Interrater Reliability Estimators Tested against True Interrater Reliabilities

Xinshu Zhao*[a] (xszhao@um.edu.mo)

Guangchao Charles Feng [a] (charlesfeng@um.edu.mo)

Song Harris Ao [a] (harrisao@um.edu.mo)

Piper Liping Liu [a] (lipingliu@um.edu.mo)

[a]: Department of Communication, Faculty of Social Sciences, University of Macau, Taipa, Macao

Corresponding Author: Xinshu Zhao

Abstract

Background: Interrater reliability, aka intercoder reliability, is defined as true agreement between raters, aka coders, without chance agreement. It is used across many disciplines including medical and health research to measure the quality of ratings, coding, diagnoses, or other observations and judgements. While numerous indices of interrater reliability are available, experts disagree on which ones are legitimate or more appropriate.

Almost all agree that percent agreement (a_o), the oldest and the simplest index, is also the most flawed because it fails to estimate and remove chance agreement, which is produced by raters' random rating. The experts, however, disagree on which of the chance-adjusted indices are legitimate or better. The experts also disagree on which of the three factors, rating category, distribution skew, or task difficulty, a good index should rely on to estimate chance agreement, or which of the factors the known indices in fact rely on.

The most popular chance-adjusted indices, according to a functionalist view of mathematical statistics, assume that all raters conduct intentional and maximum random rating while typical raters conduct involuntary and reluctant random rating. The mismatch between the assumed and the actual rater behaviors causes the indices to rely on mistaken factors to estimate chance agreement, leading to the numerous paradoxes, abnormalities, and other misbehaviors of the indices identified by prior studies.

Methods: We conducted a 4×8×3 between-subject controlled experiment with 4 subjects per cell. Each subject was a rating session with 100 pairs of rating by two raters, totaling 384 rating sessions as the experimental subjects. The experiment tested seven best-known indices of interrater reliability against the observed reliability and chance agreement. Impacts of the three factors, i.e., rating category, distribution skew, and task difficulty, on the indices were tested.

Results: The most criticized index, percent agreement (a_o), showed as the most accurate predictor of reliability, reporting directional r^2=.84. It was also the third best approximator, overestimating observed reliability by 13 percentage points. The three most acclaimed and most popular indices, Scott's π, Cohen's κ and Krippendorff's α, underperformed all other indices, reporting directional r^2=.312 and underestimated reliability by 31.4~31.8 points. The newest index, Gwet's AC_1, emerged as the second-best **predictor** and the most accurate approximator. Bennett et al's S ranked behind AC_1, and Perreault and Leigh's I_r ranked the fourth both for prediction and approximation. The reliance on category and skew and failure to rely on difficulty explain why the six chance-adjusted indices often underperformed a_o, which they were created to outperform. The evidence corroborated the notion that the chance-adjusted indices assume intentional and maximum random rating by the raters while the raters instead exhibited involuntary and unwilling random rating.

Conclusion: The authors call for more empirical studies and especially more controlled experiments to falsify or qualify this study. If the main findings are replicated and the underlying theories supported, new thinking and new indices may be needed. Index designers may need to refrain from assuming intentional and maximum random rating, and instead assume involuntary and reluctant random rating. Accordingly, the new indices may need to rely on task difficulty, rather than distribution skew or rating category, to estimate chance agreement.

Key words*:* intercoder reliability, interrater reliability, reconstructed experiment, Cohen's *kappa*, Krippendorff's *alpha*,.

Interrater Reliability Estimators Tested against True Interrater Reliabilities

Background

Intercoder or interrater reliability is used to measure measurement quality in many disciplines, including health and medical research (1–10). A search of databases including Google Scholar, Scopus, and Web of Science found dozens of terms in academic literature, such as *diagnostician* for inter-diagnostician reliability and *patient* for inter-patient reliability, showing the concept's broad reach --

> annotator, arbitrator, assessor, auditor, diagnostician, doctor, editor, evaluator, examiner, grader, interpreter, interviewer, judge, monitor, observer, operator, patient, pharmacist, physician, reader, referee, reporter, researcher, respondent, scorer, screener, student, supervisor, surgeon, teacher, tester, therapist, transcriber, translator, user, voter.

Likely the earliest index is *percent agreement*, denoted a_o (9,11). Almost all reliability experts agree that a_o inflates reliability because it fails to remove *chance agreement* (a_c) (2–5,12–14). Scores of indices have been proposed to estimate and remove a_c. Bennett and colleagues' S and Perreault and Leigh's I_r estimate a_c as functions of *category* (C) (7,15). Scott's π, Cohen's κ and Krippendorff's α estimate a_c as functions of distribution *skew* (s_k) (2,16–19). Gwet's AC_1 makes a_c a function of both category and skew. Although many other indices are available and new indices continue to emerge, only these seven are in regular use and continue to be recommended or advocated, according to comprehensive reviews (14,20–26).

Using derivation or simulation, statisticians discuss and debate three questions: 1) Which indices are valid or more accurate when estimating reliability or chance agreement? 2) What factors affect the indices? 3) What factors should affect the indices? Answers to Questions 2 and 3 explain the answers to Question 1 (14,27). Underlying the debates are five viewpoints, the first of which is widely shared by almost all experts, while the others are contested, often heatedly. The five viewpoints lead to five groups of conjectures, which we list below and leave the details to Appendix, Section I.2.

1. Percent agreement (a_o) ignores chance agreement (a_c), therefore is inflated.
2. Rating category (C) inflates S, I_r, and AC_1 by deflating the indices' a_c estimates.
3. Distribution skew (s_k) deflates π, κ & α by inflating the indices' a_c estimates.
4. Major indices overlook task difficulty, a major factor affecting a_c; consequently, they misestimate reliability.
5. Chance-adjusted indices, S, π, κ, α, I_r, and AC_1 included, assume intentional and maximum chance rating by all raters; it is under this assumption that the chance-adjusted indices share the same chance correcting formula, Equation 1, where a_o is observed %-agreement, a_c is estimated chance agreement, and r_i is estimated true agreement, i.e., reliability index.

$$r_i = \frac{a_o - a_c}{1 - a_c} \quad (1)$$

The intentional-random assumption, aka maximum-random assumption, is said to be a root cause of many known paradoxes, abnormalities, and other misbehaviors of the indices, because raters are believed to be honest and truthful. Random ratings, if any, should be involuntary rather than intentional, task-dependent rather than invariably maximized (14,21–24,26,28–30).

Chance agreement is a product of rater behavior, and the debates are ultimately about rater behavior (14,31): What behaviors are assumed by the indices' estimations? What behaviors in fact take place? Do the assumptions match the behaviors? The debaters rely on theoretical arguments, mathematical derivation, fictitious examples, naturalistic comparisons, and Monte Carlo simulation. A systematic observation of rater behavior is needed to inform the debates over rater behavior.

This paper reports a controlled experiment that manipulated category, skew, and difficulty, and observed raters' behavioral responses. The seven indices were tested against the observed behavior. The findings also apply to the two equivalents of a_o, six equivalents of S, two equivalents of π, and one equivalent of κ, covering 18 indices in total, all of which had been analyzed mathematically by Zhao, Liu and Deng (14).

Methods

Reconstructed Experiment with Golden Standard

Reconstructed Experiment on Real Data (REORD)

We conducted a 4×8×3 between-subject controlled experiment with 4 subjects per cell. Here the term "subject" refers to the unit of analysis of a study, such as a participating patient in an experiment on the effectiveness of a new drug. A "subject" in this study, however, was a rating session with 100 pairs of rating by two raters. As 4×8×3×4=384, this study was based on 384 rating sessions, aka subjects. The three manipulated factors included four levels of *category* (C=2,4,6,8), eight levels of *difficulty* (d_f ranges 0~1, 0 for the least and 1 for the most difficult), and three levels of *skew* (s_k=0.5 for 50-50 distribution, 0.75 for 75-25 or 25-75 distribution, and 0.99 for 99-1 or 1-99 distribution), as summarized in Table 1.

[Insert Table 1 about here]

Over three hundred raters, registering 383 web names, from 53 Asian, European, and North American cities judged online the lengths of bars, which served as the experimental stimulus. A total of 22,290 items were rated, of which 19,900 were successfully paired, producing 9,950 pairs of rating. Borrowing techniques from bootstrap (32,33), jackknife (34), and Monte Carlo simulation (35), we sampled and resampled from the 9,950 pairs to reconstruct the 384 rating sessions (36).

Thus, raters and rating were real, while rating sessions were reconstructed, making it a *reconstructed experiment on real data* (REORD). The Appendix at the end of this manuscript

(Section II) provides further details and rationales of this REORD experiment.

Observed True Reliability (o_{ri}) and True Chance Agreement (o_{ac}) as Golden Standards

The raters were instructed to judge the length of bars. The researchers determined the bar lengths through programming, therefore know with certainty which rating decision was right or wrong. As the lengths of the bars were set such that random guesses would occur only between the longest and the second longest bars, the true chance agreement (o_{ac}) was twice the wrong agreement (Eq. 3, Appx.), and true reliability (o_{ri}) was observed agreement a_o minus o_{ac} (Eq. 5 of Appx.). Thus, o_{ri} served as the golden standard, namely the observed estimand, against which the seven indices were evaluated, and o_{ac} served as the golden standard for the seven chance estimators (37).

Five Independent Variables and Sixteen Dependent Variables

Thus, this REORD experiment features three manipulated independent variables, *category* I, skew (s_k) and difficulty (d_f) and 16 main dependent variables, which are the seven indices' reliability and chance estimations plus the observed true reliability (o_{ri}) and true chance agreement (o_{ca}). As the two main estimands, o_{ri} and o_{ca} sometimes also serve as independent variables when assessing their impacts on the indices' estimations. Table 1, Table 2 and the Appendix provide more details and rationales of variable calculations.

[Insert Table 2 about here]

Statistical Indicators – Directional R Squared (dr^2) and Mean of Errors (m_e)

Reliability indices serve two functions. One is to evaluate measurement instruments against each other, for which an index needs to accurately predict, meaning positively and highly correlating with, true reliability. We use *directional r squared* ($dr^2=r\cdot|r|$) to gauge the predictive accuracy of the seven indices and their chance estimators (Table 2 and Eq. 10 of the Appendix). We preferred r^2 over r because r^2 has a clearer and more practical interpretation, percent of the DV variance explained by the IV; r^2 is also more conservative as $r^2\leq|r|$. We preferred dr^2 over r^2 because dr^2 indicates the direction of the relationship while r^2 does not.

The second function of the indices is to evaluate measurement instruments against fixed benchmarks, such as 0.67 and 0.80, that some reliability authorities recommend (19,30,38,39). For this function, an index needs to approximate true reliability. We use *mean of errors*, m_e, which is the indices' deviations from the observed true reliability averaged across the 384 rating sessions, to gauge the approximating accuracy of the seven indices, denoted $m_e(r_i)$ in Table **2** and Eq. 8 of the Appendix. With the same reasoning, we also use m_e to assess and compare the chance estimators of the indices, denoted $m_e(a_c)$ in Table **2** and Eq. 9 of the Appendix.

We adopted $dr^2>.8$ as the primary benchmark and $m_e<.02$ as the secondary benchmark when evaluating the seven indices. Section V of the Appendix details the calculations of and the rationales behind the benchmarks.

Functions of P Values and Statistical Pretests

This study observes the tradition of reporting $p<\alpha$, where α=.05, .01, or .001. We however also strive to follow what have been advocated as a better statistical practice (40–44):

1) avoiding the terms containing “significance, e.g., “statistical significance,” for $p<\alpha$;
2) considering $p<\alpha$ as a prescreen threshold, passing which allows us to assess, interpret, and compare effect size indicators, such as r^2, dr^2 and m_e, with some confidence;
3) using terms such as “statistical pretest” and “statistically acknowledged” where we would have traditionally used “significance test” and “statistically significant;”
4) reserving the terms containing “significant” and “significance” exclusively for effect sizes of practical or theoretical importance.

More of our views and practices regarding the functions of *p* values may be found in our prior work (45–47).

Results

Reliability Estimations Tested Against Observed Reliability

Findings are summarized in Tables 3 through 6 and Figure 1 and discussed in three sections. This section (II) reports the performance of the seven indices when predicting and approximating the observed reliability. The next section (III) analyzes the impact of four factors on the indices' performance. The following section (IV) discussed *offset* mechanism, which is a key to understand the indices' complex behavior.

Overall, 2.86% of the raters' decisions fell on the short bars (1.11%, 1.93% and 5.53% respectively for four, six, and eight categories). As expected, there were fewer agreements on short bars, averaging 0.45% (0.04%, 0.12%, and 1.18%). These agreements showed no detectable effects on the main relations we investigate. The correlations between the manipulated variables were practically zero, confirming orthogonality, which rules out confounding or multicollinearity.

Predicting Reliability

Percent agreement, a_o, the oldest and the most criticized index of interrater reliability, did well predicting true reliability, showing dr^2=.841 (Line 3, Table 3). Of the seven indices tested, a_o was the only one meeting the primary benchmark dr^2>.8 (Ineq. 11), outperforming the second best, AC_1 (dr^2=.721), and the third best, S (dr^2=.691) by more than 10 points, although the latter two met the tentative benchmark dr^2>.67.

[Insert Table 3 and Figure 1 about here]

The most respected three, π, κ and α, tied as the least accurate predictor, reporting dr^2=.312, failing the tentative benchmark by margins. They also underperformed the next worst, I_r, by 28.7 points (dr^2=.599).

The underperformances of the chance-adjusted indices, especially the popular π, κ and α, were disappointing, considering that the whole mission of the indices was to outperform a_o. The low r^2 means large predictive errors, suggesting that the three indices too often assign lower scores to more reliable instruments, and attach higher scores to less reliable ratings. They failed to differentiate reliable instruments from unreliable ones accurately and consistently.

Figure 2 visualizes the performances and ranks the indices by their dr^2 scores. It is noticed, again, that κ and α ranked among the lowest while percent agreement (a_o) ranked the highest. Figure 2 also shows a strong and positive correlation between accuracy of predicting chance agreement and accuracy of predicting interrater reliability (dr^2=.9768, p<.001), supporting a design feature of this study, which is to analyze the indices' chance estimates for the purpose of understanding the indices.

[Figure 2 About Here]

Approximating Reliability

A .555 average reliability (o_{ri}) was observed (A3, Table 5). The seven indices' estimation of reliability, however, ranged from .237 (π) to .726 (I_r), implying large approximation errors. As expected, percent agreement (a_o) overestimated reliability, reporting e_m=.13 (B6, Table 5) and m_e=.13 (A3, Table 4). The error, however, was below what's allowed by the secondary benchmark, m_e<.2 (Ineq. 13 of the Appendix). So a_o was the only index meeting both primary and secondary benchmarks.

[Insert Table 4 about here]

[Insert Table 5 about here]

Three other indices also met the m_e<.2 benchmark, of which two, AC_1 (m_e=.093) and S (m_e=.096). also outperformed a_o (Line 3 Table 4).

The trio, π, κ and α, again underperformed all others, reporting m_e .323~.327 (Line 8, Table 5). The errors equaled one third of the 0~1 scale, and more than doubled the errors of a_o (m_e=.130). I_r overestimated reliability across the board like a_o did (D6, Table 5), while κ, π and α underestimated across the boa– -- 23.7%~24.1% estimated versus 55.5% observed (Line 3, Table 5).

