories for judgment, but are nonidentified for 2 raters with 2 categories [7, 10, 11]although see Martín Andrés and Femia-Marzo [12] for an approach that is identified). Peirce's coefficient i is identified for 2 × 2 tables only because one of the margins (the observed outcomes) is assumed to be fixed and to represent the true underlying base rate of the process. Given its relative ease of calculation, theoretical foundation in a mixture model framework, and popularity in other domains of research, demonstrating the utility of Peirce's i in these single-rater situations could position it as a viable alternative to more commonly used indices in the behavioral and medical sciences.

Perhaps the most popular coefficient of agreement for 2 by 2 tables is Cohen's κ [13] (an online literature search using PsychINFO found that 1458 peer-reviewed journal articles published since 2000 in the behavioral sciences discuss κ in the context of reliability). Cohen's κ is defined as

$$\kappa = \frac{P_o - P_c}{1 - P_o}. \tag{1.3}$$

Rather than modeling underlying sources of agreement, κ instead uses the observed margins to correct the total observed agreement (P_o) for the expected agreement due to chance (P_c). In 2 × 2 contingency tables, using the terminology from (1.2), P_o and P_c represent.

$$P_o = \frac{a+d}{a+b+c+d}, \quad P_c = \frac{(a+b)(a+c) + (c+d)(b+d)}{a+b+c+d}. \tag{1.4}$$

Kappa therefore assesses the degree to which rater agreement exceeds that expected by chance, which is determined by the marginal values of the table. This represents an important conceptual difference with the mixture approaches in that κ is not really explicit about what is meant by chance agreement [7, 14, 15], whereas Peirce's definition of i delineates how both agreement and disagreement can occur.

Because the mixture derivation is specific about the data generating mechanism, it is easy to simulate contingency tables, and in this paper we will compare the performance of Peirce's i and Cohen's κ when the data are generated according to Peirce's model. Some of the results can be anticipated by an analytic comparison of the two coefficients. Loken and Rovine [3] show that, in terms of the contingency table defined above, κ can be defined as

$$\kappa = \frac{2(ad - bc)}{(a+c)(c+d) + (b+d)(a+b)}. \tag{1.5}$$

Clearly, κ and i differ in that κ is identical even if the rows and columns are interchanged, and the typical reliability assessment does not depend on rater assignment. Peirce's i, however, is not symmetric, because the columns represent the observed outcomes (or in a rating setting, this could also be taken to be the "gold-standard" rating [16]—such as a blood test being compared to a preliminary diagnosis). Thus, κ can be used to assess the reliability of two raters or of one rater compared to a gold-standard one.

In a single rater setting, i and κ can be compared analytically [3]. The expected values of i and κ are equal when (a) the observed proportions of yeses and noes are equal, (b) the population base rate of occurrence is equal to 1/2, and/or (c) the guessing parameter, j, is equal to the base rate [3]. Under Peirce's formulation, this guessing parameter can be thought of as the proportion of cases, where the completely "ignorant observer" chooses yes. When these conditions are not met, estimates of i and κ differ. If j is "more extreme" than the given base rate, the estimate of κ will be greater than i. If j is "less extreme" than the given base rate, the estimate of i will be greater than κ. In situations where the base rate is less than 1/2, j is said to be "more extreme" if it is closer to zero than the base rate, and for situations where the base rate is greater than 1/2, j is said to be "more extreme" if it is closer to 1.0 than the base rate [3].

The current study will first expand upon these findings by comparing i and κ in a "gold-standard" situation across differing sample sizes, base rates, guessing rates, and true "method of the science." A series of 2 × 2 tables are analyzed using both i and κ. The second part of the current study will describe a method for expanding the utility of Peirce's i to a two rater, 2 × 2 setting. The previous research examining interrater reliability in contingency tables has encountered identifiability problems in 2 × 2 settings [9–12]. Other researchers have dealt with the problem of nonidentifiability by adding a nominal third response category and adding .5 to every cell in the contingency table [11]. These researchers later defined an efficient and easily calculable approach by analyzing 2 × 2 tables that does not require adding an additional category [12]. We will describe an alternative method for making i useful in a two-rater setting. Our approach will be to calculate a value for i under both possible orientations of the contingency table, and then calculating i_{ave} as the mean of the two estimates. A second set of simulations compares the recovery of true agreement by the new index and κ.

