van Oest, R. (2018). A new coefficient of interrater agreement: The challenge of highly unequal category proportions. *Psychological Methods*. doi:10.1037/met0000183

A New Coefficient of Inter-Rater Agreement:
The Challenge of Highly Unequal Category Proportions

Accepted for publication in Psychological Methods

Rutger van Oest
BI Norwegian Business School

Author Note

Rutger van Oest, Department of Marketing, BI Norwegian Business School

All ideas and results of this study were presented at the 2017 INFORMS Marketing Science Conference at the University of Southern California, Los Angeles. The author thanks Auke Hunneman and George Knox for useful comments on earlier drafts.

Correspondence concerning this article should be addressed to Rutger van Oest, Department of Marketing, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway. Email: rutger.d.oest@bi.no

Abstract

We derive a general structure that encompasses important coefficients of inter-rater agreement such as the *S*-coefficient, Cohen’s kappa, Scott’s pi, Fleiss’ kappa, Krippendorff’s alpha and Gwet’s AC1. We show that these coefficients share the same set of assumptions about rater behavior; they only differ in how the unobserved category proportions are estimated. We incorporate Bayesian estimates of the category proportions and propose a new agreement coefficient with uniform prior beliefs. To correct for guessing in the process of item classification, the new coefficient emphasizes equal category probabilities if the observed frequencies are unstable due to a small sample, and the frequencies increasingly shape the coefficient as they become more stable. The proposed coefficient coincides with the *S*-coefficient for the hypothetical case of zero items; it converges to Scott’s pi, Fleiss’ kappa and Krippendorff’s alpha as the number of items increases. We use simulation to show that the proposed coefficient is as good as extant coefficients if the category proportions are equal and that it performs better if the category proportions are substantially unequal.

Keywords: inter-rater agreement, qualitative judgments, nominal categories

A New Coefficient of Inter-Rater Agreement:
The Challenge of Highly Unequal Category Proportions

Coding of qualitative items into nominal categories is a common task in the social sciences (e.g., Hughes & Garrett, 1990). Popular applications include content analysis, survey research and meta analysis (Perreault & Leigh, 1989). Content analysis classifies text, or other forms of communication, into content categories. Survey research often relies on open-ended questions of which the answers need to be categorized. Meta analysis requires measurement of the dimensions that describe the various settings in published empirical studies, which involves categorization if the dimensions are nominal.

Reproducibility, or inter-rater agreement, is essential for meaningful categorization of qualitative items (Kassarjian, 1977; Krippendorff, 2004). A situation in which different raters code the items very differently would imply poorly identified categories, threatening the validity of any subsequent analysis. Thus, inter-rater agreement should always be established and reported using appropriate coefficients (Lombard, Snyder-Duch, & Bracken, 2002; Stemler, 2004).

The literature contains a large number of easy-to-compute agreement coefficients. However, these coefficients take vastly different values for the same classification of items if the category frequencies are highly unequal. The differences arise from how the various coefficients correct for the notion that raters may agree on items by chance. If raters' judgments involve guessing, the literature disagrees about whether the categories are equally likely to be picked (Bennett, Alpert, & Goldstein, 1954; Brennan & Prediger, 1981; Perreault & Leigh, 1989) or the category probabilities should depend on the observed category frequencies (Cohen, 1960; Fleiss, 1971; Krippendorff, 2004; Scott, 1955).

Agreement coefficients relying on equal category probabilities ignore all information contained in the category frequencies. They do not capture that it is relatively easy to attain a high percentage of agreement if the category proportions are substantially unequal, resulting in a very mild correction for agreement by chance. Alternatively, frequency-based coefficients represent the unobserved true category proportions by the observed relative frequencies (Krippendorff, 2004). In so doing, these coefficients address the issue that rater agreement becomes more likely as the category proportions become more unequal (Stemler & Tsai, 2008). However, frequency-based coefficients may become unstable if raters classify a small number of items; these coefficients are relatively sensitive to small-sample noise. For example, one or two more item assignments to a category with a very low frequency may substantially alter the value of a frequency-based coefficient. The present study focuses on these two classes of chance-corrected agreement coefficients: equal-probability coefficients and frequency-based coefficients.

First, we review several important chance-corrected agreement coefficients, both based on equal probabilities and based on category frequencies. We show that these coefficients share a common set of (implicit) assumptions about rater behavior and that they follow from one general coefficient of inter-rater agreement that incorporates the unobserved category proportions; the extant coefficients only differ in how they estimate these proportions. Thus, we derive an overarching framework of agreement coefficients.

Second, we propose a new agreement coefficient that fits within the overarching framework and incorporates Bayesian estimates of the category proportions. It starts from equal category probabilities, but moves closer to traditional frequency-based coefficients as raters categorize more items and the relative frequencies thus become more precise. The new coefficient coincides with the *S*-coefficient for the hypothetical case of zero items and converges to Scott’s pi, Fleiss’ kappa and Krippendorff’s alpha as the sample size increases. It accounts for all information in the category frequencies, but discounts this information based on how imprecise it is.

Third, we employ the overarching framework to run a simulation that compares the performances of the various agreement coefficients. The proposed coefficient is as good as the other coefficients in scenarios with equal category proportions; it performs better if the category proportions are substantially unequal. For the latter set of scenarios, we show that the *S*-coefficient and Gwet's AC1 do not perform well in general, whereas Cohen's kappa, Scott's pi, Fleiss' kappa and Krippendorff's alpha are imprecise for small samples. We also provide two examples to illustrate the properties of the new coefficient.