AC_1 and S underestimated some sessions while overestimated other sessions (Line 6, Table 5). Of AC_1 and S, the under and over estimations offset each other to make the sizes (absolute values) of e_m much smaller than that of m_e. Of the other five indices, e_m and m_e are about equal in size (Line 6, Table 5 vs Line 3, Table 4).

In part because of the offsets, AC_1 and S produced near-zero or very small e_m errors (.001 and .044, respectively), much smaller than any of the other five indices did. By contrast, κ, π and α again produced the largest errors, reporting e_m ranging from -.318~-.314, much worse than the next worst, I_r (e_m=.171, Line 6, Table 5).

Pi-Kappa-Alpha Synchrony

As shown above, *π, κ* and *α* behaved like one index, despite the spirited debates on which of them is the best (10,12,48–51). This pattern of *π-κ-α synchrony* persisted throughout the data.

Impacts of Four Factors

The five viewpoints reviewed earlier discussed four factors behind reliability and/or reliability estimations. Now that we have observed rater behavior, we examine the true impacts of the four factors.

Conjecture Group 1: Chance Agreement Inflates a_o

As said, a 13% chance agreement (o_{ac}) and a 55.5% reliability (o_{ri}) were observed, while percent agreement (a_o) assumed 0% chance agreement and reported a 68.5% reliability, which means a 13-point overestimation (Tables 4 and 5). Conjecture 1 and the century-old beliefs were supported.

(1) Chance agreement exists.

(2) By completely overlooking chance agreement, a_o inflates the estimated reliability.

The data from this experiment, however, adds a third point:

(3) The chance agreement may not be as large as previously thought.

In this experiment, the chance agreement of a_o stayed below the .2 threshold, which was a main factor that allowed the predictive accuracy (r^2) of a_o to stay above the .8 threshold. As a_o outperformed all six indices on the primary benchmark (r^2) and outperformed four out of the six on the secondary benchmark (m_e), an argument could be made that overestimating and misestimating chance agreement can be as counterproductive as overlooking chance agreement.

Conjecture Group 2, Category Inflates S, I_r & AC_1

As critics of S, I_r and AC_1 would have predicted, categoI(C) had large and negative effects on chance estimations S_{ac}, Ir_{ac} and AC_{ac}, with dr^2 ranging -.863~-.661, (p<.001, Line 9, Table 3). Table 6 (K4~K7) shows more details, e.g., S_{ac} was 50% when C=2 but plunged to 12.5% when C=8. The decreases appeared large compared to the 13-point average o_{ac}.

[Insert Table 6 about here]

Negative effects on chance estimations contribute to positive effects on reliability estimations, as shown in the dr^2 ranging .599 ~.721 (p<.001, Line 3, Table 3). S jumped from 40.2% when C=2 to 64.1% when C=8 (C4~C7, Table 6). The effect (difference) of 23.9 points is large compared with the 55.5-point average o_{ri}. In contrast, category effects on the targets of estimations, o_{ri} and o_{ac}, were tiny. Coefficients dr^2 were respectively .003 (p≥.05) and -.019 (p<.01) (A4 and A9, Table 3, See Table 6, Lines 4~7, for more details).

These results support the classic theory that *S* and equivalents underestimate chance agreement when categories exceed two, even when additional categories are largely empty.

The tables also show that I_r and AC_1 relied on category in the same fashion that *S* did and shared the same deficiency. The differences between the category effect on *S*, I_r or AC_1 estimation and the category effect on observed reliability all passed the p<.001 pretest. At the meantime, category showed minimal effects (dr^2≈.001, p≥.05) on π, κ and α, as their authors intended (Line 4, Table 3).

Conjecture Group 3: Skew Depresses κ, π & α

As critics of *κ, π & α* would have predicted, skew had substantial and positive effects on chance estimators κ_{ac}, π_{ac} & α_{ac}, with dr^2 ranging .434~.437 (p<.001, Line 10, Table 3). Table 6 (Lines 8~10) shows more details, e.g., κ_{ac} was 50% when distribution was 50&50, but rose to 67.6% when distribution changed to 1&99.

The positive effects on chance estimates led to negative effects on reliability estimates. Skew effects on the three indices were all negative, with dr^2 ranging -.293 ~ -.292 (p<.001, Line 5, Table 3). When distribution changed from completely even to extremely skewed, the trio's chance agreement estimates increased from about .5 to about .68, and in parallel their reliability estimates decreased from about .37 to about .04, a drop of over 89% (Lines 8~10, Table 6). While mathematical analyses of prior studies had predicted a drop (14,26,52), the empirical evidence of this study showed the drastic magnitude of the drop.

In contrast to the large effects on the index estimators, skew showed minimal effect on the observed estimands, o_{ri} and o_{ac} ($p \geq .05$ for both dr^2, A5 & A10, Table 3), supporting the argument that chance estimates and reliability indices should not rely on skew. Each difference between the skew effect on π, κ or α estimation and the category effect on the observed estimand passes the $p<.001$ pretest.

In another contrast, skew showed practically zero effects on S, I_r or their chance estimates, and a small negative effect on AC_{ac} (dr^2=-.039, $p<.001$, Lines 5 & 10, Table 3). So I_r avoided the skew effect as its authors intended, while AC_1 reversed the effect as its author intended, although the reversed effect was small. A long-suspected pattern was confirmed empiri–lly -- κ, π & α were dependent on skew while S, I_r & AC_1 were dependent on category.

Conjecture Group 4: Indices Overlook Task Difficulty

Difficulty showed a substantial and positive effect on o_{ac} (dr^2=.585, $p<.001$, A11, Table 3), and a large and negative effect on o_{ri} (dr^2=-.774, $p<.001$, A6). A change from extremely easy to extremely difficult decreased o_{ri} by over 68 percentage points and increased o_{ac} by nearly 36 points (Columns A and I, Table 6). These effects appear large compared with 13-point average o_{ac} and 55.5-point average o_{ri}, suggesting that chance estimates and reliability indices should rely on difficulty.

In contrast, difficulty had minimal effects on S_{ac}, Ir_{ac} and AC_{ac} (dr^2=.000~.009, $p \geq .05$, Table 3) and *negative* effects on κ_{ac}, π_{ac} & α_{ac} (dr^2=-.123 or -.125, p<.001, Table 3; c.f. Columns I & N~P, Lines 11~18, Table 6), implying that the indices either failed to rely on difficulty or relied on its opposite, *easiness*, to estimate chance agreement. Each difference between the difficulty effect on chance estimation and the difficulty effect on observed chance agreement was statistically acknowledged at p<.001.

Difficulty showed weaker effects on the six chance-adjusted indices (dr^2=-.566~-.389, Line 6, Table 3) than on the estimation target o_{ri} (dr^2=-.774). Each difference between the difficulty effect on reliability estimation and the difficulty effect on observed reliability was statistically acknowledged at p<.001.

By contrast, a_o, showed a strong and negative correlation (dr^2=-.778, B6, Tables 3) with difficulty. The correlation was as strong as the correlation between o_{ri} and difficulty (dr^2=-.774, A6), suggesting the negative correlations between the chance-adjusted indices and difficulty (dr^2=-.566~-.389) are likely due to a_o embedded in the indices.

Based on derivation and simulation, Gwet concluded that the indices before AC_1 had not handled difficulty properly, and AC_1 handled it better, at least than κ (53–55). The above findings support both claims. The near zero correlation between AC_{ac} and difficulty (dr^2=.009, $p \geq .05$, E11, Table 3), however, suggests that AC_1 still does not handle difficulty well.

Conjecture Group: Indices Assume Intentional and Maximum Random Rating

The precision evidence for the behavioral assumptions behind the statistical indices comes from mathematical analysis. A 2013 study provides detailed scenarios of rater behavior assumed by each of the 22 indices analyzed (14). Readers are invited to derive mathematical formulas from the behavioral scenarios. If a reader-derived formula matches the formular for the corresponding index, then the reader may conclude that the corresponding index indeed assumes the behavioral pattern spelt out in the scenario. If, for example, a formula derived from the Kappa Scenario provided by the 2013 study matches the formula for Cohen's κ published in 1960 (2), it would confirm that κ indeed assumes the rater behavior depicted in the 2013 Kappa Scenario. Such exercises by readers have shown them that chance-adjusted indices all assume that raters regularly conduct *intentional and maximum random rating*.

This study provided corroborating empirical evidence. The indices' chance estimates were poorly correlated with their estimands, the observed chance agreements (Table 3, Line 8). The observed chance agreement (o_{ac}) explained less than 8% of the variance in each of the category-based indices' chance estimates, S_{ac} (2.1%), I_{rac} (2.1%), and AC_{ac} (7.5%). Although the correlations were stronger for the skew-based indices' chance estimates, π_{ac} (-15.1%), κ_{ac} (-15.2%), and α_{ac} (-15.1%), the dr^2 coefficients were all negative, suggesting that the three indices tended to give higher estimates when the true chance agreements were lower, and give lower estimates when the true chance agreements were higher. Clearly, the index-estimated random ratings were not the raters' random ratings observed in this study. This finding supports the argument that the chance-adjusted indices assume intentional and maximum random rating while typical raters conduct involuntary and task-dependent random rating. The mismatch between the assumptions and the observations explains the negligible or negative correlations between the estimates and the estimands.

More corroborating evidence for the maximum-random assumption came from the large overestimation of chance agreement by the six chance-adjusted indices, as shown at Line 12 of Table 5 and the right half of Table 6, summarized in Line 19.

The more situational and detailed evidence of the behavioral assumptions come from the influences of the four factors and the offset and aggravation behaviors of the indices, which are discussed below.

Summarizing the Impact of Four Factors

Each index of interrater reliability implied one or more misassumptions about chance agreement. a_o Overlooked chance agreement. S, I_r and AC_1 inappropriately relied on category. π, κ And α inappropriately relied on skew. While difficulty had a strong and positive effect on chance agreement, all chance adjusted indices failed to rely on difficulty. π, κ and α even relied on its opposite, easiness. The misassumptions, including missed, mistaken, and contra assumptions, impeded estimation. π, κ and α fared worse in part because they entailed more and more devastating misassumptions, some of which had been mistaken as signs of sophistications.

Recall that the main mission of chance adjusted indices is to remove chance agreement in order to improve on percent agreement. When they mishandled the factors affecting chance agreement, they misestimated chance agreement, thereby misestimated reliability. Misassumptions about the four factors are keys to understanding the indices' underperformance.

To understand more, we discuss below the *offsetting* mechanism, which interacts with the assumptions and misassumptions of the indices to define the indices' behavior.

Offsets in Reliability Estimation

Puzzles may arise if one peruses Tables **3** through 6, five of which discussed below.

Puzzle 1. Each chance-adjusted index relied on a wrong factor, skew or category, to estimate chance agreement; none of them relied on a right factor, difficulty. How come some approximated chance agreement far better than others (Line 12 of Table 5 and Line 7 of Table 4)?

Puzzle 2. Chance estimators barely measured the observed chance agreement o_{ac}, or even measured anti o_{ac} (C8~H8 of Table 3). How come the reliability estimations were all positively and sometimes substantially correlated with the observed reliability (C3~H3)?

Puzzle 3. Assuming a negative relation between chance agreement and reliability, one might expect that an over estimation of chance agreement leads to an under estimation of reliability. How come *S* overestimated chance agreement by 100% (o_{ac} =.130 compared to S_{ac}=.260, Line 9, Table 5) while at the same time approximated reliability almost perfectly (*S*=.556, compared to o_{ri}=.555, Line 3, Table 5)?

Puzzle 4. Continued from Puzzle 3, how come AC_1 overestimated chance agreement (e_m=.044, Line 12, Table 5) while also overestimated reliability (e_m =.044, Line 6, Table 5)?

More generally, how come across-the-board overestimations of chance agreement did not translate into across-the-board underestimations of reliability (Line 12 vs Line 6, Table 5)?

Puzzle 5. Continued from Puzzles 3 & 4, how come I_r overestimated chance agreement more than AC_1 did (Ir_{ac}=.131 vs AC_{ac}=.044, Line 12, Table 5), while also overestimated reliability more than AC_1 did (*Ir*=.171 vs AC_1=.044, Line 6, Table 5)?

The puzzles can be explained in part by *offsets*, including *partial offset*, *over offset*, and *counter offset* (*aggravation*) built into the reliability formulas, some of which discussed below.

Category offset, skew aggravation, and skew offset

To understand Puzzle 1, first recall that, under intentional-and-maximum-random assumption, chance-adjusted indices tend to overestimate chance agreement (9,14,29,38,39,56–58). In this experiment, the overestimations ranged from 4.4 percentage points by AC_1 to 44.5 points by Scott's π, all statistically acknowledged (p<.001, Line 12, Table 5).

To explain Puzzle 1, we note that the category-based indices assume that larger number of categories *decreases* chance agreement (C9~E9, Table 3), which *offset* the general overestimation. The skew-based indices assume that higher skew *increases* chance agreement (F10~H10), which *aggravated* the general overestimation. AC_1 assume both, that category and skew both decrease chance agreement (E10), thereby *offset* the overestimation even more than the other two category-based indices.

To illustrate the point, we follow the textbook tradition of starting from *ground zero*, which is the condition of two raters, two categories, and 50&50% distribution. Here, and only here, all major indices gave about the same estimates, $a_c \approx 0.5$ (K2~P2, Table 6). Under intentional-and-maximum-random assumption, two raters draw from marbles, half with one color and half another color; they rate randomly if the colors match, and honestly if mismatch (9,14,29,38,39). Task difficulty is not a factor in this view of rater behavior.

In actual rating, however, a_c=0.5 could occur only if the task is extremely difficult. In our experiment, even the most difficult (d_f=1 for 1-pixel difference) condition did not reach that theoretical maximum, reporting an o_{ac}=.38 (I18, Table 6). The less difficult sessions reported significantly smaller o_{ac}, averaging 0.13 across all levels of difficulty. This means a 37-point initial overestimation at the ground zero by each chance-adjusted index

(e_m=.5-.13=.37).

When category increased from ground zero, S_{ac}, Ir_{ac} and AC_{ac} decreased quickly under the *category assumption* (Columns K~M, Row 4~7, Table 6). While the assumption was unjustified given the small change in o_{ac} (I4~I7), the decrease partially offset the 37-point overestimation, making S_{ac}, Ir_{ac} and AC_{ac} less inaccurate. By contrast, κ_{ac}, π_{ac} & α_{ac} rejected the category assumption to remain unchanged (Columns N~P), hence did not benefit from the partial offset. Thus, S_{ac}, Ir_{ac} & AC_{ac} became less inaccurate than κ_{ac}, π_{ac} & α_{ac}.

Now return to ground zero, then increase skew. Under the skew assumption, κ_{ac}, π_{ac} & α_{ac} increased with skew (Columns N~P, Row 8~10, Table 6). While the assumption was unjustified given the small change in o_{ac} (I8~I10), the increase further aggravated the 37-point overestimation, making κ_{ac}, π_{ac} & α_{ac} even more inaccurate. By contrast, S_{ac} and Ir_{ac} rejected the skew assumption to remain unchanged (K~L, 8~10), hence did not suffer from the aggravation. Thus, κ_{ac}, π_{ac} & α_{ac} became even more inaccurate than S_{ac} & Ir_{ac}.