2. Simulating i and κ in "Gold-Standard" Situations

We simulated 2 × 2 contingency tables using the following definitions: (1) the sample size is N; (2) the population base rate of "yeses," or true occurrence, is τ; (3) the true "science of

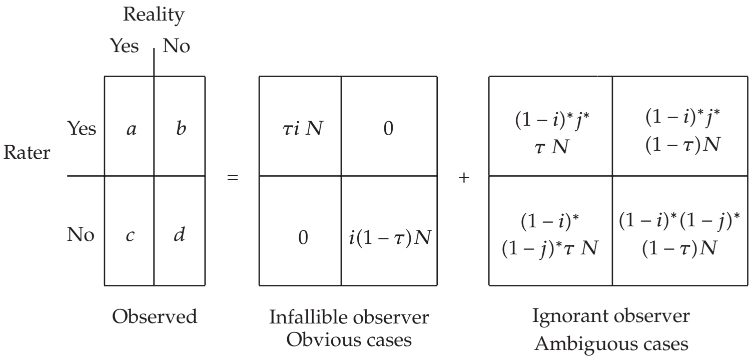

Figure 1: Gold standard data generating mechanism.

the method" (or cases classified "for cause") is i; (4) the guessing rate of the single observer for the remainder of the cases is j. This notation for Peirce's i is consistent with that used by Rovine and Anderson [2] and Loken and Rovine [3].

A 2 by 2 table was generated by drawing N binary events with probability τ. Of the true yeses, proportion i were classified as yes/yes. The remaining proportion of $(1 - i)$, the "true yeses," was classified by the "ignorant" observer as yes with probability j. The same procedure was used for the "truenoes." Therefore, Peirce's i makes the assumption that the ability of the rater to correctly identify "true yeses" is equal to his/her ability to identify "true noes." In reality, this assumption may not hold across all situations (e.g., potentially easier to identify days in which a tornado is unlikely to occur than days when a tornado is more likely to occur). Figure 1, adapted from Loken and Rovine [3], presents a graphical representation of how the underlying model of Peirce's i was used to generate the data. The simulated tables were then used to calculate i and κ.

Simulations were performed using small (25), moderate (100), and large (500) sample sizes, and several combinations of values for τ, i, and j. Since the parameters are symmetrical about .5, there was no additional need to examine values below .5. For a given set of fixed parameters, we generated 1000 tables and calculated the means and standard deviations of Peirce's i and Cohen's κ.

3. Peirce's i and Cohen's κ in "Gold-Standard" Situations

An illustrative subset of the simulations is summarized in Table 1. The cases included in the table are representative of the global trends observed across simulations. In most cases, the mean estimates of i and κ are essentially identical. In general, as sample size increased, estimates of both Peirce's i and κ became closer to the data-generating "science of the method" and the standard deviations decreased in the expected manner. There were, however, situations where substantial mean differences were observed. When the guessing parameter (j) is less extreme than the population base rate (τ), the mean of Peirce's i is closer to the true "science of the method" than κ. This difference increases as the discrepancy between τ and j increases. For example, for $\mathrm{N} = 500$, $\tau = .7$, and $j = .5$, the mean difference is .04; with $\tau = .9$ and $j = .5$, the mean difference = .23. The more the guessing rate underestimates the base rate, the more κ is downwardly biased relative to i. When j is "more extreme" than τ, there also appears to be a difference in estimates provided by Peirce's

Table 1: Descriptive statistics of i and κ.