A Review of Agreement Coefficients

We consider *R* raters (or coders or judges) who classify each of *N* items (or units or subjects) into one of *C* mutually exclusive categories, that is, one of the *C* levels of a categorical variable. The simplest agreement coefficient is the hit rate *H*, which is the number of agreements across all *pairs* of raters and all items, expressed as a fraction of the maximum attainable; this maximum is the number of items, *N*, times the number of rater pairs,

$$\binom{R}{2} = R(R-1)/2.$$

If $F_c^{(i)}$ denotes the number of raters assigning item *i* to category *c*, the number of pairwise rater agreements for item *i* equals

$$\textstyle\sum_{c=1}^{C} \binom{F_c^{(i)}}{2} = \sum_{c=1}^{C} F_c^{(i)}(F_c^{(i)}-1)/2,$$

so the number of pairwise agreements across *all* items becomes $\sum_{i=1}^{N}\sum_{c=1}^{C} F_c^{(i)}(F_c^{(i)}-1)/2$ (Fleiss, 1971). Thus, the hit rate can be written as

$$H = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C} F_c^{(i)}(F_c^{(i)}-1)/2}{NR(R-1)/2} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{F_c^{(i)}(F_c^{(i)}-1)}{R(R-1)}. \tag{1}$$

If $R = 2$, (1) reduces to the fraction of items on which both raters agree (Fleiss, 1971). Unfortunately, despite its ongoing popularity, the hit rate (or percentage agreement) is not a suitable agreement coefficient and should never be reported as the only measure (Grayson & Rust, 2001; Hughes & Garrett, 1990; Krippendorff, 2004; Lombard et al., 2002). The problem is that the hit rate in (1) does not exclude agreements that occur merely by chance. For example, a situation with $C = 2$ equally large categories and $R = 2$ raters who guess the categories of all items corresponds to classification without intrinsic value, but with an expected hit rate of 50%.

S-Coefficient (and Equivalent Coefficients)

As good agreement coefficients discard agreement by chance, they usually take the following form:

$$I = \frac{H - H_{\text{chance}}}{1 - H_{\text{chance}}}, \tag{2}$$

where H_{chance} is the hit rate that is expected by chance alone. The numerator in (2) is the extent to which the hit rate H exceeds H_{chance}; the denominator in (2) rescales the coefficient to ensure that the (theoretical) maximum value is $I = 1$; the other benchmark value $I = 0$ corresponds with $H = H_{\text{chance}}$. Assuming that all *C* categories are equally likely to be picked, there are C^2 possible combinations for any pair of raters of which *C* combinations result in agreement. Thus, $H_{\text{chance}} = C/C^2 = 1/C$. Substitution into (2) results in the *S*-coefficient by Bennett, Alpert and Goldstein (1954):

$$I_S = \frac{H - 1/C}{1 - 1/C} = \left(H - \frac{1}{C}\right)\frac{C}{C - 1}. \tag{3}$$

This agreement coefficient has been rediscovered several times (Zwick, 1988), resulting in Guilford’s *G*, Maxwell’s RE coefficient, Janson and Vegelius’ *C* and Brennan and Prediger’s κ_n (Brennan & Prediger, 1981; Holley & Guilford, 1964; Janson & Vegelius, 1979; Maxwell, 1977).

The *S*-coefficient and equivalent coefficients assume equal category probabilities to correct for agreement by chance; they do not account for the category frequencies. Proponents consider this a desirable property and base their arguments on symmetric treatment of the categories and lack of prior knowledge about the true category proportions (Perreault & Leigh, 1989). Their perspective is reasonable if the relative frequencies are unstable due to a very small sample or if the relative frequencies are approximately equally large. However, insensitivity to the category proportions seems undesirable if the relative frequencies are stable and highly skewed. For example, a hit rate of 95% is excellent if two categories are equally large, but the same percentage becomes less impressive if one of the two categories contains, say, 99% of the items; a hit rate of 99% is attainable by assigning all items to the large category, though researchers usually want to classify *both* categories (Krippendorff, 2004; Morrison, 1969). Thus, agreement coefficients relying on equal category probabilities apply only a very mild correction for agreement by chance; they do not capture that it is relatively easy to attain a high hit rate H if the category proportions are substantially unequal.

Cohen's Kappa

Cohen's kappa assumes $R = 2$ raters (Cohen, 1960). Though lacking a natural extension to more than two raters, it is feasible to take the average of kappa values across all pairs of raters (Conger, 1980). Cohen's kappa defines the hit rate expected by chance as

$$H_{\text{chance}} = \sum_{c=1}^{C} \left(F_{1,c}/N\right)\left(F_{2,c}/N\right), \tag{4}$$

where $F_{1,c}$ is the *number* of items that rater 1 assigns to category c, and $F_{1,c}/N$ is the corresponding *fraction* of items; frequency $F_{2,c}$ and fraction $F_{2,c}/N$ are defined similarly for rater 2. If the two raters pick categories in accordance with these fractions, $(F_{1,c}/N)(F_{2,c}/N)$ is the probability that both raters pick category c; H_{chance} is the sum across all categories. By substituting (4) into (2), we obtain

$$I_{\text{Cohen}} = \frac{H - \sum_{c=1}^{C}(F_{1,c}/N)(F_{2,c}/N)}{1 - \sum_{c=1}^{C}(F_{1,c}/N)(F_{2,c}/N)}. \quad (5)$$

Because H_{chance} in (4) is directly computed from the observed frequencies, Cohen's kappa in (5) is potentially unstable for small samples, with low *N*. The extant literature provides extensive discussions of the advantages and disadvantages of Cohen's kappa relative to other agreement coefficients (e.g., Banerjee, Cappozzoli, McSweeney, & Sinha, 1999; Feinstein & Cicchetti, 1990; Krippendorff & Fleiss, 1978; Zwick, 1988).