Rather than accepting or rejecting the skew assumption, AC_{ac} reversed it, by assuming that skew reduced a_c (M8~M10). While the assumption also mismatched the observed skew effects (I8~I10), the decrease further reduced the once 37-point overestimation. Here two unjustified assumptions, *category* and *reversed skew*, joined hands to partially offset another unjustified assumption, *intentional and maximum random*. Thus, AC_{ac} became even less inaccurate than S_{ac} & Ir_{ac}, hence the least inaccurate of the six. As the effect of intentional-and-maximum-random assumption was stronger than the other two effects combined, a net effect was that even *ACac* still overestimated chance agreement.

There were other under-offsets, over-offsets, and counter-offsets, i.e., aggravations, some of which discussed below. Behind multifarious offsets were multifarious assumptions about rater behaviors, which fought or allied with each other or stayed neutral to produce the multifarious outcomes. Two wrongs sometimes made one right, sometimes half right, and often three, four, or more wrongs.

Chance-removal offset

To understand Puzzle 2, we first recall that, assuming intentional and maximum random rating, index designers want to remove maximum amount of chance agreement from all considerations, which requires to remove a_c not only from percent agreement (a_o), but also from the realm of consideration (9,14,23,24,29,38,39). Accordingly, a_c is subtracted twice in Eq. 1, first from a_o in the numerator, and second from 1 in the denominator, which represents 100% of the realm of consideration. Two offsets occurred as a result. First, a_c offsets a_o in the numerator. Second, a_c in the denominator offsets its own impact in the numerator. As the self-offsets weaken a_c's effects, a_o dominates Eq. 1, the indices' estimation of reliability. That explains Puzzle 2: the minimal or negative a_c–o_{ac} correlations exerted weaker effects than the strong and positive a_o-o_{ri} correlation.

The weaker effects still hinder. The chance estimators not only failed to fulfill their prescribed mission of improving on percent agreement, but the estimators worked against the mission. Consequently, all six indices underperformed percent agreement when predicting observed true chance agreement. Ironically, it was the supposedly “most primitive” and “flawed” percent agreement (a_o) that worked inside the indices to keep them from performing and looking even worse (2 p38,12 p80).

The offsets also help to explain Puzzle 3. While *S* overestimated chance agreement by an averaged 13.1 points (Line 12, Table 5), the chance-removal offset helped to bring down the scalar error of reliability estimation to 9.6 points (Line 3, Table 4). This across-session error contains over- and under-estimations of individual sessions, which offset each other in averaging to reduce the vector error to near zero (e_m=.001, Line 6, Table 5. See also the discussion of aggregation bias earlier).

By setting estimated reliability (r_i in Eq. 1) equal to observed reliability (o_{ri} in Eq. 5 of Appendix), $r_i = o_{ri}$, we derive a threshold (t_h) for a_c, which is Eq. 2:

$$t_h = \frac{o_{ac}}{1 - o_{ri}} \qquad 0 \leq o_{ac} \leq t_h \leq \infty \qquad (2)$$

For any rating session, an index accurately estimates reliability when $a_c = t_h$, underestimates when $a_c > t_h$, and overestimates when $a_c < t_h$. Therefore, when $o_{ac} < a_c < t_h$, the index overestimates both the chance agreement and the reliability, explaining Puzzle 4. Across the 384 sessions, average t_h would be .292 if we plug o_{ac} (.13) and o_{ri} (.555) into Eq. 2. As Table 5 shows, of the six chance-adjusted indices, the three (κ, π, α) reporting a_c>.292

(Line 9) also underestimated reliability (Line 6), and the three (S, I_r, AC_1) reporting a_c<.292 also overestimated reliability. At the same time, all six overestimated chance agreement (Line 12). Due to the chance-removal offset, it is possible and possibly common for some category-based indices to overestimate both chance agreement and reliability.

A previously undocumented paradox emerges from this analysis (Eq. 1 and Eq. 2). An index estimates reliability accurately (r_i=o_{ri}) *only* when it overestimates chance agreement (a_c>o_{ac}), an index that estimates chance agreement accurately (a_c=o_{ac}) inevitably underestimates reliability (r_i<o_{ri}), except in the extreme and impractical situation when r_i=o_{ri}=0. The paradox, applicable for all known chance-adjusted indices, is rooted in the chance-removal offset imposed by Eq. 1, which traces back to the intentional and maximum random assumption (14,23,24,26).

Square-root over offset

To understand Puzzle 5, recall that Perreault and Leigh's I_r adopts the chance estimator of S, Ir_{ac}=S_{ac}, and takes the square root of S as the reliability estimation (7). S≤I_r, as I_r=$S^{½}$ for 1≥S≥0 and I_r=0 for -1≥S<0. When chance agreement is overestimated, the square root operation constitutes an additional offset (14). Due to the category-based over-offset of S, I_r overestimates chance agreement more than AC_1; at the meantime, due to the square root over-offset of I_r, I_r overestimates reliability more than AC_1. The two offsets explain Puzzle 5.

A rating session in this experiment simulates a study. In practice, errors do not offset across studies, e.g., one study's overestimation of Disease A does not offset another study's underestimation of Disease B. We should not overemphasize the near-zero aggregated error by *S* shown in e_m or overlook the sizable individual errors by *S* shown in m_e.

Discussion

Main Findings

Of the seven indices, percent agreement (a_o) stood out as the most accurate predictor of reliability (dr^2=.841, Table 3) and the third most accurate approximator (m_e=.130, Table 4). AC_1, the newest and the least known, was the second-best predictor (dr^2=.721) and the best approximator (m_e=.093). *S* ranked behind AC_1 for both functions (dr^2=.691, m_e=.096).

The most respected, the most imposed, and the most applied indices, π, κ and α, ranked the last for both functions (dr^2=.312, m_e=.323~.327).

The indices' underperformances appeared attributable to mismatches between the assumed and observed rater behaviors, and multifarious offsets and aggravations between the misassumptions. Percent agreement assumed zero random rating, leading to the 13-point overestimation of reliability. The other six indices assumed intentional and maximum random rating, leading to a 37-point initial overestimation of chance agreement at "ground zero" for interrater reliability (Line 3, Table 6).

Away from ground zero, *S*, I_r and AC_1 assumed larger number of categories produced less chance agreement, which offset the initial overestimation, while π, κ and α assumed skewer distributions produced more chance agreement, which aggravated the overestimation. The opportune offsets and the austere aggravations explain the smaller approximation errors by the category-based indices than by the skew-based indices. Contrary to the assumptions, neither rating category nor distribution skew showed meaningful effects on the observed true chance agreement.

Difficulty exhibited a substantial and positive effects on chance agreement (dr^2=.585, p<.001, Table 3), while S, I_r, and AC_1 did not rely on difficulty to estimate chance agreement (dr^2=.000~.009, p≥.05). Failing to rely on difficulty further explains the three indices' underperformance in prediction. Moreover, π, κ & α relied on the opposite, easiness, to estimate chance agreement (dr^2 =-.125~-.123, p<.001), which contributed another part to π, κ & α's worse performance than S, I_r, and AC_1.

What Did the Indices Indicate?

An index indicates a certain concept. What did the seven indices indicate? Did they indicate what they purport to indicate?

Percent agreement a_o was the only index meeting the primary benchmark (dr^2>.8), thereby also meeting the competitive benchmark. By overlooking chance agreements, a_o overestimated reliability by 13 percentage points (e_m=m_e=.130, Tables 4 & 5). The error, however, was within the margin allowed by the secondary benchmark (m_e<.2). The overestimation appeared across the board, as shown in Columns A and B (Lines 4 through

18) of Table 6, which implies that researchers and reviewers may manage a_o's deficiency by discounting a certain amount, such as 15 points, treating a_o-0.15 as a crude estimation of reliability. Overall, in this experiment percent agreement behaved as a good predictor and a 13-point over-approximator of interrater reliability.

The other six indices set out to outperform a_o by removing estimated chance agreement a_c. Unfortunately, their a_c estimations failed to accurately estimate true chance agreement o_{ac}. S_{ac}, Ir_{ac}, and AC_{ac} were slightly influenced by o_{ac} (dr^2=.021~.075, p<.01 or p<.001, Table 3). They were instead strongly and negatively influenced by category (dr^2=-.863~-.661, p<.001), suggesting they indicated fewness of category more than they indicated chance agreement. The other three chance estimators, π_{ac}, κ_{ac} & α_{ac}, predicted far less accurately. They indicated mostly skew (dr^2=.434~.437) and, to a lesser extent, easiness, the opposite of o_{ac} (Lines 8-10, Columns F-H, Table 3).

When Eq. 1 was used to remove a_c, a_o offset some impact of a_c, which also self-offset some. The offsets reduced the category and skew effects and kept the index-o_{ri} correlations positive (Line 3-5, Table 3). But still, a_c, the unique core of each index, all impeded the reliability estimation. S_{ac}, Ir_{ac} and AC_{ac} impeded less than π_{ac}, κ_{ac}, & α_{ac} did, allowing S, I_r and AC_1 to predict reliability better than π, κ, & α did (Line 3, Table 3). But the reduced impediments were still impediments. Consequently, none of the chance-adjusted indices had a good chance of outperforming a_o when predicting reliability. Two indices, AC_1 (m_e=.093) and S (m_e=.096), did outperform a_o (m_e=.13) for approximation, which was due more to opportune offsets between misassumptions, and less to removing chance agreements (Line 3, Table 4).

At the end, no chance-adjusted index passed the primary benchmark $dr^2>0.8$. Two, AC_1 (.721) and S (.691), passed the threshold $dr^2>0.67$ for tentative acceptance (Table 3). Being the best approximator, AC_1 (m_e=.093) was the one meeting the competitive benchmark. AC_1 and S were also two of the four indices meeting the secondary benchmark, $m_e<.2$ (Line 3, Table 3).

Category exerted some effects on AC_1 (dr^2=.123) and S (dr^2=.175). Fortunately for the two indices, the category effects were much smaller than the estimand effects of o_{ri} (dr^2=.721 & .691). The two indices underestimated reliability when C=2, and overestimated when $C\geq4$ (Columns A, C and E, Lines 4~7, Table 6). Overall, AC_1 and S were acceptable predictors of interrater reliability, and under- or over-approximators when category was respectively under or over 3.

I_r (dr^2=.599, m_e=.18) failed the tentative benchmark for prediction but satisfied the secondary benchmark for proximity. It overestimated reliability across the board. Overall, I_r was a poor predictor and an 18-point over-approximator of interrater reliability. I_r's overestimation was worse when the number of categories was increased.

The performances of π, κ and α belong to another class. The trio's estimation-estimand correlations (dr^2=.312) were far below the primary benchmark of dr^2>.8 or the tentative benchmark of dr^2>.67; and their approximation errors (m_e=.323~.327) were far above the secondary benchmark m_e<.2. Furthermore, evenness (1-skew) exerted nearly as large effects on the trio (dr^2=.292~.293, Line 5) as their estimand o_{ri} did (dr^2=.312), suggesting that the trio indicated distribution evenness nearly as much as they indicated interrater reliability. More even distributions raised π, κ and α nearly as effectively as higher reliability did, even though skew or evenness showed no effect on observed reliability or chance agreement.

Overall, π, κ & α were crude predictors of reliability and evenness, and 31-point under-approximators of reliability. They were crude because they showed large errors when predicting reliability (dr^2=.312) or evenness (dr^2=.292~.293).

While dr^2 (.292~.293) were too low to make π, κ & α precise indicators of evenness or skew, they were too high to allow the trio to be pure indicators of reliability. The correlation can be even more disconcerting if one considers its impact on the creation and selection of scientific knowledge. Reviewers and researchers use the trio to screen measurements and manuscripts, while trio systematically favor more even distributions, making the world appear flatter. It would be a collective version of the conservative bias, except this one permeates scientific knowledge (59,60). By contrast, a_o showed none of this disparaging deficiency (dr^2=.000).

Conclusion

Like most controlled experiments, this study had limited external validity. The raters made visual judgments, which did not represent all tasks. The categories stopped at eight. The short-bar categories were largely empty by design. Each session had only two raters. The list could go on. To avoid unwarranted generalization, we used past tense to describe the indices' behaviors and their impact.

Our findings, however, have been speculated or predicted by the theoretical analyses, mathematical derivations and Monte Carlo simulations (14,29,64,65,53,55–58,61–63). These studies used no actual measures, specific tasks, human raters, or other specifics that may limit external validity. What some other studies lack in internal validity, this study provides. The validity of our collective knowledge is significantly strengthened by adding empirical studies based on observing rater behavior.

The indices were advertised to be "standard" and "global" for "general purpose" (12,14,66,67). Now that some reigning indices did not perform as advertised against one set of observed behavior, it is good evidence that indices are not general or global or standard. The burden is not on doubters to prove that the indices always fail, but on defenders to demonstrate that the indices perform, at least sometimes.

Despite the lack of empirical evidence in support of the reigning indices, the spiral of inertia may continue, forcing some and enticing others to work with the indices (26,52). In that event, the interpretation of π, κ and α may warrant more caution, and the application of a_o and AC_1 may deserve more credence, to the extent that findings of this experiment will be replicated.

Future Research

Replication studies. More controlled experiments are called for to falsify or qualify the findings of and the theories behind this experiment, and to test the other reliability indices against their estimands (66,68,69).

New Indices. New indices may be needed. Index designers may be more cautious about the assumptions that raters conduct intentional and maximum chance rating, or their chance rating is determined by skew or category. More thoughts may be given to the possibility that raters conduct instead involuntary and task-dependent random rating, and more weights given to task difficulty. The index designers are encouraged to assess and adjust their ideas and indices against behavioral data, including the data from this experiment, which will be made public upon publication of this manuscript.

REORD and Behavior-based statistical methods. Mathematical statistics use a system of axioms and theorems to build tools for analyzing behavioral data. The REORD (reconstructed experiment on real data) methodology reverses the logic, using observed behavior to inform statistical methods. The application might not be limited to interrater reliability. REORD, for example, may open a new front for the studies of sensitivity and specificity measures, two practical tools often used in medical and health research. REORD may also help to investigate the empirical relationship between reliability and validity, two of the most fundamental concepts of scientific enquiry.

Rater expectations of prevalence or skew. The researchers in this REORD experiment told the raters nothing about the prevalence or the skew of the long and short bars. As prevalence and skew were programmed to vary randomly between trials and between rating sessions, the researchers themselves did not know about the prevalence or skew until data analysis, and the raters could not have guessed accurately. This design feature was chosen because it resembled one type of research condition, under which raters don't know what to expect, therefore they don't expect.

For some tasks, however, raters do expect about prevalence and skew, due to their prior experience with the same tasks or their prior exposure to second-hand information. A follow-up study may investigate the impact of such expectations on raters' rating or the indices of reliability, sensitivity, and specificity.

Human vs machine raters. Expectations about distribution, prevalence, and skew can be programmed into artificial intelligence (AI) to aid automated diagnoses, judgements, scorings, evaluations, ratings, and other decisions by machines. Unlike human decisions and human expectations that are often vague and varying, machine decisions and machine expectations can be programmed to be super clear and super consistent (70,71). Topics of human-machine reliability and inter-machine reliability versus inter-human reliability could be fruitful and fascinating for research using REORD, and so could topics of sensitivity, specificity, and validity with human and/or machine raters.