	Peirce's i		Cohen's κ	
Parameters (n, i, j)	Mean	SD	Mean	SD
$\tau = .5$				
500, .5, .5	.50	.03	.50	.03
100, .5, .5	.50	.07	.50	.07
25, .5, .5	.50	.14	.49	.14
500, .7, .5	.70	.03	.70	.03
100, .7, .5	.69	.06	.69	.06
25, .7, .5	.70	.11	.69	.11
$\tau = .7$				
500, .5, .5	.50	.04	.46	.03
100, .5, .5	.50	.08	.45	.08
25, .5, .5	.50	.18	.45	.17
500, .7, .5	.70	.03	.66	.03
100, .7, .5	.70	.06	.66	.06
25, .7, .5	.71	.13	.65	.14
500, .5, .9	.50	.02	.55	.02
100, .5, .9	.50	.05	.55	.05
25, .5, .9	50	.11	.55	.11
$\tau = .9$				
500, .5, .5	.50	.05	.27	.04
100, .5, .5	.50	.12	.26	.08
25, .5, .5	.48	.31	.27	.17
500, .7, .5	.70	.04	.46	.04
100, .7, .5	.70	.10	.45	.10
25, .7, .5	.70	.21	.45	.18

i and Cohen's κ. For example, when $\tau = .7$ and $j = .9$, the mean difference is .05, with κ overestimating the "science of the method."

4. Discussion of the Utility of Peirce's i in Gold-Standard Situations

The results illustrate the utility of Peirce's i in a "gold-standard" situation, where one rater works against a known outcome, or a definitive standard. In general, when data are generated under the model presented by Peirce, i and κ tend to provide similar estimates of rater accuracy, with similar variability. This similarity, however, does not hold when the guessing rate for the random ratings does not match the population base rate. Because the population base rate and the guessing rate are confounded in Cohen's κ, estimates of reliability will be affected by mismatches. The discrepancy between i and κ can be substantial, with the potential for qualitatively different interpretations of the extent of reliability observed. However, the most serious bias in κ appears to occur only for mismatches that would seem less likely to occur in real data (e.g., a random guessing rate of .9 even though the population base rate is .5).

While these findings provide support for the use of i as a viable alternative to κ in a "gold-standard," one-rater setting, the question of its utility in a "nongold standard," or two-rater, setting remains [10]. In the behavioral sciences, it is common to have two equally qualified raters of the same event (e.g., two teachers evaluating the same student, two coders rating a videotape, two doctors classifying an MRI, etc.). The issue of reliability then centers on their combined agreements and disagreements, without reference to an absolute criterion. As stated before, κ is symmetrical but Peirce's i is not, as it explicitly treats one of the margins as the fixed standard. The purpose of the second part of the current study is to extend the utility of Peirce's i to two-rater, 2 × 2 contingency tables.

5. Simulating i and κ in Two-Rater Situations

As mentioned above, Peirce's i is unidentified in a two-rater setting. When τ is not given, there are too many parameters to estimate relative to the degrees of freedom. Our approach is to estimate the tables under two different assumptions, and then take the average measure. We first altered the original formula for i by rearranging the table margins (i.e., rows are treated as columns and vice versa). The resulting formula, i^*, is illustrated in

$$i^* = \frac{ad - bc}{(a+b)(c+d)}. \tag{5.1}$$

This formula reverses the assumptions about what is the fixed margin. Simulations run using the definitions discussed in part 1 of this study show that estimates of i and i^* bracket κ, such that if i is greater than κ, i^* is smaller and vice versa (but κ is often not found precisely in the middle of the bracket).

Peirce's i and i^* are then averaged to estimate the reliability for 2 rater, 2 × 2 contingency tables. The formula for i_{ave} is

$$i_{\text{average}} = \frac{1}{2}\left[\frac{ad - bc}{(a+c)(b+d)} + \frac{ad - bc}{(a+b)(c+d)}\right]. \tag{5.2}$$

Additional simulations were performed examining the association between i_{ave} and κ.