Scott's Pi, Fleiss' Kappa and Krippendorff's Alpha

Scott's pi is similar to Cohen's kappa, but the category probabilities are *the same* for both raters and reflect the true category proportions (Scott, 1955). By combining the frequencies from both raters, the proportion of category *c* can be estimated as $(F_{1,c} + F_{2,c})/2N$, so the probability that both raters pick category *c* is $([F_{1,c} + F_{2,c}]/2N)^2$. The probability of agreement by chance becomes

$$H_{\text{chance}} = \sum_{c=1}^{C}\left([F_{1,c} + F_{2,c}]/2N\right)^2. \quad (6)$$

This structure is similar to the proportional chance criterion in discriminant analysis (Morrison, 1969). Substitution into (2) yields Scott's pi:

$$I_{\text{Scott}} = \frac{H - \sum_{c=1}^{C}\left([F_{1,c} + F_{2,c}]/2N\right)^2}{1 - \sum_{c=1}^{C}\left([F_{1,c} + F_{2,c}]/2N\right)^2}. \quad (7)$$

It coincides with Cohen's kappa if $F_{1,c} = F_{2,c}$ for $c = 1, \dots, C$.

Fleiss' kappa is a straightforward extension of Scott's pi to more than $R = 2$ raters:

$$I_{\text{Fleiss}} = \frac{H - \sum_{c=1}^{C}\left(\sum_{r=1}^{R} F_{r,c}/RN\right)^2}{1 - \sum_{c=1}^{C}\left(\sum_{r=1}^{R} F_{r,c}/RN\right)^2}. \quad (8)$$

Furthermore, we obtain Krippendorff's alpha (in the context of nominal data) by applying a small-sample correction to (8), where RN is the sample size:

$$I_{\text{Kripp}} = \frac{H - \sum_{c=1}^{C}\left(\sum_{r=1}^{R} F_{r,c}/RN\right)\left([\sum_{r=1}^{R} F_{r,c} - 1]/[RN - 1]\right)}{1 - \sum_{c=1}^{C}\left(\sum_{r=1}^{R} F_{r,c}/RN\right)\left([\sum_{r=1}^{R} F_{r,c} - 1]/[RN - 1]\right)}. \tag{9}$$

For large RN, I_{Kripp} converges to I_{Scott} for two raters and I_{Fleiss} for any number of raters. Whereas Fleiss' kappa in (8) immediately uses the observed relative frequencies to define the probability distribution of the categories for *both* raters, Krippendorff's alpha in (9) recalculates this probability distribution for the *second* rater by excluding the first rater's choice for the item. Thus, the probability that the first rater picks category *c* remains $\sum_{r=1}^{R} F_{r,c}/RN$, but the probability that the second rater also picks category *c* becomes $(\sum_{r=1}^{R} F_{r,c} - 1)/(RN - 1)$, where both the frequency of category *c* and the total frequency across categories have been decreased by one. Put differently, the probability of agreement by chance is based on sampling *with* replacement for Fleiss' kappa, whereas it is based on sampling *without* replacement for Krippendorff's alpha. A powerful feature of the latter coefficient is that its general form extends to non-nominal coding, though the calculations tend to become complex (Lombard et al., 2002; Stemler & Tsai, 2008). Similar to Cohen's kappa, all three agreement coefficients are highly dependent on the observed category frequencies, making them potentially unstable for small samples. They usually provide values that are very similar to Cohen's kappa (Stemler & Tsai, 2008).

Gwet's AC1 Coefficient

Gwet's AC1 defines the probability of agreement by chance as the *joint* probability that at least one rater needs to guess the item's category and the outcome is agreement, where the probability of agreement *conditional* on guessing is defined as $1/C$ for all categories (Gwet, 2008):

$$I_{\text{Gwet}} = \frac{H - H_{\text{chance}}}{1 - H_{\text{chance}}} \quad \text{with} \quad H_{\text{chance}} = \frac{1}{C-1}\sum_{c=1}^{C}\left(\frac{\sum_{r=1}^{R} F_{r,c}}{RN}\right)\left(1 - \frac{\sum_{r=1}^{R} F_{r,c}}{RN}\right). \tag{10}$$

For $R = 2$ and $C = 2$, H_{chance} in (10) can be written as $1 - H_{\text{chance}}$ in (6) for Scott's pi. In general, I_{Gwet} uses the relative frequencies of categories in the opposite direction, resulting in a correction for agreement by chance that is even milder than the correction used in the S-coefficient; I_{Gwet} takes values that are far from Scott's pi and similar frequency-based coefficients if the category proportions are highly unequal.

Perreault-Leigh Coefficient

Unlike the previously discussed coefficients, Perreault and Leigh (1989) do not start from agreement coefficient (2), but take a model-based approach for $R = 2$ raters. Their coefficient, $I_{\text{PerrLeigh}}$, is the probability that a rater's judgment of an item's category is accurate, meaning that the rater assigns the item to its correct category without guessing. All raters are assumed to have the same probability of accurate judgment. Whenever a rater's judgment of an item's category is *not* accurate, the rater is assumed to randomly pick a category, with equal category probabilities. For fraction ${I_{\text{PerrLeigh}}}^2$ of the items, both raters make accurate judgments and thus choose the same category. For the other fraction $1 - {I_{\text{PerrLeigh}}}^2$, at least one of the two raters is not able to make an accurate judgment. Because raters, uncertain about an item's category, pick any category with probability $1/C$, the probability of pairwise rater agreement is also $1/C$. The hit rate H is the fraction of items for which both raters make accurate judgments plus the fraction of items for which at least one rater is not able to make an accurate judgment but there is agreement by chance:

$$H = {I_{\text{PerrLeigh}}}^2 + \left(1 - {I_{\text{PerrLeigh}}}^2\right)\frac{1}{C}. \tag{11}$$

Solving (11) with respect to $I_{\text{PerrLeigh}}$ yields the Perreault-Leigh coefficient:

$$I_{\text{PerrLeigh}} = \sqrt{\left(H - \frac{1}{C}\right)\frac{C}{C-1}}, \tag{12}$$

which is defined as zero if $H < 1/C$. Because the Perreault-Leigh coefficient is the square root of the *S*-coefficient, the *S*-coefficient becomes the square of the Perreault-Leigh coefficient. This confirms intuition: The chance-corrected probability of *pairwise* rater agreement is the probability that *both* raters make accurate judgments and thus assign an item to its correct category without guessing. We conclude that the *S*-coefficient is the Perreault-Leigh coefficient in the form of an agreement coefficient.