Declarations

Ethics approval and consent to participate

The survey study received ethical approval under the ethics procedures of University of Macau Panel on Research Ethics (reference SSHRE22-APP016-FSS). Written consent for the survey was also taken.

All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects/participants and/or their legal guardian(s).

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research is supported in part by grants of University of Macau, including CRG2021-00002-ICI, ICI-RTO-0010-2021, CPG2021-00028-FSS and SRG2018-00143-FSS, ZXS PI; Macau Higher Education Fund, HSS-UMAC-2020-02, ZXS PI; Jiangxi 2K Initiative through Jiangxi Normal University School of Journalism and Communication, 2018-08-10, Zhao PI.

Acknowledgements

The authors gratefully acknowledge the contributions of Hui Huang and Chi Yang to the execution of the reconstructed experiment.

Authors' contributions

XZ designed the study, supervised the construction of the experimental site, organized the data collection, conducted the data analysis, and drafted the manuscript. GCF provided feedback for the research design, assisted with data analysis, and provided comments. SHA and LPL provided input into the manuscript writing. All authors read and approved the final manuscript.

Appendix to "Interrater reliability estimators tested against true interrater reliability"

V: Benchmarks and Thresholds

V.1. Ideal index outperforms all others

V.2. Reliability over chance agreement

V.3. Prediction (dr^2) over approximation (m_e & e_m)

V.4. m_e *over* e_m

V.5. Primary Requirement

V.6. Secondary Requirement

V.7. Tentative Requirement

V.8. Competitive requirement

I. Five Concepts and Five Viewpoints about Interrater Reliability

Section I discusses five concepts and five viewpoints about interrater reliability to supplement the literature review and design sections of the manuscript.

I.1. Five Concepts

Five of the fundamental concepts, interrater reliability (r_i), chance agreement (a_c), rating categories (C), distribution skew (s_k), and task difficulty (d_f), is explicated below (72). Indicators of r_i and a_c were measured as dependent variables in this experiment, while C, s_k and d_f were manipulated as the three independent variables.

I.1.1. Interrater Reliability (r_i). *Interrater reliability* (r_i) refers to the true agreement between raters, aka coders, engaged in systematic and task-driven rather than random rating, aka coding. Indices of interrater reliability are meant to estimate this true agreement. As chance agreement (a_c), defined below, is believed to inflate reliability estimate, many indices attempts to estimate and remove a_c (7,53,55,73,74). All major indices, including the six examined in this study other than *%-agreement* a_o, share Eq. 1 to remove a_c and estimate r_i.

$$\boldsymbol{r_i = \frac{a_o - a_c}{1 - a_c}} \qquad (1)$$

A main objective of this study is to assess the seven indices of interrater reliability against observed true reliability (o_{ri}). The eight measures also serve as dependent variables, on which the effects of category, skew, and difficulty are assessed and compared.

I.1.2. Chance Agreement (a_c). This study also measured seven chance agreement (a_c) variables, one chance estimate for each of the six chance-adjusted indices plus observed true chance agreement (o_{ac}). There was an implied eighth chance indicator, by percent agreement (a_o), which is, by definition, a constant at zero.

Chance agreement (a_c) refers to the agreement produced by random rather than systematic and task-driven rating. Five indices invented their own chance estimators while I_r adopted the estimator from *S*.

In addition to comparing the indices with observed reliability, it is important to also compare the indices' chance estimates with observed chance agreement. In Equation 1, subtraction of a_c in the nominator decreases r_i, while the subtraction in the denominator increases r_i. The varying offsetting obscures the differences between indices (14). Since the main or only difference between many indices is in chance estimators (a_c), comparing a_c with its estimation target (o_{ac}) may tell us more about the inside mechanism at the core of the indices.

The seven chance estimate measures also serve as dependent variables, on which the effects of category, skew, and difficulty are assessed and compared.

I.1.3. Category (C). *Category* (*C*) was defined as the number of choices available to a rater on a nominal scale. For example, variable *gender* often has two categories, while *party affiliation* in U.S. may have four, democrat, republican, independent, and others.

I.1.4. Distribution Skew (s_k). Distribution, aka base rate, frequency, marginal, or prevalence, refers to the pattern of percentage occurrences, e.g. 49% female and 51% male, or 5% unhealthy and 95% healthy (27,29,76,77,53–58,65,75). Major indices are symmetrical, centered on 50&50% distribution. Accordingly, this study folded the original distribution to create *distribution skew* (s_k), which served as a main independent variable.

I.1.5. Difficulty (*df*). *Difficulty* (d_f) represents the combination of all factors that make rating inaccurate, including 1) *task difficulty*: Some tasks are more difficult than others; 2) *rater difficulty*: Some raters are less capable, focused, or motivated than others, which increases difficulty; 3) *instrument difficulty*: Instruments are means that help raters to accomplish a task, including organization, instruction, training, and equipment. Deficient instruments increase difficulty. This study fixed instrument difficulty at the lower end by giving easily understood tasks and instructions. We manipulated task difficulty and assumed variation in rater difficulty.

I.2. Five Viewpoints

Five viewpoints have influenced experts' understanding of interrater reliability. They are also the theoretical focal points of this study.

I.2.1. Chance agreement inflates a_o. In the academic literature on interrater-interrater reliability, likely the earliest and the most widely received viewpoint is that percent agreement (a_o) inflates reliability by overlooking *chance agreement* (a_c). Consequently a_o is considered “the most primitive,” (2 p38) “inadequate,” (13 pp187&193) and “flawed,” (12 p80) therefore “should *not* be used.” (3–5,13 p187). Removing chance agreement is the core or the stated mission of early indices, e.g., Benini’s β (11). Bennett et al’s *S* (15), Goodman & Kruskal’s λ_r (78) and Guttman’s ρ (79). Of these, only *S* remains in regular use today (28, 29).

I.2.2. Rating category inflates *S, I_r, and AC_1*. Another widely shared viewpoint is that *S* depends on category while it should not. Large number of *categories*, even if empty, deflates chance estimates of *S* (S_{ac}), thereby inflates *S* (16,49,56–58,66). The criticism also applies to six equivalents or special cases of *S*, namely *C* (80), *G* (81,82), k_n (61), *PABAK* (83), *RE* (84), and *redefined Pi* (85).

Perreault & Leigh took the square root of *S* to produce I_r (7). Gwet[41–44] incorporated the entire *S* into his AC_1. So category affects I_r and AC_1 in a similar way as it affects *S*, according to mathematical analysis and simulation (14,56–58), although some consider I_r “the best” (87,88,89 p.384). As I_r regularly produces higher scores than other indices, its popularity has grown fast in some fields (25).

Eliminating category effect was a main justification for Scott (16) to offer π, which in turn inspired Cohen’s κ (2) and Krippendorff’s α (19,67). Not suffering from category effect is a main reason that methodologists recommend π, κ or α over alternatives (12,14).

I.2.3. Distribution skew deflates π, κ & α. Considered the "statistics of choice" (90 p140), κ is by far the most often used index across disciplines, followed by π and α (3,4,94,14,51,56–58,91–93).

A controversial viewpoint is that π, κ and α depends on distribution skew while they should not. The trio, critics argue, mistakenly assumes that more skewed distributions create more chance agreements. Consequently, higher or lower prevalence of a variable, e.g., disease, produces larger estimates of chance agreement, thereby deflates estimated reliability (3,4,56–58,61–63,65,74,76,77,5,83,93–100,7,10,27,29,51,53,55).

By contrast, AC_1 assumes a negative skew effect on chance agreement, while I_r follows *S* to assume no skew effect.

The alleged dependence of π, κ and α on skew ignited repeated and spirited debates. Experts defended κ by reaffirming its validity, extending its application, or teaching its use (48,75,101–108). Rogot & Goldberg introduced A_2, a mathematical equivalent of κ (109). Byrt and colleagues introduced *BAK* (83), and Siegel & Castellan introduced *Revised K* (110), which are two equivalents of π. Krippendorff advocated and defended α vigorously (12,49,50). Zwick recommended π over κ and *S* (10), while Hsu & Field recommended κ over π (48). Vach opined that the dependence on skew is harmless (107 p655), and Krippendorff acclaimed that the dependence is desirable and by design (49,50).

I.2.4. Reliability indices overlook task difficulty. An emerging viewpoint is that indices of interrater reliability should depend on *task difficulty,* but they do not. More difficult tasks induce more chance rating, therefore more chance agreements (9,14,86,95,29,38,53–58). Krippendorff , however, opined the opposite, that "more complex" tasks lead to "very small" chance agreement (50 p488).

I.2.5. Indices assume intentional and maximum random rating. Among the most fundamental hence the most forcefully debated views is that the chance-adjusted indices all assume intentional and maximum random rating by conspiring raters, which include all raters for all ratings, all the time (9,14,23,24,26,28,52,111). The raters, according to this assumption, agree *a priori* to do the following -

1) To "rate" at the commands of randomization devices, e.g., randomly thrown coins, rolled dice, or drawn marbles, virtual or actual, without looking at the subjects under rating,

2) To rate truthfully *only* when the randomization devices disagree with each other, therefore rendering no consistent command for raters to follow.

Krippendorff rejected this view regarding Krippendorff's α, and characterized the discussion as "strange, almost conspiratorial uses of language." (50).

Bipolar all-or-nothing assumptions were detected hidden in the indices. Percent agreement assumes absolutely no random rating, while the chance-adjusted indices assume intentional and maximum random rating. The latter group assume that raters draw virtual or actual marbles before any “rating;” they “rate” by the order of the marbles whenever the marbles agree to give a consistent order; they rate honestly only when the marbles disagree with each other thereby giving no consistent order (9,14,29,38,39,56–58).

Different indices assume different ways that raters arrange the virtual or actual marbles for the random drawing and rating. S, I_r and AC_1 assume that raters arrange the marbles evenly across color types that are matched with rating categories, causing the triad's dependence on rating category. π, κ And α assume that raters match the distribution of marble colors to the pre-determined but post-reported target distribution, causing the trio’s dependence on target distribution and skew. As said, Krippendorff denied that α makes such assumptions (50,112–114).

The key questions, therefore, are about rater behavior. What behaviors are assumed? What behaviors take place? Do the assumptions match the behaviors? Reliability researchers rely on theoretical arguments, mathematical derivation, fictitious examples, naturalistic comparisons, and Monte Carlo simulation. A systematic observation of rater behavior is needed to inform the debates over rater behavior.

This paper reports a controlled experiment that manipulated category, skew, and difficulty, and observed raters’ behavioral responses. Seven indices of interrater reliability were tested against the observed behavior. The findings also apply to the two equivalents of

a_o, six equivalents of S, two equivalents of π, and one equivalent of κ, covering 18 indices in total.

II. Reconstructed Experiment with Golden-Standard Task

Section II details the design and the execution of the reconstructed experiment that provided the main empirical evidence for this study.

We programmed a website that asked raters to identify the longest bar from several bars (Figure 1). Two of the independent variables, category, and difficulty, were manipulated by programming the website.

II.1. Manipulating Category (*C*). *Category* (*C*) was manipulated by giving raters two, four, six or eight bars to choose from. Thus, *C* had four values, 2, 4, 6 and 8.

II.2. Manipulating Difficulty (d_f). *Task difficulty* (d_f) was manipulated by varying the differences between two longest bars. The differences ranged from one pixel, the smallest controllable element on a computer screen, to eight pixels, which were clear to nearly everyone. The variable d_f was linearly transformed to a 0~1 scale where 1 represents the most difficult.

The two longest bars (*long bars*) were 200 pixels long plus or minus 0~4 pixels for the manipulation of difficulty. The lateral distance between long bars was fixed at 150 pixels to minimize distance effect.

We confined the main competition between the long bars. Few raters chose the short bars as they were clearly shorter, which made this experiment very close to Scott's empty-category assumption and minimized the correlation between *category* and *difficulty* (16).

II.3. Creating One-way Golden Standard. A *gold standard* is a consensus criterion under which judgments can be made with certainty. Reliability indices are standards to evaluate instruments. Now that we are to evaluate the standards, a golden standard would be helpful if available. The longest-bar task provides such a golden standard. Through programming codes, we the researchers always know with certainty which bar was the longest, and whether each rating decision was right or wrong, based on which chance agreement and true reliability can be calculated and analyzed. We use "golden standard" as a stronger term than "gold standard." The latter term was borrowed by Rudd in 1979 from economics where it referred to the value of gold as a monetary standard (37,115).

So that variables vary, the golden standard needs to be equipped with a one-way mirror that is always crystal clear to researchers, but variably clear to participants. The longest-bar task also provides this figurative or virtual mirror, as the task was designed such that raters sometimes knew with near certainty, but sometimes did not, thereby they had opportunities to rate randomly and agree by chance.

II.4. Pairing Rater Responses. Each rater rated 10 items per period and was given summary statistics of right and wrong at the end of each period. The task was made to resemble an online game or IQ test to maintain raters' attention and focus. Items per period were limited to 10 to reduce clutter effect (116,117). Number of bars, level of difficulty, and the location of long bars were randomly rotated to minimize the effects of learning, fatigue, boredom, serial position, rater idiosyncrasies, and other possible confounders (117–121).

The same 10 items were rated again in the same order by the next rater available. After completing 10 items, a rater may choose to rate 10 more. He or she might be given 10 unpaired items rated by another rater, or 10 new items if all rated items had been paired. The process repeated until the end of data collection.

The data collection took place in a three-month period. Students, teachers, researchers, technicians, managers, office workers and other professionals from 15 colleges and two research firms in America, China mainland, Hong Kong, Macau and Singapore participated as a part of their class exercises, professional training, or work assignments. They registered 383 web names and logged on from 53 Asian, European and North American cities. They rated a total of 22,290 items, of which 19,900 were successfully paired, producing 9,950 paired responses, from which we sampled and resampled to reconstruct 384 rating sessions to form a between-subject (session) experiment that we report below.

II.5. Manipulating *Skew* (s_k). As the longest bar is either at the left or right side of the second longest bar, we defined *distribution* as the left-and-right percentage. For example, when 1% of the rated screens had the longest bar at the left, the distribution is denoted 1&99. Five levels were chosen: 1&99, 25&75, 50&50, 75&25, and 99&1, the last of which represented 99% left & 1% right. 0&100 and 100&0 were omitted as π, κ and α would be undefined.

It is skew, but not the unfolded distribution, that's expected to affect the indices (14,29,61,65,76). Therefore, *skew* (s_k) was operationalized as *distribution folded in the middle*. 1&99 and 99&1 were both assigned s_k=0.99, for the highest skew. 50&50 was assigned s_k=0.5 for the lowest skew, and 25&75 and 75&25 were both assigned s_k=0.75 for moderate skew. Variable *skew* (s_k) ranged 0.5~0.99.

II.6. Reconstructing Rating Sessions. To reconstruct the first rating session, we randomly sampled without replacement 100 paired rating responses (N_t=100) requiring two *categories* (C=2), lowest *difficulty* (d_f=0), and highest *skew* (s_k=.99). After recording the variable and response information, we returned the sample to the population of 9,950.

To reconstruct the second rating session, we drew another random sample of 100 pairs requiring four *categories* (C=4) while the other two variables, *difficulty* and *skew*, remained d_f=0 and s_k=.99. Again, we returned each pair back to the population after recording the needed information. We then reconstructed the third session, then the fourth, and so on. We repeated the process for every combination of *category*, *difficulty*, and *skew*, producing 4*8*3=96 sessions.