We began with the same set of definitions as simulation 1, except that two fixed guessing rates are required, j and f,where f represents the guessing rate of the second rater. A 2 × 2 table was again generated by drawing N binary events with probability τ. Of the true yeses, proportion i was classified as yes/yes. The same proportion of no cases was classified as no/no. The remaining cases were randomly dispersed across the 4 cells using the joint probabilities of j and f. For example, the probability of an ambiguous case being classified yes/no is equal to $j\,(1-f)$. Figure 2 presents a graphical representation of how the underlying model of Peirce's i was used to generate tables in a two-rater situation. The resulting observed tables were then used to calculate i_{ave} and κ.

Simulations were performed using small (25), moderate (100), and large (500) sample sizes, and several combinations of values for τ, i, j, and f. In order to explore the effect of differing guessing rates, both j and f were examined from .1 through .9. Similar to the first simulation, τ and i were only examined between .5 and .9. For a given set of fixed parameters, we generated 1000 tables and calculated the means, standard deviations, and minimum and maximum differences between i_{ave} and κ.

$$\begin{array}{cc|c|c|}
 & & \multicolumn{2}{c}{\text{Rater 2}} \\
 & & \text{Yes} & \text{No} \\
\hline
\text{Rater 1} & \text{Yes} & a & b \\
\hline
 & \text{No} & c & d \\
\hline
 & & \multicolumn{2}{c}{\text{Observed}}
\end{array}
=
\begin{array}{|c|c|}
\hline
\tau i\, N & 0 \\
\hline
0 & i(1-\tau)N \\
\hline
\multicolumn{2}{c}{\text{Obvious cases}}
\end{array}
+
\begin{array}{|c|c|}
\hline
(1-i)^{*} j^{*} f^{*} N & (1-i)^{*}(1-j)^{*} f^{*} N \\
\hline
(1-i)^{*} j^{*} (1-f)^{*} N & (1-i)^{*}(1-j)^{*} (1-f)^{*} N \\
\hline
\multicolumn{2}{c}{\text{Ambiguous cases}}
\end{array}$$

Figure 2: Two-rater data generating mechanism.

Table 2: Descriptive statistics of i_{ave} and κ.

	i_{ave}		Cohen's κ		Difference	
Parameters (n, i, j, f)	Mean	SD	Mean	SD	Min	Max
$\tau = .5$						
500, .5, .5, .5	.50	.03	.50	.03	.00	.01
100, .5, .5, .5	.50	.07	.49	.07	.00	.04
25, .5, .5, .5	.51	.15	.50	.15	.00	.08
500, .5, .9, .5	.55	.03	.50	.03	.02	.07
100, .5, .9, .5	.55	.06	.50	.06	.00	.10
25, .5, .9, .5	.55	.11	.50	.12	.00	.16
500, .5, .3, .7	.48	.03	.44	.03	.01	.06
100, .5, .3, .7	.48	.06	.44	.06	.00	.10
.25, .5, .3, .7	.48	.13	.43	.13	.00	.17
500, .5, .1, .9	.40	.02	.29	.02	.09	.13
100, .5, .1, .9	.40	.03	.29	.04	.06	.14
25, .5, .1, .9	.41	.07	.29	.08	.03	.17
500, .9, .5, .5	.90	.01	.90	.01	.00	.00
100, .9, .5, .5	.90	.03	.90	.03	.00	.01
25, .9, .5, .5	.91	.06	.90	.07	.00	.04
500, .9, .3, .7	.89	.01	.88	.01	.00	.01
100, .9, .3, .7	.89	.03	.88	.03	.00	.01
$\tau = .9$						
500, .5, .5, .5	.41	.04	.40	.04	.00	.01
100, .5, .5, .5	.40	.09	.40	.09	.00	.03
25, .5, .5, .5	.40	.19	.39	.19	−.07	.17
500, .5, .3, .7	.39	.04	.34	.04	.02	.09
100, .5, .3, .7	.39	.08	.34	.08	.01	.16

6. Peirce's i and Cohen's κ in Two-Rater Situations

An illustrative subset of the simulations is summarized in Table 2. As before, the cases included in the table are representative of the global trends observed across simulations. In general, i_{ave} and κ provide similar estimates of reliability. Specifically, in situations where j and f are equal, the mean and variance of i_{ave} are nearly identical to those of κ. When j and f differ and are each greater than or equal to the population base rate (τ), i_{ave} is slightly upwardly biased, while κ is not. When guessing rates differ and bracket the base rate (e.g., $\tau =$

$.5, j = .3, f = .7$), both i_{ave} and κ are downwardly biased, with κ being slightly more affected (in the example above, M diff $\approx .04$). When j and f differ and each are less than or equal to τ, both estimates are also downwardly biased, with κ again being mildly more affected.