Summary

Popular agreement coefficients can be divided into three broad classes: (1) coefficients that do not correct for raters' guesses in the process of item classification (i.e., the simple hit rate, or equivalently, the percentage agreement), (2) coefficients that correct for agreement by chance without accounting for the observed category frequencies (i.e., the *S*-coefficient and equivalent coefficients), and (3) coefficients that correct for agreement by chance and account for the category frequencies (i.e., Cohen's kappa and other frequency-based coefficients). In the next sections, we first present a general framework that encompasses the chance-corrected agreement coefficients in classes 2 and 3; we show that these coefficients have a common structure that is consistent with an extension of the Perreault-Leigh model of rater behavior. Next, we refine this framework to obtain a new agreement coefficient that incorporates all information in the observed frequencies (the attractive feature of class 3), but discounts this information based on how imprecise it is (avoiding the weakness of class 3). Although Krippendorff's alpha also entails a (minor) small-sample correction, it does not use Bayesian updating principles to smooth out small-sample noise. We will use simulation to compare the performances of the various coefficients.

General Framework

We take a model-based approach that resembles the influential work of Perreault and Leigh (1989), which has received over 1500 Google Scholar citations. Rust and Cooil (1994) used a similar set of assumptions to derive a coefficient that is equivalent to the Perreault-Leigh coefficient for $R = 2$ raters but considers the *modal* judgment across raters, instead of individual rater judgments. Whereas Perreault and Leigh (1989) assumed equal category probabilities, we adjust this assumption to derive a general structure that is compatible with all discussed agreement coefficients in classes 2 and 3. We need the following assumptions:

1. Each rater acts independently.
2. Each rater makes an accurate judgment of an item's correct category with probability I_r.
3. If a judgment is not accurate, the rater picks category *c* for the item with probability $p_c, c = 1, \ldots, C$.
4. The researcher represents $(p_1, \ldots., p_C)$ by the category proportions $(\pi_1, \ldots., \pi_C)$.

Assumption 1 is a standard assumption and implies that raters do not communicate with each other (Krippendorff, 2004; Rust & Cooil, 1994). Assumptions 2 and 3 follow Perreault and Leigh (1989), where I_r is the probability of accurate judgment and raters pick a category probabilistically whenever they are unable to make an accurate judgment. Assumption 4 captures the researcher's belief that the category probabilities in Assumption 3 can be approximated by the corresponding category proportions: Raters are motivated to do the coding well and they use the notion that an item's correct category is more likely a large category, with large π_c, than a small one. Raters may already have some initial ideas about the category proportions or they may remember previous outcomes during the coding process and use availability heuristics (Tversky & Kahneman, 1974).

It follows from the assumptions that each rater's probability of assigning an item to category *c*, *not* conditional on whether the rater's judgment is accurate, coincides with the category's proportion π_c, as this probability is the outcome of two possible situations:

1. The rater makes an accurate judgment and the item's correct category is c, with joint probability $I_r\pi_c$.
2. The rater does not make an accurate judgment and picks category c in accordance with the category's proportion, with joint probability $(1 - I_r)\pi_c$.

Thus, the relative frequencies of categories converge to the unobserved category proportions if raters classify many items. In contrast, the original Perreault-Leigh process (leading to the *S*-coefficient) would imply unconditional probability $I_r\pi_c + (1 - I_r)(1/C)$ for category c, which is unequal to π_c if $I_r \neq 1$ and $\pi_c \neq 1/C$; the Perreault-Leigh process implies that the relative frequencies of categories tend to be more equal than the category proportions.

Derivation of General Agreement Coefficient for Two Raters

The hit rate *H* is the fraction of items on which both raters agree. Building on Assumptions 1-4, agreement occurs in three possible situations:

1. Both raters make accurate judgments for an item, which occurs with probability ${I_r}^2$.
2. No rater makes an accurate judgment for an item, but the raters agree by chance. The probability that both raters pick category c by chance is $(1 - I_r)^2{p_c}^2$. Summing over all possible categories $c = 1, \dots, C$ yields probability $\sum_{c=1}^{C}(1 - I_r)^2{p_c}^2$.
3. One rater makes an accurate judgment for an item and the other rater is correct by chance. If the item's correct category is c, this probability is $2I_r(1 - I_r)p_c$; multiplication by two occurs because the two raters are interchangeable. Summing over all possible correct categories and their probabilities of occurrence yields the probability *not* conditional on the correct category: $\sum_{c=1}^{C} p_c[2I_r(1 - I_r)p_c]$.

The hit rate is the result of all three scenarios:

$$H = {I_r}^2 + (1 - I_r)^2\sum_{c=1}^{C}{p_c}^2 + 2I_r(1 - I_r)\sum_{c=1}^{C}{p_c}^2 = {I_r}^2 + \left(1 - {I_r}^2\right)\sum_{c=1}^{C}{p_c}^2. \qquad (13)$$

Solving (13) with respect to I_r yields

$$I_r = \sqrt{\frac{H - \sum_{c=1}^{C} p_c{}^2}{1 - \sum_{c=1}^{C} p_c{}^2}}. \tag{14}$$

It immediately follows that the squared probability of accurate judgment, $I_r{}^2$, takes the form of a chance-corrected agreement coefficient:

$$I_r{}^2 = \frac{H - \sum_{c=1}^{C} p_c{}^2}{1 - \sum_{c=1}^{C} p_c{}^2}. \tag{15}$$

This again confirms intuition: the probability that *both* raters make accurate judgments is the chance-corrected probability of *pairwise* agreement.