A few cases can significantly affect π, κ and α when distribution is skewed (51,56–58,62,63,77,93,108,122). To assure stable effects, we resampled three more times to quadruple the number of sessions, so N_c=96*4=384, which was the total number of the reconstructed rating sessions that constituted the “subjects” for this experiment. Each skew condition had an equal number of high- and low- prevalence sessions, that is, each skew=.99 condition had two 1&99 sessions and two 99&1 sessions, and each skew=.75 condition had two 25%75 conditions and two 75&25 sessions.

II.7. Reconstructed Experiment in Summary. This was a 4X8X3 between-subject controlled experiment with 4 subjects per cell where each subject was a rating session, as shown in Table 1. The execution took two stages. The first was *individual-level treatment-response*, during which individual-level independent variables, category and difficulty, were manipulated, stimulus and treatment were administered, and individual responses were recorded. The second was *group-level reconstruction*, during which individual responses were sampled and resampled, and the group-level independent variable, skew, was manipulated.

While the treatment and response collection followed the procedure of typical controlled experiment (36), the sampling and resampling benefited from the theories and techniques of bootstrap (32,33); jackknife (34) and Monte Carlo simulation (35).

Simulation is a powerful tool for understanding reliability. But simulations do not measure behavior. They presume certain behaviors then examine their consequences (53,55–58,123). A typical individual-level experiment is unsuitable because reliability indices are meaningful only for rating sessions. A session-level experiment would require hundreds of rating sessions, which would be too costly and too difficult to administer. Each rating session would require a fixed level for each independent variable, e.g., all tasks are extremely difficult, have eight categories, and 99% are left, which would deviate too much from realistic rating. Reconstructed experiment offers a useful and feasible addition to our toolkit, allowing observed rater behaviors to be factored into the debate over how raters behave.

III. Variable Measurements and Calculations

III.1. Calculating Chance Agreement (o_{ac}). The raters reported few agreements on short bars (0.45%, Table 2), confirming that the main competition was successfully limited between the long bars. It also simplifies the calculation for chance agreement. Assuming no deliberate and systematic errors, each *erroneous agreement* (o_{ae}), the agreement between two raters choosing a same wrong bar, is considered random. Because there were only two real choices, the probability theory predicates an equal number of agreements falling on the longest bars, thus being correct by chance. Therefore, *observed chance agreement* (o_{ac}) was calculated by doubling the directly observed erroneous agreement o_{ae}:

$$o_{ac} = 2 * o_{ae} \qquad (3)$$

To be sure, we derived another formula for o_{ac} assuming that sometimes raters had four, six, or eight real choices, as described in Section III.2 below. The two measures yielded essentially the same results. As Eq. 3 is simpler and easier to trace back to the directly observed o_{ae}, we report statistics based on Eq. 3.

III.2. Alternative Calculation of Observed Chance Agreement (o_{ac}). We identified two formulas for calculating the observed chance agreement (o_{ac}). The findings section of the manuscript reports the results based on the simpler formula (Eq. 3). All analyses involving o_{ac} were performed twice using the two different formulas, which produced essentially the same results. We describe the alternative formula (Eq. 4) below.

Some agreements are right, some are erroneous. This study directly observed erroneous agreement (o_{ae}). As we assume no systematic error, all o_{ae} are assumed to have come from chance rating, which constitutes the first part of the chance agreement to be estimated.

The observed right agreement (o_{ar}) includes randomly and systematically right agreement. We need to estimate the former. Due to our design of two long bars and several (0, 2, 4, 6) short bars, the chance agreement came from two types of random selection: between two long bars, and among all bars. When the latter results in an agreement on the longest bar, we call it *right agreement from random choices among all bars* (a_{ra})

All agreement on the short bars resulted from raters choosing randomly among all bars. With C categories, 1/C of such random choices should fall on each bar, including the longest bar. Suppose there are four bars (C=4), and o_{s4} represents observed agreement on the two short bars, the right agreement (on the longest bar) from choosing randomly among four bars equals the agreement on each short bar, which is $o_{s4}/2$. Similarly, the right agreement from choosing randomly among six or eight bars is $o_{s6}/4$ or $o_{s8}/6$, respectively. So the total amount of right agreement from random selection among all bars is $a_{ra}=(o_{s4}/2)+(o_{s6}/4)+(o_{s8}/6)$, which constitutes the second part of the chance agreement we want to estimate.

Of all observed agreements on the second longest bar (o_{a2}), some came from random selection among all bars (a_{ra}), and the rest (o_{a2}-a_{ra}) came from random selection between the two long bars. The same amount (o_{a2}-a_{ra}) should fall on the longest bar, which constitutes the last part of the chance agreement we want to estimate.

Adding up the three parts, the observed chance agreement o_{ac} is:

$$o_{ac} = o_{ae} + a_{ra} + (o_{a2} - a_{ra}) = o_{ae} + o_{a2} \quad (4)$$

As mentioned, the two approaches of calculating o_{ac} produced very small differences in means and even smaller differences in correlations. The two formulas therefore corroborate each other.

III.3. Calculating Observed Reliability (o_{ri}). Observed reliability (o_{ri}) is observed agreement (a_o) minus observed chance agreement (o_{ac}):

$$\boldsymbol{o_{ri} = a_o - o_{ac}} \quad (5)$$

IV. Statistical Indicators

Typical studies calculate estimators to estimate estimands, the targets of estimations. This study observed estimands to evaluate their estimators. We adopted and adapted common indicators, *mean*, *error*, and r^2, to analyze data from this novel design with novel objectives. To guide our choices, we first review the two functions of interrater reliability as estimators.

IV.1. Approximating and predictive functions of reliability indices. Reliability indices serve two functions. One is to compare an instrument with fixed benchmarks, such as 0 for absence of reliability, 0.67 for highly tentative reliability, 0.8 for acceptable reliability, and 1 for perfect reliability (19 p147). This function requires an index to *approximate* true reliability in order to *place* accurate scores on instruments, and we need a proximity measure(s) to assess and analyze indices' ability to approximate true reliability.

Another function is to compare instruments with each other in order to *differentiate* them. This function requires an index to accurately *predict* true reliability, which means to be highly and positively correlated with its estimation target, so that it almost always gives higher scores to more reliable instruments and lower scores to less reliable instruments. We need a correlational measure(s) to evaluate the indices' ability to predict true reliability.

If an index always approximates the reliability of every individual session perfectly, it also predicts perfectly. Assuming no perfection, however, the prediction-proximity relation is more complicated. A good predictor is not necessarily a good approximator. For example, if a perfect predictor always overestimates by a constant, it's still a perfect predictor, because all instruments benefit equally. Conversely, a good approximator is not necessarily a good predictor. While a dreadful approximator gives higher score to worse instruments and lower scores to better instruments, its errors could offset each other to make it a perfect approximator on average. Therefore, both proximity and prediction measures are needed.

IV.2. Proximity Measure I -- Error of Mean (e_m). An intuitive proximity measure is *error of mean* (e_m), defined as the difference between the grand average (*mean*) of estimations (r_i or a_c) and the grand average (*mean*) of estimation targets (o_{ri} and o_{ac}) . For any reliability index r_i and chance estimator a_c, the error of mean (e_m) calculations are shown as Eqs. 6 and 7.

$$\boldsymbol{e_m}(\boldsymbol{r_i}) = \mathbf{mean}(\boldsymbol{r_i}) - \mathbf{mean}(\boldsymbol{o_{ri}}) \qquad -1 \le e_m(r_i) \le 1 \qquad (6)$$

$$\boldsymbol{e_m}(\boldsymbol{a_c}) = \mathbf{mean}(\boldsymbol{a_c}) - \mathbf{mean}(\boldsymbol{o_{ac}}) \qquad -1 \le e_m(a_c) \le 1 \qquad (7)$$

For example, the difference ($e_m(r_i)$) between κ estimation (r_i) and observed reliability (o_{ri}), averaged across 384 sessions, would indicate one aspect of κ's inaccuracy.

As a vector, a positive e_m indicates overestimation, while a negative e_m indicates underestimation. A near zero e_m, however, does not necessarily indicate accuracy for individual rating sessions. Overestimations and underestimations of individual sessions may offset each other to create a small e_m, a phenomenon known as *aggregation bias* or *ecological fallacy* (124,125).

In typical research, however, overestimation of one study does not offset the underestimation of another study. Errors of all directions accumulate or even multiply in terms of social impact. We need an additional measure, which is described below.

IV.3. Proximity Measure II -- Mean of Errors (m_e). To avoid aggregation bias, we took the absolute value of the estimation error of each session, $|r_i\text{-}o_{ri}|$ and $|a_c\text{-}o_{ac}|$, and averaged them across all 384 sessions. The results are *mean of errors* (m_e) for reliability and chance estimations for reliability (r_i) and chance errors (a_c), as shown in Eqs. 8 & 9:

$$\boldsymbol{m_e(r_i)} = \mathbf{mean}(|\boldsymbol{r_i} - \boldsymbol{o_{ri}}|) \qquad 0 \le m_e(r_i) \le 1 \qquad (8)$$

$$\boldsymbol{m_e(a_c)} = \mathbf{mean}(|\boldsymbol{a_c} - \boldsymbol{o_{ac}}|) \qquad 0 \le m_e(a_c) \le 1 \qquad (9)$$

Smaller m_e indicates a smaller error hence a better estimator. As a scalar, however, m_e does not differentiate overestimations from underestimations, which vector e_m does.

The spreads of our main variables varied significantly (Lines 4,5,10 &11 of Table 3), which presents another concern. A narrower spread makes e_m and m_e look closer to zero because their baselines (-1~1 or 0~1) do not change with spreads, producing a statistical version of *baseline bias* (126) or *scale of reference bias* (127).

IV.4. Predictive Accuracy and Share of Influence -- Directional r^2 (dr^2). As a ratio of regression prediction over total variance, r^2 is commonly used to measure *predictive accuracy* (128–130). As a percent of dependent variance explained by independent variable(s), r^2 also indicates *share of influence* (128,129,131). As a scalar, however, r^2 does not signal direction, while direction is important for this study. There are conflicting expectations about how difficulty or skew affects chance agreement, for example. We added the sign of r to r^2 to produce a *directional r squared* (dr^2):

$$\boldsymbol{dr^2 = r * |r|} \qquad \boldsymbol{-1 \le dr^2 \le 1} \qquad (10)$$

We use dr^2 as the main indicator of indices' predictive accuracy and various variables' share of influence.

IV.5. Regression vs ANOVA. Experimenters often employ ANOVA for analyzing data. The independent variables of this experiment are on ratio scales, which can be more efficiently analyzed with regression. As regression and ANOVA are mathematically equivalent, there is no loss in essential information or accuracy.

V. Benchmarks and Thresholds

This is the first time interrater reliability estimators and their chance agreement estimators are evaluated against their respective estimands, the observed true reliability and observed true chance agreement. No preestablished benchmarks or thresholds are available. Before reporting the outcome, this section lays out the principles that guide the evaluation. Besides helping the reviewers to evaluate our evaluation, we also hope that explicating the principles, if published, may start a conversation about what criteria and principles are appropriate for this type of evaluations.

V.1. Ideal index outperforms all others. An ideal index outperforms all other indices on all indicators, producing the largest dr^2 and smallest m_e and e_m for both reliability and chance estimations. Since no such index emerged, the following principles applied.

V.2. Reliability over chance agreement. While chance estimation is important for understanding an index's inside, an index's value is ultimately judged by the accuracy of its reliability estimation.

V.3. Prediction (*dr^2*) over approximation (*m_e* & *e_m*). As said, a good predictor usually gives more reliable instruments higher scores, and less reliable instruments lower scores. A good predictor can be a poor approximator only when its estimations deviate from the true reliability by a near constant across all studies. If the constant can be estimated, such as in studies like this, researchers can add the constant to the estimations to improve the approximation. If the constant cannot be estimated, researchers may collectively adjust the benchmarks to reduce the impact of the across-the-board miss-approximation.

When a good approximator is a poor predictor, its consequences are more severe and harder to remedy. A poor predictor often gives more reliable instruments lower scores, and less reliable instruments higher scores. A poor predictor can be a good approximator only when its errors on individual studies offset each other to lower the across-study errors. The offsetting through averaging does not remedy the underlying cause of the large estimation errors shown in the low correlation.

If we cannot have both, we would trade approximating precision for differentiating precision. When evaluating reliability indices, therefore, more weights should be placed on dr^2 than m_e or e_m.

V.4. m_e over e_m. To evaluate the indices' approximation accuracy, we place more weights on mean of errors (m_e) because it is less influenced by aggregation bias.

V.5. Primary Requirement. Some disciplines honor r_i>0.8 as the criterion for acknowledging reliability, and r_i>0.67 for highly tentative acknowledgment (19,49,132). Without more reasonable precedents to following, this study tentatively adopts 0.8 and 0.67 as thresholds for dr^2, m_e and e_m. In accordance with Reasoning VI.3 above, we consider Inequality 11 a primary requirement for accepting an index's validity, where $dr^2_{(ori\&ri)}$ represents directional r^2 between observed reliability (o_{ri}) and an index's estimated reliability (r_i):

$$\boldsymbol{dr^2}_{(\boldsymbol{o_{ri}}\&\boldsymbol{r_i})} > \mathbf{0.8} \qquad -\mathbf{1} \leq \boldsymbol{dr^2} \leq \mathbf{1} \tag{11}$$

The stated mission of chance-adjusted indices is to outperform percent agreement (a_o), which requires Inequality 12, where $dr^2_{(ori\&ao)}$ represents directional r^2 between o_{ri} and a_o.

$$\boldsymbol{dr^2}_{(\boldsymbol{o_{ri}}\&\boldsymbol{r_i})} \geq \boldsymbol{dr^2}_{(\boldsymbol{o_{ri}}\&\boldsymbol{a_o})} \qquad -\mathbf{1} \leq \boldsymbol{dr^2} \leq \mathbf{1} \tag{12}$$

Inequality 11 applies when $dr^2_{(ori\&ao)}$<0.8, otherwise Inequality 12 applies.

V.6. Secondary Requirement. Inequalities 13 &14 serve as the secondary requirement, where $m_{e\,(ri)}$ and $m_{e\,(ao)}$ represent respectively approximation errors (m_e) of an index (r_i) and a_o.

$$\boldsymbol{m_{e_{(r_i)}} < 0.2 \qquad 0 \leq m_e \leq 1} \tag{13}$$

$$\boldsymbol{m_{e_{(r_i)}} \leq m_{e_{(a_o)}} \qquad 0 \leq m_e \leq 1} \tag{14}$$

Inequality 13 applies when $m_{e\,(ao)}$>0.2; Inequality 14 applies otherwise. The threshold 0.2 in Ineq. 13 comes from 1-0.8=0.2, where 0.8 is borrowed from, again, from Krippendorff's criteria (19,49,132).

V.7. Tentative Requirement. In case no index meets the primary and secondary requirements, thresholds of 0.67 for dr^2 and 0.33 for m_e may be applied for tentative acceptance, again borrowing Krippendorff's criteria (19,49,132).

V.8. Competitive requirement. To be among the recommended, an index also needs to outperform all other indices on at least one of the major indicators.