7. Discussion of the Utility of i_{ave} in Two-Rater Situations

When data were simulated using the mixture framework of obvious and ambiguous cases, i_{ave} tends to be as stable an estimate as κ, providing a very similar estimate of the true interrater reliability seen in 2×2 tables. However, similar to the results of the first simulation, there are situations where substantial discrepancy between i_{ave} and κ was observed. Specifically, when raters employ drastically different guessing rates (e.g., .1 and .9), κ shows a more severe downward bias from the true "science of the method" than does i_{ave}. However, it continues to be the case that the most serious bias in κ occurs for mismatches less likely to occur in real data. For example, in a two rater setting, it is not likely that one rater would overwhelmingly choose "yes" for ambiguous cases while the other overwhelmingly chooses "no."

8. Empirical Conclusions and Associations with Other Mixture Models

The current study compares Peirce's i [1] to Cohen's κ [13] for examining interrater reliability in 2×2 contingency tables. In the "gold-standard," one-rater setting, Peirce's i [1] and the commonly used κ performed similarly. Under certain conditions described above, however, i tended to do a better job of recovering the true "science of the method" than κ. We point out that Peirce's i was designed to examine single-rater predictions while κ was designed to index interrater reliability, and that the data were generated under the assumption that i represented the true model. In the two-rater, interrater reliability setting, i_{ave} appeared to perform as well κ across multiple scenarios. In addition to the empirical equivalences and benefits observed, the theory associated with i and i_{ave} more clearly articulates what is meant by "agreement due to chance" than does Cohen's κ [3]. This formal definition of the data generating process allows researchers to simulate 2×2 contingency tables based on the guidelines of a modern, mixture model framework [3, 7–10].

One known problem with the use of the i statistics and/or κ (discussed in regard to κ by Martín Andrés and Femia Marzo [11] andNelson and Pepe, 2000 [17]), was also encountered in the current study, was the negative estimates of reliability provided by each index when either of the agreement cells (a or d) were empty. An alternative index of reliability, Δ, proposed by Martín Andrés and Femia-Marzo [12] has been shown to provide accurate estimates of reliability in these situations.

In a broader context, there has been considerable interest recently in applying mixture models to issues in measurement and reliability [7, 9, 10, 18–21]. For example, one approach to evaluating model fit is to view the data as a mixture of cases that do and do not conform to the model [19, 20]. The estimate of the proportion of a sample that must be removed in order for the data to perfectly fit a hypothesized model H (a quantity called π^*) has a strong intuitive appeal and also has the advantage over a traditional χ^2 of being insensitive to sample size [20].

Other mixture model approaches to contingency tables have explicitly examined rater agreement and reliability. As mentioned above, both Aickin [7] and Schuster and Smith [9, 10] have described the population of rated cases as a 2-class mixture of some classified "for cause" and others classified by chance, and these models fall under the broader mixture category of latent agreement models of reliability [8, 18, 21]. Peirce's approach of viewing agreement as stemming from an infallible observer and a completely ignorant observer is directly analogous to the approaches discussed above, where cases/items are viewed as obvious or ambiguous [3, 7–9]. We believe that Peirce's i and our adjusted indicator for 2 by 2 tables generated by equivalent raters offer an intuitive, appealing, and accessible way to evaluate rater reliability.

Acknowledgment

This research was partially supported by Grant DA 017629 from the National Institute on Drug Abuse.