Going Beyond Two Raters

The structures in (14) and (15) remain the same if more than two raters are involved in the coding process. The appendix proves that Assumptions 1-4 indeed imply probability coefficient (14), and thus agreement coefficient (15), for $R = 3$ raters. That is, for any feasible set of category proportions $(p_1, \ldots, p_C)$, incorporating (14) into Assumption 2 yields a process of item classification with expected hit rate *H*. Because the derivations become cumbersome for more than three raters, we use simulation to show that (14) is valid for any number of raters. We generate 10,000 scenarios by combining four hit rates *H* with five values of *R*, five values of *C*, and 100 sets of category probabilities $(p_1, \ldots, p_C)$. We take *H* = .60, .70, .80, or .90; *R* = 4, 5, 6, 10, or 20; *C* = 2, 3, 4, 10, or 20; and we draw $(p_1, \ldots, p_C)$ from the uniform Dirichlet(1, …, 1) distribution that accounts for the logical property that $\sum_{c=1}^{C} p_c = 1$. In each of the 10,000 scenarios, we take Assumptions 1-4, with I_r defined by (14), as the data generating process and simulate raters' classifications of 500,000 items. We establish that I_r in (14) is correct by showing that the simulated hit rate *H* matches the scenario's true hit rate. Across all scenarios, the mean absolute deviation between the simulated hit rate and the actual hit rate is a negligible .0002; the largest absolute deviation is .0017.

Same Structure, Different Estimates of Category Proportions

The chance-corrected agreement coefficients in classes 2 and 3 can be written as (15) and therefore have the same structure. However, they use different estimates of the category proportions p_c, $c = 1, \ldots, C$, in (15). For example, the S-coefficient imposes $p_c = 1/C$, Scott's pi and Fleiss' kappa estimate p_c as the category's average relative frequency across raters, and Cohen's kappa uses the corresponding geometric average for two raters. Concretely, we have

$$I_S = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \frac{1}{C};$$

$$I_{\text{Scott}} = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \frac{F_{1,c} + F_{2,c}}{2N};$$

$$I_{\text{Fleiss}} = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \frac{\sum_{r=1}^{R} F_{r,c}}{RN};$$

$$I_{\text{Cohen}} = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \sqrt{\left(\frac{F_{1,c}}{N}\right)\left(\frac{F_{2,c}}{N}\right)};$$

$$I_{\text{Kripp}} = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \sqrt{\left(\frac{\sum_{r=1}^{R} F_{r,c}}{RN}\right)\left(\frac{\sum_{r=1}^{R} F_{r,c} - 1}{RN - 1}\right)};$$

$$I_{\text{Gwet}} = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \sqrt{\frac{1}{C-1}\left(\frac{\sum_{r=1}^{R} F_{r,c}}{RN}\right)\left(1 - \frac{\sum_{r=1}^{R} F_{r,c}}{RN}\right)}, \qquad (16)$$

where the left part is the general structure (15), derived from Assumptions 1-4, and the right part is the specific estimate of category proportion p_c that is substituted into (15) to obtain the corresponding agreement coefficient. Thus, Assumptions 1-4 lead to all chance-corrected agreement coefficients, implying that these assumptions can act as the data generating process in a simulation that compares the performances of the various coefficients.

Proposed Agreement Coefficient

We introduce a new agreement coefficient by extending the general framework with a fifth assumption:

5. The researcher does not know the category proportions $(\pi_1, \ldots., \pi_C)$, but holds prior beliefs that are captured by the Dirichlet$(\alpha_1, \ldots., \alpha_C)$ distribution, with probability density function

$$f(\pi_1, \ldots., \pi_C) = \frac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} {\pi_c}^{\alpha_c - 1}, \qquad \pi_C = 1 - \sum_{c=1}^{C-1} \pi_c.$$

A key property of this distribution is that π_c has expected value $\alpha_c / \sum_{\tilde{c}=1}^{C} \alpha_{\tilde{c}}$, $c = 1, \ldots, C$. Furthermore, the sum of parameter values, $\sum_{\tilde{c}=1}^{C} \alpha_{\tilde{c}}$, reflects the degree of certainty in the prior beliefs; larger $\sum_{\tilde{c}=1}^{C} \alpha_{\tilde{c}}$ corresponds to less variance in the distribution. An important special case occurs if $(\alpha_1, \ldots., \alpha_C) = (1, \ldots, 1)$, as the Dirichlet distribution becomes uniform.

For $C = 2$, the Dirichlet distribution coincides with the more familiar beta distribution, with two parameters, $\alpha_1 > 0$ and $\alpha_2 > 0$. Figure 1 visualizes the beta density of π_1 (with $\pi_2 = 1 - \pi_1$) for different values of α_1 and α_2. The figure shows that the beta distribution, and thus the Dirichlet distribution, is able to capture a wide range of possible prior beliefs about the category proportions, including bimodal U-shapes (small α_1 and α_2, bottom left), inverted-U shapes (large α_1 and α_2, top right), a flat uniform pattern ($\alpha_1 = 1$ and $\alpha_2 = 1$, middle), inverted-J shapes (small α_1 and large α_2, top left), and J-shapes (large α_1 and small α_2, bottom right). On the diagonal, from bottom left to top right, the two parameters α_1 and α_2 take the same values, implying that the expected value of π_1 (and of $\pi_2 = 1 - \pi_1$) is ½. Whereas the prior beliefs on the diagonal are always that the categories have equal proportions, the degree of certainty of these beliefs increases when moving in the top right direction; the distribution ultimately becomes a zero-variance spike as both α_1 and α_2 tend to infinity.