References

1. Artstein R. Inter-annotator agreement. In: Ide N, editor. Handbook of Linguistic Annotation [Internet]. Springer Netherlands; 2017 [cited 2022 Jan 15]. p. 297–313. Available from: https://link.springer.com/chapter/10.1007/978-94-024-0881-2_11
2. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas [Internet]. 1960 [cited 2022 Jan 15];20(1):37–46. Available from: http://psycnet.apa.org/index.cfm?fa=search.displayRecord&uid=1960-06759-001
3. Feng GC. Estimating intercoder reliability: a structural equation modeling approach. Qual Quant [Internet]. 2014 Jul 20 [cited 2022 Jan 15];48(4):2355–69. Available from: http://link.springer.com/10.1007/s11135-014-0034-7
4. Feng GC. Intercoder reliability indices: Disuse, misuse, and abuse. Qual Quant [Internet]. 2014 [cited 2022 Jan 15];48(3):1803–15. Available from: http://link.springer.com/article/10.1007/s11135-013-9956-8
5. Feng GC. Mistakes and how to avoid mistakes in using intercoder reliability indices. Methodology [Internet]. 2015 [cited 2022 Jan 15];11(1):13–22. Available from: http://econtent.hogrefe.com/doi/full/10.1027/1614-2241/a000086
6. Grayson K, Rust R. Interrater reliability. J Consum Psychol [Internet]. 2001 [cited 2022 Jan 15];10(1/2):71–3. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/22114173%5Cnhttp://linkinghub.elsevier.com/retrieve/pii/S1057740801702471

7. Perreault WD, Leigh LE. Reliability of nominal data based on qualitative judgments. J Mark Res. 1989;26(2):135–48.

8. Popping R. On agreement indices for nominal data. In: Saris WE, Gallhofer IN, editors. Sociometric research: Volume I, data collection and scaling [Internet]. 1st ed. New York, NY: St. Martin’s / Springer; 1988 [cited 2022 Jan 15]. p. 90–105. Available from: http://link.springer.com/chapter/10.1007/978-1-349-19051-5_6

9. Riffe D, Lacy S, Fico FG. Analyzing Media Messages: Using Quantitative Content Analysis in Research [Internet]. 2nd ed. Mahwah, New Jersey and London, New Jersey and London: Lawrence Erlbaum Associates, Publishers; 2005 [cited 2022 Jan 15]. Available from: https://books.google.com.hk/books?hl=en&lr=&id=enCRAgAAQBAJ&oi=fnd&pg=PP1&ots=B00EbKHtj7&sig=e_EdXbsENFS9VfNJR62OrQ00_MM&redir_esc=y#v=onepage&q&f=false

10. Zwick R. Another look at interrater agreement. Psychol Bull [Internet]. 1988;103(3):374–8. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0024005773&partnerID=tZOtx3y1

11. Benini R. Principii di Demongraphia: Manuali Barbera Di Scienze Giuridiche Sociali e Politiche (No. 29)[Principles of demographics (Barbera Manuals of Jurisprudence and Social Policy)]. Firenze, Italy: G. Barbera; 1901.

12. Hayes AF, Krippendorff KH. Answering the call for a standard reliability measure for coding data. Commun Methods Meas [Internet]. 2007 [cited 2022 Jan 15];1(1):77–89. Available from: http://www.tandfonline.com/doi/abs/10.1080/19312450709336664

13. Hughes MA, Garrett DE. Intercoder reliability estimation approaches in marketing: A generalizability theory framework for quantitative data. J Mark Res [Internet]. 1990 [cited 2022 Jan 15];27(2):185–95. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=9602260627&site=ehost-live

14. Zhao X, Liu JS, Deng K. Assumptions behind intercoder reliability indices. Ann Int Commun Assoc [Internet]. 2013;36(1):419–80. Available from: http://www.tandfonline.com/doi/abs/10.1080/23808985.2013.11679142?journalCode=rica20

15. Bennett EM, Alpert R, Goldstein AC. Communications through limited response questioning. Public Opin Q [Internet]. 1954 [cited 2022 Jan 15];18:303–8. Available from:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2189948

16. Scott WA. Reliability of content analysis: The case of nominal coding. Public Opin Q [Internet]. 1955 [cited 2022 Jan 15];19(3):321–325. Available from: http://www.jstor.org/stable/2746450

17. Krippendorff KH. Estimating the Reliability, systematic error and radnom error of interval data. Educ Psychol Meas [Internet]. 1970 [cited 2022 Jan 15];30:61–70. Available from: http://epm.sagepub.com/content/30/1/61.short

18. Krippendorff KH. On generating data in communication research. J Commun [Internet]. 1970 [cited 2022 Jan 15];20:241–69. Available from: http://dx.doi.org/10.1111/j.1460-2466.1970.tb00883.x

19. Krippendorff KH. Content Analysis: An Introduction to its Methodology. Thousand Oaks, CA: Sage; 1980.

20. Button CM, Snook B, Grant MJ. Inter-rater agreement, data reliability, and the Crisis of confidence in psychological research. Quant Methods Psychol. 2020;16(5):467–71.

21. Checco A, Roitero A, Maddalena E, Mizzaro S, Demartini G. Let's agree to disagree: Fixing agreement measures for crowdsourcing. Proc Fifth AAAI Conf Hum Comput Crowdsourcing [Internet]. 2017 [cited 2022 Jan 15];(Hcomp):11–20. Available from: www.aaai.org

22. ten Hove D, Jorgensen TD, van der Ark LA. On the usefulness of interrater reliability coefficients. In: Wiberg M, Culpepper S, Janssen R, Gonzalez J, Molenaar D, editors. Quantitative Psychology: The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017. Cham, Switzerland: Springer; 2018. p. 67–75.

23. Zhao X. When to use Cohen's κ, if ever? [Internet]. [Boston, USA, May, https://repository.hkbu.edu.hk/coms_conf/2/]: Paper presented at the 61st annual conference of International Communication Association.; 2011 [cited 2022 Jan 15]. Available from: https://repository.um.edu.mo/handle/10692/102423

24. Zhao X. When to use Scott's π or Krippendorff's α, if ever? [Internet]. [St. Louis, USA, August, https://repository.hkbu.edu.hk/coms_conf/3/]: Paper presented at the annual conference of Association for Education in Journalism and Mass Communication; 2011 [cited 2022 Jan 15]. Available from: https://repository.um.edu.mo/handle/10692/102434

25. Zhao X, Deng K, Feng GC, Zhu L, Chan VKC. Liberal-conservative hierarchies for indices of inter-coder reliability [Internet]. Paper presented at the 62nd annual conference of International Communication Association, Phoenix, Arizona, USA, May; 2012 [cited 2022 Jan 15]. Available from: https://repository.um.edu.mo/handle/10692/102423

26. Zhao X, Feng GC, Liu JS, Deng K. We agreed to measure agreement - Redefining reliability de-justifies Krippendorff's alpha. China Media Res [Internet]. 2018 [cited 2022 Jan 15];14(2):1. Available from: https://repository.um.edu.mo/handle/10692/25978

27. Conger AJ. Kappa and Rater Accuracy: Paradigms and Parameters. Educ Psychol Meas [Internet]. 2016 [cited 2022 Jan 15];0013164416663277. Available from: http://epm.sagepub.com/content/early/2016/08/18/0013164416663277.abstract%255Cnhttp://epm.sagepub.com/content/early/2016/08/18/0013164416663277%255Cnhttp://epm.sagepub.com/content/early/2016/08/18/0013164416663277.full.pdf

28. Delgado R, Tibau XA. Why Cohen's Kappa should be avoided as performance measure in classification. PLoS One [Internet]. 2019 [cited 2022 Jan 15];14(9):1–26. Available from: http://dx.doi.org/10.1371/journal.pone.0222916

29. Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of psychiatric diagnosis: Theory and practice. Arch Gen Psychiatry. 1981;38(4):408–13.

30. Riffe D, Lacy S, Fico FG, Watson B. Analyzing media messages: Using quantitative content analysis in research (4th ed.). [Internet]. 4th ed. New York: Routledge; 2019 [cited 2022 Jan 15]. Available from: https://www.taylorfrancis.com/books/9780429464287

31. Zhao X. Selective spiral ——A mega, meta, predictive and presumptive theory of communication. Chinese J Journal Commun [Internet]. 2018 [cited 2022 Jan 15];40(2):140–53. Available from: http://cjjc.ruc.edu.cn/EN/Y2018/V40/I2/140

32. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Stat [Internet]. 1979 [cited 2022 Jan 15];7(1):1–26. Available from: http://projecteuclid.org/euclid.aos/1176344552%5Cnhttps://projecteuclid.org/euclid.aos/1176344552

33. Efron B, Robert J. Tibshirani. An Introduction to the Bootstrap [Internet]. New York and London: Chapman & Hall; 1993 [cited 2022 Jan 15]. 257 p. Available from: http://books.google.com/books?id=gLlpIUxRntoC&pgis=1

34. Shao J, Tu D. The Jackknife and Bootstrap [Internet]. Springer Series in Statistics. New York: Springer Science & Business Media; 1995 [cited 2022 Jan 15]. 516 p. Available from: http://www.loc.gov/catdir/enhancements/fy0815/95015074-d.html

35. Liu JS. Monte Carlo strategies in scientific computing. New York: Springer; 2001.

36. Montgomery DC. Design and Analysis of Experiments, 7th Edition. John Wiley & Sons. 2009.

37. Claassen JAHR. The gold standard: not a golden standard. BMJ [Internet]. 2005 [cited 2022 Jan 15];330(7500):1121. Available from: bmj.com

38. Riffe D, Lacy S, Fico FG. Analyzing Media Messages: Using Quantitative Content Analysis in Research. Mahwah, N J: Lawrence Erlbaum Associates; 1998.

39. Riffe D, Lacy S, Fico FG. Analyzing Media Messages : Using Quantitative Content Analysis in Research. 3rd ed. New York: Routledge; 2014.

40. Wasserstein RL, Lazar NA. The ASA’s statement on p -Values: context, process, and purpose. Am Stat [Internet]. 2016 [cited 2022 Jan 15];70(2):129–33. Available from: https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

41. Amrhein V, Greenland S, McShane B, others. Retire statistical significance. Nature. 2019 Mar 21;567:305–7.

42. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “$p < 0.05$.” Am Stat. 2019;73(sup1):1–19.

43. Wilkinson L, APA Task Force. Statistical methods in psychology journals: Guidelines and explanations (Report by Task Force on Statistical Inference, APA Board of Scientific Affairs). Am Psychol [Internet]. 1999 [cited 2022 Jan 15];54(8):594–604. Available from: http://psycnet.apa.org/journals/amp/54/8/594/

44. Lazar N. Time to say goodbye to “statistically significant” and embrace uncertainty, say statisticians. Retraction Watch [Internet]. 2019 Mar 21 [cited 2022 Jan 15]; Available from: https://retractionwatch.com/2019/03/21/time-to-say-goodbye-to-statistically-significant-and-embrace-uncertainty-say-statisticians/

45. Liu PL, Zhao X, Wan B. COVID-19 information exposure and vaccine hesitancy: The influence of trust in government and vaccine confidence. Psychol Heal Med [Internet]. 2021 [cited 2022 Jan 15];00(00):1–10. Available from: https://doi.org/10.1080/13548506.2021.2014910

46. Zhao X. Four functions of statistical significance tests [Internet]. Presentation at the School of Statistics and Center for Data Sciences Beijing Normal University, 25th December.; 2016 [cited 2022 Jan 15]. Available from: https://repository.um.edu.mo/handle/10692/95184

47. Zhao X, Ye J, Sun S, Zhen Y, Zhang Z, Xiao Q, et al. Best Title Lengths of Online Postings for Highest Read and Relay. Journal Commun Rev [Internet]. 2022 [cited 2022 Jul 21];75(3):5–20. Available from: https://repository.um.edu.mo/handle/10692/95320

48. Hsu LM, Field R. Interrater agreement measures: Comments on Kappan, Cohen's Kappa, Scott's π, and Aickin's α. Underst Stat. 2003;2(3):205–19.

49. Krippendorff KH. Reliability in content analysis: Some common misconceptions and recommendations. Hum Commun Res. 2004;30(3):411–33.

50. Krippendorff KH. A dissenting view on so-called paradoxes of reliability coefficients. Ann Int Commun Assoc [Internet]. 2013 [cited 2022 Jan 15];36(1):481–99. Available from: http://www.tandfonline.com/doi/pdf/10.1080/23808985.2013.11679143

51. Lombard M, Snyder-Duch J, Bracken CC. Content analysis in mass communication: Assessment and reporting of intercoder reliability. Hum Commun Res [Internet]. 2002 [cited 2022 Jan 15];28(4):587–604. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2958.2002.tb00826.x/abstract

52. Feng GC, Zhao X. Do not force agreement – A response to Krippendorff. Methodol Eur J Res Methods Behav Soc Sci [Internet]. 2016 [cited 2022 Jan 15];12(4):145–8. Available from: https://repository.um.edu.mo/handle/10692/26008

53. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol [Internet]. 2008 [cited 2022 Jan 15];61(1):29–48. Available from: http://doi.wiley.com/10.1348/000711006X126600

54. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. 2nd ed. Gaithersburg, MD: STATAXIS Publishing Company; 2010. 197 p.

55. Gwet KL. Variance estimation of nominal-scale inter-rater reliability with random selection of raters. Psychometrika [Internet]. 2008 [cited 2016 Feb 7];73(3):407–30. Available from: http://link.springer.com/article/10.1007/s11336-007-9054-8

56. Feng GC. Indexing versus Modeling Intercoder Reliability. Hong Kong Baptist University; 2013.

57. Feng GC. Factors affecting intercoder reliability: A Monte Carlo experiment. Qual Quant [Internet]. 2013 [cited 2022 Jan 15];47(5):2959–82. Available from: http://link.springer.com/article/10.1007/s11135-012-9745-9

58. Feng GC. Underlying determinants driving agreement among coders. Qual Quant. 2013;47(5):2983–97.

59. Attneave F. Psychological probability as a function of experienced frequency. J Exp Psychol. 1953;46(2):81–6.

60. Fischhoff B, Slovic P, Lichtenstein S. Knowing with certainty: The appropriateness of extreme confidence. J Exp Psychol Hum Percept Perform [Internet]. 1977 [cited 2022 Jan 15];3(4):552–64. Available from: http://content.apa.org/journals/xhp/3/4/552

61. Brennan RL, Prediger DJ. Coefficient kappa: Some uses, misuses, and alternatives. Educ Psychol Meas [Internet]. 1981 [cited 2022 Jan 15];41(3):687–99. Available from: http://journals.sagepub.com/doi/10.1177/001316448104100307

62. Feinstein AR, Cicchetti D V. High agreement but low Kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6):551–8.

63. Feinstein AR, Cicchetti D V. High agreement but low Kappa: I. the problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9.

64. Lantz CA, Nebenzahl E. Behavior and interpretation of the Kappa statistic: Resolution of the two paradoxes. J Clin Epidemiol. 1996;49(4):431–4.