References

[1] C. S. Peirce, "The numerical measure of the success of predictions," *Science*, vol. 4, no. 93, pp. 453–454, 1884.

[2] M. J. Rovine and D. R. Anderson, "Peirce and Bowditch: an American contribution to correlation and regression," *The American Statistician*, vol. 58, no. 3, pp. 232–236, 2004.

[3] E. Loken and M. J. Rovine, "Peirce's 19th century mixture model approach to rater agreement," *The American Statistician*, vol. 60, no. 2, pp. 158–161, 2006.

[4] D. B. Stephenson, "Use of the "odds ratio" for diagnosing forecast skill," *Weather and Forecasting*, vol. 15, no. 2, pp. 221–232, 2000.

[5] F. W. Wilson, "Measuring the decision support value of probabilistic forecasts," in *Proceedings of the 12th Conference on Aviation Range and Aerospace Meteorology and the 18th Conference on Probability and Statistics in the Atmospheric Sciences*, Atlanta, Ga, USA, 2006.

[6] A. Martín Andrés and J. D. Luna del Castillo, "Tests and intervals in multiple choice tests: a modification of the simplest classical model," *British Journal of Mathematical and Statistical Psychology*, vol. 42, pp. 251–263, 1989.

[7] M. Aickin, "Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa," *Biometrics*, vol. 46, no. 2, pp. 293–302, 1990.

[8] I. Guggenmoos-Holzmann and R. Vonk, "Kappa-like indices of observer agreement viewed from a latent class perspective," *Statistics in Medicine*, vol. 17, no. 8, pp. 797–812, 1998.

[9] C. Schuster and D. A. Smith, "Indexing systematic rater agreement with a latent-class model," *Psychological Methods*, vol. 7, no. 3, pp. 384–395, 2002.

[10] C. Schuster and D. A. Smith, "Estimating with a latent class model the reliability of nominal judgments upon which two raters agree," *Educational and Psychological Measurement*, vol. 66, no. 5, pp. 739–747, 2006.

[11] A. Martín Andrés and P. Femia Marzo, "Delta: a new measure of agreement between two raters," *The British Journal of Mathematical and Statistical Psychology*, vol. 57, no. 1, pp. 1–19, 2004.

[12] A. Martín Andrés and P. Femia-Marzo, "Chance-corrected measures of reliability and validity in 2 × 2 tables," *Communications in Statistics*, vol. 37, no. 3–5, pp. 760–772, 2008.

[13] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[14] R. L. Brennan and D. J. Prediger, "Coefficient Kappa: some uses, misuses and alternatives," *Educational and Psychological Measurement*, vol. 41, pp. 687–699, 1981.

[15] R. Zwick, "Another look at interrater agreement," *Psychological Bulletin*, vol. 103, no. 3, pp. 374–378, 1988.

[16] S. C. Weller and N. C. Mann, "Assessing rater performance without a "gold standard" using consensus theory," *Medical Decision Making*, vol. 17, no. 1, pp. 71–79, 1997.

[17] J. C. Nelson and M. S. Pepe, "Statistical description of interrater variability in ordinal ratings," *Statistical Methods in Medical Research*, vol. 9, no. 5, pp. 475–496, 2000.

[18] A. Agresti and J. B. Lang, “Quasi-symmetric latent class models, with application to rater agreement,” *Biometrics*, vol. 49, no. 1, pp. 131–139, 1993.

[19] C. M. Dayton, “Applications and computational strategies for the two-point mixture index of fit,” *The British Journal of Mathematical and Statistical Psychology*, vol. 56, no. 1, pp. 1–13, 2003.

[20] T. Rudas, C. C. Clogg, and B. G. Lindsay, “A new index of fit based on mixture methods for the analysis of contingency tables,” *Journal of the Royal Statistical Society. Series B*, vol. 56, no. 4, pp. 623–639, 1994.

[21] J. S. Uebersax and W. M. Grove, “A latent trait finite mixture model for the analysis of rating agreement,” *Biometrics*, vol. 49, no. 3, pp. 823–835, 1993.