Bayesian Estimates of Category Proportions

We use Bayesian principles to obtain estimates of the true category proportions and substitute these estimates into (15) to obtain a new agreement coefficient. Bayesian updating blends the (Dirichlet) prior beliefs about the category proportions with the observed category frequencies to obtain overall estimates. It acknowledges that the observed frequencies contain valuable information about the true category proportions, but also recognizes that the information is imperfect due to a limited sample size. Prior beliefs largely shape the Bayesian estimates of the category proportions if the number of items *N* is very small, but the observed relative frequencies will increasingly dominate as they become more stable due to larger *N*.

The category frequencies $F \equiv (\sum_{r=1}^{R} F_{r,1}, \dots, \sum_{r=1}^{R} F_{r,C})$ are outcomes of a multinomial distribution with probabilities $(\pi_1, \dots., \pi_C)$. Conjugacy of this multinomial distribution and the Dirichlet$(\alpha_1, \dots., \alpha_C)$ distribution of $(\pi_1, \dots, \pi_C)$ implies that the distribution of $(\pi_1, \dots, \pi_C)$, after incorporating F, is again Dirichlet (e.g., Rossi, Allenby, & McCulloch, 2005):

$$(\pi_1, \dots., \pi_C)|F \sim \text{Dirichlet}\left(\alpha_1 + \sum_{r=1}^{R} F_{r,1}, \dots, \alpha_C + \sum_{r=1}^{R} F_{r,C}\right). \tag{17}$$

The researcher's best estimate of the proportion of category *c* is the expected value of π_c from (17):

$$p_c = E(\pi_c|F) = \frac{\alpha_c + \sum_{r=1}^{R} F_{r,c}}{\sum_{\tilde{c}=1}^{C}(\alpha_{\tilde{c}} + \sum_{r=1}^{R} F_{r,\tilde{c}})} = \frac{\alpha_c + \sum_{r=1}^{R} F_{r,c}}{\sum_{\tilde{c}=1}^{C} \alpha_{\tilde{c}} + RN}. \tag{18}$$

Substitution into (15) yields a Bayesian agreement coefficient that incorporates both prior expectations, characterized by the parameters $(\alpha_1, \dots, \alpha_C)$, and the observed frequencies in *F*:

$${I_r}^2 = \frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2} \quad \text{with} \quad p_c = \frac{\alpha_c + \sum_{r=1}^{R} F_{r,c}}{\sum_{\tilde{c}=1}^{C} \alpha_{\tilde{c}} + RN}. \tag{19}$$

We note that (19) captures Scott's pi, Fleiss' kappa and the *S*-coefficient as limiting cases.

Limiting Case: Scott's Pi and Fleiss' Kappa

To eliminate the impact of prior beliefs, one may impose maximum uncertainty on these beliefs by letting $(\alpha_1, \dots., \alpha_C) \to (0, \dots, 0)$ in the Dirichlet$(\alpha_1, \dots., \alpha_C)$ distribution; the observed category frequencies would completely determine the value of the agreement coefficient. Substituting $(\alpha_1, \ldots., \alpha_C)$ into (19) yields Fleiss' kappa, that is, I_{Fleiss} in (16). Furthermore, Fleiss' kappa becomes Scott's pi, I_{Scott}, if $R = 2$.

Limiting Case: *S*-Coefficient

To eliminate the role of the observed category frequencies, and thus avoid Bayesian updating, one may assume complete certainty in the prior beliefs by letting $(\alpha_1, \ldots., \alpha_C) \rightarrow (\infty, \ldots, \infty)$; the prior beliefs would completely determine the value of the agreement coefficient. Substituting $(\alpha_1, \ldots., \alpha_C)$ into the right part of (19) yields $p_c = \alpha_c / \sum_{\tilde{c}=1}^{C} \alpha_{\tilde{c}}$, $c = 1, \ldots, C$. This, in turn, becomes $p_c = 1/C$, $c = 1, \ldots, C$, if the prior beliefs are equal category proportions, that is, $\alpha_1 = \cdots = \alpha_C$. We indeed obtain the *S*-coefficient, I_S in (16).

Agreement Coefficient With Uniform Prior Beliefs

Scott's pi and Fleiss' kappa require multimodal prior beliefs about the category proportions in order to reflect maximum uncertainty (i.e., the most bottom left location on the diagonal in Figure 1); the *S*-coefficient requires a zero-variance spike distribution for the prior beliefs in order to reflect complete certainty (i.e., the most top right location on the diagonal in Figure 1). Thus, both types of coefficients are located on the diagonal, but they occupy extreme locations; both multimodal and zero-variance prior beliefs are unrealistic. We propose a uniform prior distribution to operationalize the Bayesian agreement coefficient (19). The uniform prior is still located on the diagonal in Figure 1, but it takes its position between the two extremes. Similar to the *S*-coefficient, the prior beliefs are equal category proportions, but unlike the *S*-coefficient, these beliefs are held with uncertainty. The amount of updating of the uniform prior depends on the number of items *N*, and thus stability of the observed relative frequencies. The estimated category proportions remain relatively equal if *N* is small, whereas they become proportional to the category frequencies for large *N*. Substituting uniform prior beliefs, that is, $(\alpha_1, \ldots., \alpha_C) = (1, \ldots, 1)$, into (19) yields

$$I_r{}^2 = \frac{H - \sum_{c=1}^{C} p_c{}^2}{1 - \sum_{c=1}^{C} p_c{}^2} \quad \text{with} \quad p_c = \frac{1 + \sum_{r=1}^{R} F_{r,c}}{C + RN}. \tag{20}$$

This coefficient coincides with the *S*-coefficient if $N = 0$; it converges to Scott's pi, Fleiss' kappa and Krippendorff's alpha if $N \rightarrow \infty$. The new agreement coefficient (20) is located between these extant agreement coefficients.