65. Spitznagel EL, Helzer JE, John E. Helzer., Helzer JE. A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry [Internet]. 1985 [cited 2022 Jan 15];42(7):725–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4015315

66. Cousineau D, Laurencelle L. An unbiased estimate of global interrater agreement. Educ Psychol Meas [Internet]. 2016 [cited 2022 Jan 15];0013164416654740. Available from: http://journals.sagepub.com/doi/abs/10.1177/0013164416654740

67. Krippendorff KH. Estimating the reliability, systematic error and random error of interval data. Educ Psychol Meas [Internet]. 1970 [cited 2022 Jan 15];30(1):61–70. Available from: http://epm.sagepub.com/cgi/doi/10.1177/001316447003000105

68. Cousineau D, Laurencelle L. A ratio test of interrater agreement with high specificity. Educ Psychol Meas [Internet]. 2015 [cited 2022 Jan 15];75(6):979–1001. Available from: http://epm.sagepub.com/content/75/6/979.abstract?&location1=all&location2=all&row_operator2=and&term1a=simulation&term_operator1=and&term_operator2=and&ct

69. Kirilenko AP, Stepchenkova S. Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa. PLoS One [Internet]. 2016 [cited 2022 Jan 15];11(3):e0149787. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149787

70. Meehl PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis, MN: University of Minnesota Press; 1954.

71. Dawes RM, Faust D, Meehl. PE. Clinical Versus Actuarial Judgment. Science (80-). 1989;243(4899):1668-1674.

72. Chaffee SH. Communication Concept: Explication. Newbury Park, CA: Sage Publications, Inc.; 1991. 96 p.

73. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke. 1988;19(5):604–7.

74. Uebersax JS. The Myth of Chance Corrected Agreement [Internet]. 2009 [cited 2012 Oct 18]. Available from: http://www.john-uebersax.com/stat/kappa2.htm

75. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. Psychometrika [Internet]. 1979 [cited 2022 Jan 15];44(4):461–72. Available from: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0001586045&partnerID=40

76. Shrout PE, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. Arch Gen Psychiatry. 1987;44(2):172–7.

77. von Eye A, von Eye M. On the marginal dependency of Cohen's κ. Eur Psychol. 2008;13(4):305–15.

78. Goodman LA, Kruskal WH. Measures of association for cross classifications. J Am Stat Assoc. 1954 Dec;49(268):732–64.

79. Guttman L. The test-retest reliability of qualitative data. Psychometrika. 1946;11(2):81–95.

80. Janson S, Vegelius J. On generalizations of the G index and the Phi coefficient to nominal scales. Multivariate Behav Res. 1979;14(2):255–69.

81. Guilford JP. Preparation of item scores for correlation between individuals in a Q factor analysis. Paper Presented at the Annual Convention of the Society of Multivariate Experimental Psychologists; 1961.

82. Holley JW, Guilford JP. A note on the G-index of agreement. Educ Psychol Meas. 1964;24(4):749–53.

83. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol [Internet]. 1993 [cited 2022 Jan 15];46(5):423–9. Available from: http://www.sciencedirect.com/science/article/pii/089543569390018V

84. Maxwell AE. Coefficients of agreement between observers and their interpretation. Br J Psychiatry. 1977;130(1):79–83.

85. Potter WJ, Levine-Donnerstein D. Rethinking validity and reliability in content analysis. J Appl Commun Res [Internet]. 1999 [cited 2022 Jan 15];27(3):258–84. Available from: http://www.tandfonline.com/doi/abs/10.1080/00909889909365539

86. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters [Internet]. 3rd ed. Gaithersburg, MD, USA: Advanced Analytics, LLC; 2012 [cited 2022 Jan 15]. 197 p. Available from: https://books.google.com/books?hl=en&lr=&id=fac9BQAAQBAJ&oi=fnd&pg=PP1&dq=Gwet+K+L&ots=UUdriDAp0a&sig=mKjbb_IW1eNG474Cb0Omp3n5BMk

87. Ji MF, McNeal JU. How chinese children's commercials differ from those of the united states: A content analysis. J Advert [Internet]. 2001 [cited 2022 Jan 15];30(3):79–92. Available from: http://web.ebscohost.com/ehost/detail?hid=106&sid=c0d4783a-f726-4eea-9dc2-4022b157f163@sessionmgr112&vid=3&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ==#db=buh&AN=5507388

88. Kolbe RH, Burnett MS. Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. J Consum Res. 1991;18(2):243–50.

89. Okazaki S, Rivas JA. A content analysis of multinationals' Web communication strategies: cross-cultural research framework and pre-testing. Internet Res. 2002;12(5):380–90.

90. Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. Psychol Bull. 1987;101(1):140–6.

91. Andsager JL, Schwartz J. Explicating time: toward making content analysis research describing time frames more meaningful. [Chicago]: Paper Presented at Annual Conference of Association for Education in Journalism and Mass Communication, Chicago; 2012.

92. Bakeman R. Behavioral observation and coding. Handb Res methods Soc Personal Psychol [Internet]. 2000 [cited 2022 Jan 15];138–59. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2000-07611-006&site=ehost-live

93. von Eye A. An alternative to Cohen's kappa. Vol. 11, European Psychologist. 2006. p. 12–24.

94. Warrens MJ. On marginal dependencies of the 2 × 2 kappa. Adv Stat [Internet]. 2014 [cited 2022 Jan 15];2014:1–6. Available from: http://www.hindawi.com/archive/2014/759527/

95. Aickin M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. Biometrics [Internet]. 1990 [cited 2022 Jan 15];46:293–302. Available from: http://www.jstor.org/stable/2531434

96. Gwet KL. Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity [Internet]. Gaithersburg, MD, USA; 2002 [cited 2022 Jan 15]. Available from: http://hbanaszak.mjr.uw.edu.pl/TempTxt/smirra2.pdf

97. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters [Internet]. 4th ed. Gaithersburg, MD: Advanced Analytics, LLC; 2014 [cited 2022 Jan 15]. 429 p. Available from: https://books.google.com/books?hl=en&lr=&id=fac9BQAAQBAJ&oi=fnd&pg=PP1&dq=Gwet+K+L&ots=UUdriDAp0a&sig=mKjbb_IW1eNG474Cb0Omp3n5BMk

98. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol [Internet]. 2013 [cited 2022 Jan 15];13:61. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3643869&tool=pmcentrez&rendertype=abstract

99. Powers DWM. The problem with Kappa. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; 2012. p. 345–55.

100. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data - Recommendations for the use of performance metrics. Proc - 2013 Hum Assoc Conf Affect Comput Intell Interact ACII 2013. 2013;245–51.

101. Bloch DA, Kraemer HC. 2 x 2 Kappa coefficients: Measures of agreement or association. Biometrics. 1989;45(1):269–87.

102. Dewey ME. Coefficients of agreement. Br J Psychiatry. 1983;143(5):487–9.

103. Feuerman M, Miller AR. Relationships between statistical measures of agreement: sensitivity, specificity and kappa. J Eval Clin Pract [Internet]. 2008 [cited 2022 Jan 15];14(5):930–3. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2753.2008.00984.x/full

104. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: An appraisal of a reappraisal. J Clin Epidemiol. 1988;41(10):959–68.

105. Kraemer HC, Periyakoil VS, Noda A. Tutorial in Biostatistics: Kappa coefficients in medical research. Stat Med [Internet]. 2002 [cited 2022 Jan 15];21(14):2109–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12111890

106. Roberts C. Modelling patterns of agreement for nominal scales. Stat Med. 2008;27(6):810–30.

107. Williamson JM, Lipsitz SR, Amita K. Manatunga. Modeling kappa for measuring dependent categorical agreement data. Biostatistics [Internet]. 2000 [cited 2022 Jan 15];1(2):191–202. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12933519

108. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. Vol. 58, Journal of Clinical Epidemiology. 2005. p. 655–61.

109. Rogot E, Irving D. Goldberg. A proposed index for measuring agreement in test-retest studies. J Chronic Dis. 1966;19(9):991–1006.

110. Siegel S, Castellan JNJ. Nonparametric statistics for the behavioural sciences [Internet]. 2nd ed. MacGraw Hill; 1988 [cited 2022 Jan 15]. 213–214 p. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Non+parametric+statistics+for+the+behavioural+sciences#9

111. Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of psychiatric diagnosis: Theory and practice. Arch Gen Psychiatry. 1981;38(4):408–13.

112. Krippendorff KH. Misunderstanding reliability. Methodology [Internet]. 2016 [cited 2022 Jan 15];12(4):139–44. Available from: http://econtent.hogrefe.com/doi/full/10.1027/1614-2241/a000119

113. Krippendorff KH. The changing landscape of content analysis: Reflections on social construction of reality and beyond. So CYK, editor. Commun Soc [Internet]. 2019 [cited 2022 Jan 15];47(47):1–27. Available from: https://repository.upenn.edu/asc_papers/604

114. Krippendorff KH. Reliability in content analysis. Hum Commun Res [Internet]. 2004 [cited 2022 Jan 15];30(3):411–33. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2958.2004.tb00738.x/abstract

115. Rudd P. In Search of the Gold Standard for Compliance Measurement. Arch Intern Med [Internet]. 1979 Jun 1 [cited 2022 Jan 15];139(6):627–8. Available from: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/589465

116. Webb PH, Michael L. Ray. Effects of TV Clutter. J Advert Res. 1979;19(3):7–12.

117. Zhao X. Clutter and serial order redefined and retested. J Advert Res [Internet]. 1997 [cited 2022 Jan 15];37(5):57–73. Available from: https://works.bepress.com/xinshu-zhao/11/

118. Jeong Y, Tran H, Zhao X. How much is too much? J Advert Res. 2012;52(1):87–101.

119. Li C. Primacy effect or recency effect? A long-term memory test of super bowl commercials. J Consum Behav. 2010;9(1):32–44.

120. Pieters RGM, Bijmolt THA. Consumer Memory for Television Advertising: A Field Study of Duration, Serial Position, and Competition Effects. J Consum Res. 1997;23(4):362.

121. Terry WS. Serial position effects in recall of television commercials. J Gen Psychol [Internet]. 2005 [cited 2022 Jan 15];132(2):151–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15871298

122. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol. 2000;53(5):499–503.

123. Cicchetti D V., Shoinralter D, Peter J. Tyrer. The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. Appl Psychol Meas. 1985;9(1):31–6.

124. Achen CH, Shively WP. Cross-Level Inference. In: Cross-Level Inference. 1995. p. 1–29.

125. Heerink N, Mulatu A, Bulte E, Mulatu A. Income inequality and the environment: Aggregation bias in environmental Kuznets curves. Ecol Econ. 2001;38(3):359–67.

126. Kinsman RA, Staudenmayer H. Baseline levels in muscle relaxation training. Appl Psychophysiol Biofeedback. 1978;3(1):97–104.

127. Groot W. Adaptation and scale of reference bias in self-assessments of quality of life. J Health Econ. 2000;19(3):403–20.

128. Cohen J. Statistical Power Analysis for the Behavioral Sciences [Internet]. 2nd ed. Vol. 2nd. Hillsdale, New Jersey: Erihaum; 1988 [cited 2022 Jan 15]. 567 p. Available from: http://books.google.com/books?id=Tl0N2lRAO9oC&pgis=1

129. Cohen J. A power primer. Psychol Bull [Internet]. 1992 [cited 2022 Jan 15];112(1):155–9. Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.112.1.155

130. Stigler SM. Francis Galton’s Account of the Invention of Correlation. Stat Sci. 1989;4(2):73–9.

131. Steel RG, James H. Torrie. Principles and Procedures of Statistics. New York: McGraw-Hill; 1960.

132. Krippendorff KH. Content Analysis: An Introduction to its Methodology. 3rd ed. Thousand Oaks, CA: Sage Publications; 2012.

133. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol [Internet]. 2008 [cited 2022 Jan 15];61(1):29–48. Available from: http://onlinelibrary.wiley.com/doi/10.1348/000711006X126600/full

134. Krippendorff KH. Bivariate agreement coefficients for reliability of data. Sociol Methodol [Internet]. 1970 [cited 2022 Jan 15];2:139–50. Available from: http://www.jstor.org/stable/270787

Tables and Figures

Table 1 A Category (C) by Difficulty (d_f) by Skew (s_k) - Reconstructed Experiment *

Across: *Distribution & Skew* (s_k)			50&50 s_k=0.5					25&75, 75&25 s_k=0.75					1&99, 99&1 s_k=0.99			
Across: *Category (C)*			2	4	6	8		2	4	6	8		2	4	6	8
difference in pixels (p_x)	*Difficulty* d_f=(8-p_x)/7															
1	=1.000		4	4	4	4		4	4	4	4		4	4	4	4
2	≈0.8571		4	4	4	4		4	4	4	4		4	4	4	4
3	≈0.7143		4	4	4	4		4	4	4	4		4	4	4	4
4	≈0.5714		4	4	4	4		4	4	4	4		4	4	4	4
5	≈0.4286		4	4	4	4		4	4	4	4		4	4	4	4
6	≈0.2857		4	4	4	4		4	4	4	4		4	4	4	4
7	≈01429		4	4	4	4		4	4	4	4		4	4	4	4
8	=0.0000		4	4	4	4		4	4	4	4		4	4	4	4

* Main cell entries are number of reconstructed rating sessions (subjects) in each experimental condition (cell).

Table 2 Concepts and Variables

		Down: Author or Origin	*Reliability (True Agreement)*	*Chance Agreement*
		generic for any index	r_i	a_c
Dependent Variables	Index Estimation	*%-Agreement* (unknown author)	a_o	ao_{ac}
		Bennett et al (1954)(15)	S	S_{ac}
		Perreault & Leigh (1989) (7)	I_r	Ir_{ac}
		Gwet (2002, 2008, 2010, 2012)(54),(96,133),(86)	AC_1	AC_{ac}
		Scott (1955) (16)	π	π_{ac}
		Cohen (1960) (2)	κ	κ_{ac}
		Krippendorff (1970, 1980)(19,67,134)	α	α_{ac}
	Empirical Observation	Primary Indicator	o_{ri} observed interrater reliability	o_{ac} observed chance agreement
		Secondary Indicator (used in calculation)	o_{ar} observed right agreement	o_{ae} observed erroneous agreement
			a_o observed agreement	d_o observed disagreement
Independent Variables	Denotation	C	s_k	d_f or e_s
	Concept	*Category*	*Distribution Skew*	*Difficulty* or *Easiness*

Other Concepts	Denotation	e_m	m_e	s_{dm}	dr^2	N_c	N_d
	Concept	error of means (mean estimation minus mean target)	mean of errors (mean of differences between estimation and target)	standard deviation of an observed target of estimation (o_{ae} o_{ri})	directional r^2 ($dr^2 = r*\|r\|$)	No. of rating sessions	No. of rating decisions within a session

Table 3 Effects of Estimation Targets, Category, Skew & Difficulty on Observed or Estimated Chance Agreement and Reliability (dr^2)

			A.	B.	C.	D.	E.	F.	G.	H.
	1	Right: Source or Author	Obser-vation	%-agreement	*Bennett et al.*	*Perreault & Leigh*	*Gwet*	*Scott*	*Cohen*	*Krippen-dorff*
Effects on Intcdr Reliability Obsv & Ests	2	Right: Obsd / Estd Interrater Reliability as Dependent Variables Down: Independent Variables	o_{ri}	a_o	S	I_r	AC_1	π	κ	α
	3	Observed Reliability (o_{ri})	1.00***	.841***	.691***	.599***	.721***	.312***	.312***	.312***
	4	Category (C)	.003	-.002	.175***	.185***	.123***	.001	.001	.001
	5	Distribution Skew (s_k)	.000	.000	.000	-.000	.003	-.293***	-.292***	-.293***
	6	Difficulty (d_f)	-.774***	-.778***	-.566***	-.434***	-.554***	-.389***	-.389***	-.389***
Effects on Chance Agrt Obsv & Ests	7	Right: Obsd / Estd. Chance Agreement as Dependent Variables Down: Independent Variables	o_{ac}	ao_{ac}=0†	S_{ac}	Ir_{ac}	AC_{ac}	π_{ac}	κ_{ac}	α_{ac}
	8	Observed Chance Agreement (o_{ac})	1.00***	---	.021**	.021**	.075***	-.151***	-.152***	-.151***
	9	Category (C)	-.019**	---	-.863***	-.863***	-.661***	-.013*	-.014*	-.013*
	10	Distribution Skew (s_k)	-.001	---	.000	.000	-.039***	.437***	.434***	.437***
	11	Difficulty (d_f)	.585***	---	.000	.000	.009	-.123***	-.125***	-.123***
N	12	N_c (number of rating sessions)	384	384	384	384	384	384	384	384
	13	N_d (number items within each session)	100	100	100	100	100	100	100	100

Main cell entries are directional r squared (dr^2), which are r squared with the directional sign of r, $dr^2=r\bullet|r|$.
*: $p<.05$; **: $p<.01$; ***: $p<.001$.† As ao_{ac}, the chance estimate of a_o, is a constant, its correlations (dr^2) with other variables cannot be calculated.

Table 4 Mean of Errors (m_e) / Distance Between Index Estimations and Targets of Estimation

			A.	B.	C.	D.	E.	F.	G.
	1	Author or Source	*%-agreement*	*Bennett et al.*	*Perreault & Leigh*	*Gwet*	*Scott*	*Cohen*	*Krippen-dorff*
Interrater Reliability	2	Interrater Reliability Estimator	a_o	S	I_r	AC_1	π	κ	α
	3	m_e (r_i)=mean ($\|r_i$-$o_{ri}\|$) (0≤m_e≤1)	.130***	.096***	.180***	.093***	.327***	.324***	.323***
	4	*Standard Deviation of* m_e (r_i)	.145	.099	.148	.104	.221	.220	.220
	5	95% confidence interval of m_e (r_i)	.115~.144	.086~.106	.164~.194	.082~.103	.304~.349	.302~.346	.301~.345
Chance Agreement	6	Chance Agreement Estimator	ao_{ac}	S_{ac}	Ir_{ac}	AC_{ac}	π_{ac}	κ_{ac}	α_{ac}
	7	m_e (a_c):=mean ($\|a_c$-$o_{ac}\|$) (0≤m_e≤1)	.130***	.182***	.182***	.130***	.450***	.448***	.448***
	8	*Standard Deviation of* m_e (a_c)	.145	.141	.141	.127	.201	.201	.202
	9	95% confidence interval of m_e (a_c)	.115~.144	.168~.196	.168~.196	.117~.143	.429~.470	.428~.469	.427~.468
N	10	N_c (number of rating sessions)	384	384	384	384	384	384	384
	11	N_d (number items within each session)	100	100	100	100	100	100	100

*: *p*<.05, **: *p*<.01, ***: *p*<.001

Table 5 Means and Error of Means (e_m): Index Estimations Against Observations

			A.	B.	C.	D.	E.	F.	G.	H.
	1	Right: Author or Source	*Observed Agreement*	*%-agreement*	*Bennett et al.*	*Perreault & Leigh*	*Gwet*	*Scott*	*Cohen*	*Krippendorff*
Interrater Reliability	2	Observed or Estimated Reliability (denotation)	o_{ri}	a_o	S	I_r	AC_1	π	κ	α
	3	Observed / Estimated Interrater Reliability	.555	.685	.556	.726	.600	.237	.240	.241
	4	Standard Deviation	.248	.122	.203	.173	.192	.249	.247	.248
	5	Range (minimum~maximum)	-.20~.90	.42~.92	-.10~.856	.0~.925	-.045~.912	-.177~.778	-.173~.778	-.17~.779
	6	$e_m(r_i)$=mean(r_i)-mean(o_{ri}) $(-1 \leq e_m \leq 1)$	.000	$.130^{***}$	.001	$.171^{***}$	$.044^{***}$	$-.318^{***}$	$-.315^{***}$	$-.314^{***}$
	7	95% confidence interval	.00~.00	.115~.144	-.013~.015	.155~.186	.031~.058	-.341~-.295	-.338~-.292	-.338~-.291
Chance Agreement	8	Chance Agreement (denotation)	o_{ac}	ao_{ac}	S_{ac}	Ir_{ac}	AC_{ac}	π_{ac}	κ_{ac}	α_{ac}
	9	Observed or Estimated Chance Agreement	.130	.000	.260	.260	.173	.575	.573	.572
	10	Standard Deviation	.145	.000	.146	.146	.148	.109	.109	.110
	11	Range (minimum~maximum)	.0~.72	.0~.0	.125~.50	.125~.50	.022~.50	.448~.905	.447~.905	.445~.905
	12	$e_m(a_c)$=mean(a_c)-mean(o_{ac}) $(-1 \leq e_m \leq 1)$	.000	$-.130^{***}$	$.131^{***}$	$.131^{***}$	$.044^{***}$	$.445^{***}$	$.443^{***}$	$.443^{***}$
	13	95% confidence interval	.00~.00	-.144~-.115	.111~.15	.111~.15	.026~.061	.423~.466	.422~.465	.421~.464
N	14	N_c (number of rating sessions)	38	384	384	384	384	384	384	384
	15	N_d (number items within each session)	100	100	100	100	100	100	100	100

*: p<.05, **: p<.01, ***: p<.001

Table 6 Effects of Category, Skew, and Difficulty on Observed Chance Agreement, Reliability, and Index Estimations (Average Scores)

			A.	B.	C.	D.	E.	F.	G.	H.	I	J	K	L	M	N	O	P	Q
			Reliability Observation or Estimation								Chance Agreement Observation or Estimation								
1	Author/ Source		Observed	*%-Agreement*	*Bennett et al.*	*Perreault & Leigh*	*Gwet*	*Scott*	*Cohen*	*Krippen-dorff*	Observed	*%-Agreement*	*Bennett et al.*	*Perreault & Leigh*	*Gwet*	*Scott*	*Cohen*	*Krippen-dorff*	
2	Estimator:		o_{ri}	a_o	S	I_r	AC_1	π	κ	α	o_{ac}	ao_{ac}	S_{ac}	Ir_{ac}	AC_{ac}	π_{ac}	κ_{ac}	α_{ac}	N_c
3	Ground 0		**.555**	.685	.370	.608	.371	.369	370	.373	**.130**	0	.500	.500	.499	.501	.500	.498	32
4	Category (C)	2	**.537**	.701	.402	.584	.470	.230	.232	.234	**.164**	0	.500	.500	.401	.598	.597	.596	96
5		4	**.550**	.678	.571	.747	.621	.226	.230	.230	**.128**	0	.250	.250	.142	.573	.571	.571	96
6		6	**.557**	.676	.612	.777	.644	.239	.241	.242	**.119**	0	.167	.167	.087	.562	.561	.561	96
7		8	**.578**	.686	.641	.796	.664	.254	.257	.257	**.108**	0	.125	.125	.062	.564	.563	.562	96
8	Skew (sk)	.50	**.550**	.688	.560	.732	.592	.370	.372	.374	**.138**	0	.260	.260	.203	.501	.500	.498	128
9		.75	**.556**	.678	.547	.722	.588	.302	.304	.305	**.122**	0	.260	.260	.186	.545	.543	.543	128
10		.99	**.560**	.690	.561	.723	.619	.040	.044	.045	**.130**	0	.260	.260	.132	.678	.676	.676	128
11	Difficulty (d_f)	.000	**.824**	.844	.782	.884	.810	.482	.484	.485	**.020**	0	.260	.260	.152	.630	.629	.628	48
12		.143	**.783**	.805	.728	.852	.761	.404	.406	.407	**.021**	0	.260	.260	.158	.616	.615	.615	48
13		.286	**.721**	.757	.659	.808	.697	.341	.343	.344	**.036**	0	.260	.260	.164	.599	.598	.600	48
14		.429	**.659**	.721	.600	.765	.643	.273	.275	.277	**.062**	0	.260	.260	.169	.591	.589	.588	48
15		.571	**.543**	.659	.518	.706	.563	.196	.199	.200	**.116**	0	.260	.260	.180	.565	.563	.563	48
16		.714	**.439**	.606	.444	.647	.495	.117	.121	.121	**.168**	0	.260	.260	.182	.548	.546	.546	48
17		.857	**.331**	.567	.387	.591	.440	.068	.071	.072	**.236**	0	.260	.260	.189	.534	.533	.532	48
18		1.00	**.142**	.523	.332	.552	.389	.018	.022	.022	**.380**	0	.260	.260	.194	.514	.512	.511	48
19		Mean	**.555**	.685	.556	.726	.600	.237	.240	.241	**.130**	0	.260	.260	.173	.575	.573	.572	384
20		N_d	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

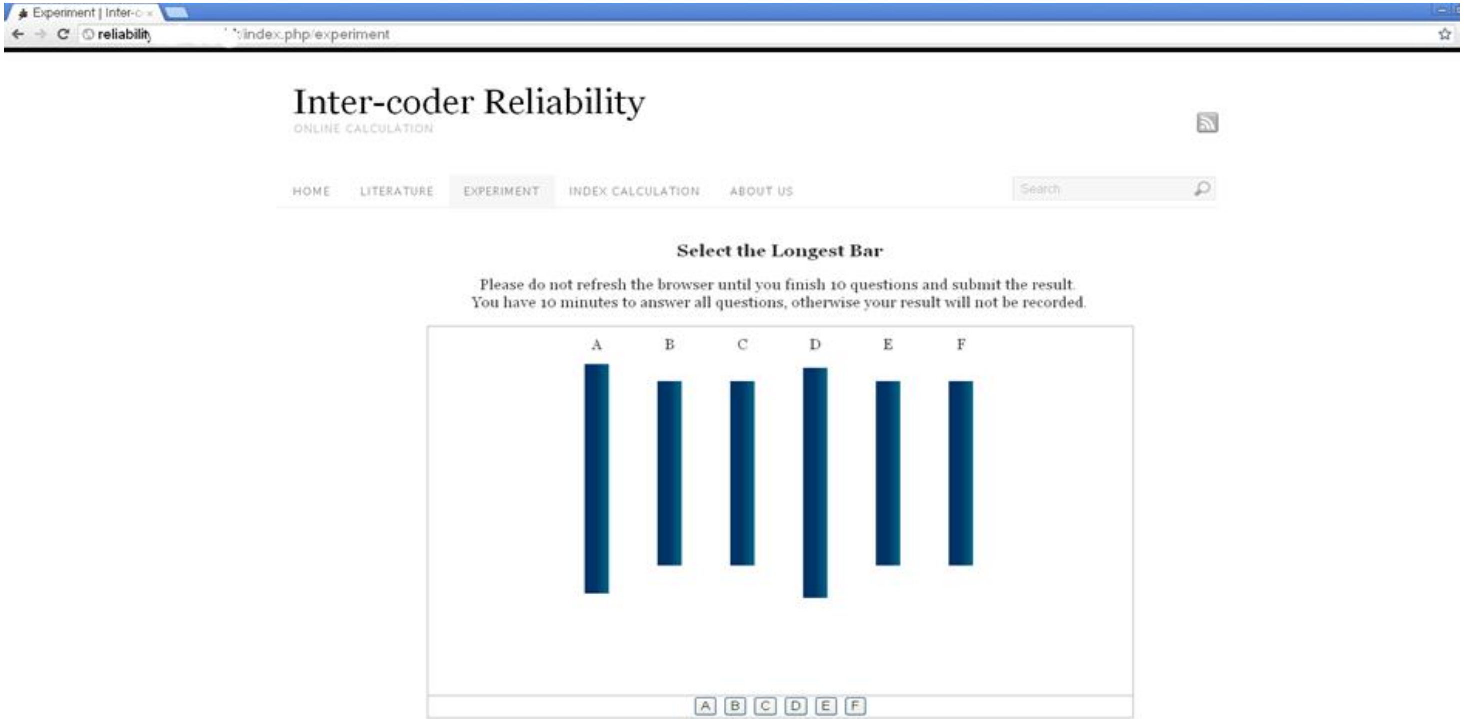

Figure 1 A sample screen seen by some raters (for category = 6, difficulty = 1).

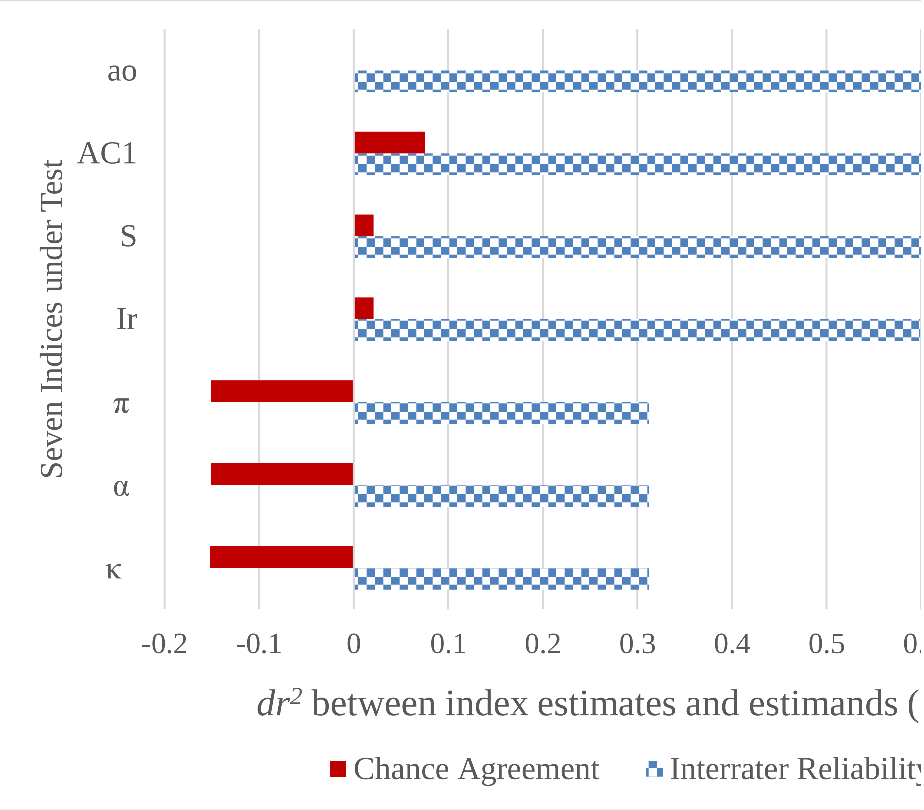

Figure 2 Accuracies of Interrater Reliability Indices.

Notes to Figure 2:

1. Solid red bars are dr^2 between estimated chance agreement & observed chance agreement.
2. Dotted blue bars are dr^2 between estimated interrater reliability & observed interrater reliability.
3. Primary benchmark: $dr^2 > 0.8$.
4. Data source: Lines 3 & 8, Table 3.