Performance Comparison

We run a simulation to assess how well the proposed agreement coefficient (20) performs relative to other agreement coefficients. We take Assumptions 1-4 as the data generating process; we use the scenario's *true* probability of accurate judgment, I_{true}, in Assumption 2; we use the scenario's *true* category proportions p_1 and $p_2 = 1 - p_1$ in Assumptions 3-4. All discussed chance-corrected agreement coefficients are compatible with this data generating process, as all of them obey structure (15) and thus follow from Assumptions 1-4; these coefficients only incorporate different estimates of the *true* category proportions, p_1 and $p_2 = 1 - p_1$, leading to different estimates of ${I_{\text{true}}}^2$, the scenario's *true* level of chance-corrected agreement.

We obtain the scenarios by picking I_{true} = .50, .70, or .90; or equivalently, ${I_{\text{true}}}^2$ = .25, .49, or .81; and taking p_1 = .50, .70, .90, or .95 (equal category proportions, moderate symmetry, high asymmetry, or very high asymmetry). Furthermore, we vary the sample size by taking *N* = 50, 100, 200, or 1000; we set the number of raters at *R* = 2, 3, or 4. In each scenario, we generate 100,000 *samples* of item categorizations and compute the mean absolute error (MAE), that is, the mean absolute deviation between ${I_{\text{true}}}^2$ and the value of the agreement coefficient computed from each sample. Tables 1 to 3 report the differences in MAE between the proposed agreement coefficient and each of the other agreement coefficients; negative values mean that our coefficient has smaller MAE and is therefore more accurate. As Scott's pi is the two-rater version of Fleiss' kappa, we do not report it separately.

The tables show that all coefficients perform well if the category proportions are equal (i.e., if $p_1 = .50$); differences in MAE are .002 or less. It is striking that the *S*-coefficient does not perform better than the other coefficients, as scenarios with equal category proportions seem ideal for a coefficient imposing equal category probabilities. We conclude that choosing the right agreement coefficient is no major issue if the categories are (approximately) equally large.

Table 1 shows that the S-coefficient and Gwet's AC1 do not perform well if the category proportions are highly unequal ($p_1 = .90$ or $p_1 = .95$). This is already evident for small samples, like $N = 50$. Furthermore, the differences in performance between these two coefficients and the proposed coefficient increase as the sample size (in terms of both N and R) increases.

Tables 2 and 3 show that the proposed agreement coefficient also performs better than Fleiss' kappa (Scott's pi), Krippendorff's alpha and Cohen's kappa if the category proportions are highly unequal ($p_1 = 0.90$ or $p_1 = 0.95$). In particular, this holds for small N; we indeed argued that frequency-based coefficients are imprecise for small samples. However, the differences in performance disappear as the sample size increases, because the proposed coefficient ultimately converges to Fleiss' kappa and Krippendorff's alpha.

Thus, all agreement coefficients perform similarly if the category proportions are (approximately) equal, but the proposed coefficient outperforms the other coefficients if the category proportions are substantially unequal.

Two Examples

To further illustrate the new agreement coefficient (20), we provide two examples with highly unequal category proportions; these examples have been taken from published work. Table 4 shows the contingency tables; Table 5 provides the values of the agreement coefficients.

Example 1: Perreault and Leigh (1989)

Perreault and Leigh's study considers a situation in which two raters code 100 items into two categories. Both raters independently classify 90 items (90%) into the large category and 10 items (10%) into the small category; they agree on the small category only once. Because the raters agree on 82 out of 100 items, the hit rate is 82%. Table 5 shows that the proposed agreement coefficient (20) takes value .034, whereas Scott's pi, Fleiss' kappa and Cohen's kappa are exactly zero and Krippendorff's alpha is virtually zero. Furthermore, the *S*-coefficient and Gwet's AC1 are much larger: .64 and .78. Thus, the value of the proposed coefficient is between the values of extant coefficients that rely on equal category probabilities (in class 2) and the values of traditional frequency-based coefficients (in class 3). This is due to the Bayesian small-sample correction. The Bayesian estimates of the true category proportions, computed from the right part of (20), are .896 and .104. These numbers are close to the empirical relative frequencies, 90% and 10% (as used in Scott's pi and Fleiss' kappa), but are slightly adjusted toward the prior of equal category proportions (as used in the *S*-coefficient) to account for the somewhat small sample size.

Example 2: Gwet (2008)

Example 2a in Tables 4 and 5 is based on 125 items. For 118 items (94.4%), both raters conclude that these items belong to the large category. However, the two raters disagree on the other seven items and thus never agree on the small category. Again, the proposed agreement coefficient (20) is very different from the *S*-coefficient and Gwet's AC1 (.089 versus .89 and .94). Furthermore, it is substantially larger than the traditional frequency-based coefficients that even take negative values. The new agreement coefficient acknowledges raters' complete disagreement on items belonging to the small category, but its Bayesian small-sample correction discounts this outcome; zero agreement may have been a coincidence.

In Example 2b, we multiply all cells in the contingency table by four. Thus, the pattern remains the same (i.e., complete disagreement on the small category), but the sample size increases substantially. Whereas Krippendorff’s alpha decreases marginally, from −.025 to −.028, because of a minor small-sample correction, the proposed agreement coefficient decreases substantially, from .089 to .004. Because of the larger sample size, the new coefficient recognizes that the complete misclassification of the small category is unlikely a coincidence. The other agreement coefficients do not entail small-sample corrections and keep the same values as before.

Discussion

We presented an overarching framework of agreement coefficients and showed that popular equal-probability coefficients and frequency-based coefficients share a common set of assumptions about rater behavior. All discussed coefficients can be obtained from the derived general coefficient of inter-rater agreement by incorporating different estimates of the category proportions. Furthermore, we put forward Bayesian estimation of the category proportions and developed a new coefficient. An extensive simulation showed that the proposed coefficient is as good as the other coefficients if the category proportions are (approximately) equal, and that the coefficient outperforms the other coefficients if the category proportions are substantially unequal. The simulation identified conditions in which extant coefficients perform well, but also identified conditions in which these coefficients become less suitable.

By incorporating Bayesian estimates of the unobserved category proportions, the proposed coefficient accounts for all information in the observed category frequencies but discounts this information based on how imprecise it is. The Bayesian small-sample correction combines the strengths of equal-probability coefficients and frequency-based coefficients. The new coefficient is *relatively* similar to the *S*-coefficient (i.e., equal category probabilities) if the number of items is small, whereas it converges to several frequency-based coefficients as the number of items increases. The coefficient reconciles the two classes of chance-corrected agreement coefficients. Standard reference tables can be used to interpret its value. For example, a value above .60 may be interpreted as substantial rater agreement and a value above .80 may be interpreted as excellent rater agreement (Landis & Koch, 1977). However, others posit that agreement coefficients should only be used to establish that chance-corrected agreement exceeds zero (Uebersax, 2002).

The overarching framework, with its explicit assumptions that are shared by all discussed coefficients, demonstrates the boundaries of application of popular chance-corrected agreement coefficients. First, these coefficients are based on the assumption that raters make either completely accurate or completely random assignments, whereas actual item classifications tend to be characterized by *degrees* of uncertainty. Raters usually have enough information to avoid completely random assignment, but may not have enough information to be entirely certain about the correct category. For example, raters are often able to eliminate candidate categories but may be in doubt about the final category, because of either imprecisely defined categories or item characteristics that are consistent with multiple categories (Varki, Cooil, & Rust, 2000). Second, it follows from the assumptions that popular chance-corrected agreement coefficients treat all raters and all items in the same way; these coefficients are computed from *aggregate* numbers: category frequencies and counts of pairwise agreement. Thus, these coefficients ignore rater-specific characteristics, item-specific characteristics, and possible interactions (e.g., Kenny, 2004); rater characteristics make item classification easier (or harder) for some raters than for other raters; item characteristics make some items easier (or harder) to categorize than other items.

We do not propose a new model of how raters assign items to categories, but show that one has to accept a set of assumptions in order to use either popular extant agreement coefficients or the new coefficient. The assumptions can be relaxed, but this would result in classes of coefficients that are much harder to compute and that often require numerical optimization (e.g., Cooil & Rust, 1995; Dillon & Mulani, 1984; Varki et al., 2000).

Conclusion

The present study makes a theoretical contribution by developing an overarching framework capturing important coefficients of inter-rater agreement. Furthermore, it makes a practical contribution by presenting a new and easy-to-compute coefficient that is particularly suitable for highly unequal category proportions. We hope that the proposed coefficient will become an important tool for assessing the quality of qualitative judgments.

References

Banerjee, M., Capozzoli, M, McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27, 3-23. doi:10.2307/3315487

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, 18, 303-308. https://doi.org/10.1086/266520

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

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

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88, 322-328. http://dx.doi.org/10.1037/0033-2909.88.2.322

Cooil, B., & Rust, R. T. (1995). General estimators for the reliability of qualitative data. *Psychometrika*, 60, 199-220. https://link.springer.com/article/10.1007/BF02301413

Dillon, W. R., & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research*, 19, 438-458. https://doi.org/10.1207/s15327906mbr1904_5

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

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382. http://dx.doi.org/10.1037/h0031619

Grayson, K., & and Rust, R. (2001). Interrater reliability. *Journal of Consumer Psychology*, 10, 71-73. doi:10.1207/15327660151043998

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29-48. doi:10.1348/000711006X126600

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement*, 24, 749-753. http://journals.sagepub.com/doi/abs/10.1177/001316446402400402

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research*, 27, 185-195. doi:10.2307/3172845

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research*, 14, 255-269. https://doi.org/10.1207/s15327906mbr1402_9

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research*, 4, 8-18. http://dx.doi.org/10.1086/208674

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, 8, 265-280. http://journals.sagepub.com/doi/abs/10.1207/s15327957pspr0803_3

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30, 411-433. doi:10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K., & Fleiss, J. L. (1978). Reliability of binary attribute data. *Biometrics*, 34, 142-144.

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

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28, 587-604. doi:10.1111/j.1468-2958.2002.tb00826.x

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, 130, 79-83. doi:10.1192/bjp.130.1.79

Morrison, D. G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, 6, 156-163. https://archive.ama.org/archive/ResourceLibrary/JournalofMarketingResearch(JMR)/Pages/1969/6/2/5001475.aspx

Perreault, W. D., & Leigh, L. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, 26, 135-148. http://dx.doi.org/10.2307/3172601

Rossi, P. E., Allenby G. M., & McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Chichester, England: Wiley. doi:10.1002/0470863692

Rust, R. T., & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research*, 31, 1-14. http://www.jstor.org/stable/3151942

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

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment*, *Research & Evaluation*, 9 (4), 11 pages. http://PAREonline.net/getvn.asp?v=9&n=4

Stemler, S. E., & Tsai, J. (2008). Best practices in estimating interrater reliability. In J. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 29-49). Thousand Oaks, CA: Sage publications. http://dx.doi.org/10.4135/9781412995627.d5

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131. http://www.jstor.org/stable/1738360

Uebersax, J. (2002). *Statistical Methods for Diagnostic Agreement*. http://www.john-uebersax.com/stat/agree.htm

Varki, S, Cooil, B., & Rust, R. T. (2000). Modeling fuzzy data in qualitative marketing Research. *Journal of Marketing Research*, 37, 480-489. https://doi.org/10.1509/jmkr.37.4.480.18785

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103, 374-378. http://dx.doi.org/10.1037/0033-2909.103.3.374

Table 1

Comparison of Proposed Coefficient, S-Coefficient, and Gwet's AC1

			MAE of I_r^2 relative to I_S				MAE of I_r^2 relative to I_{Gwet}			
			p_1 =.50	p_1 =.70	p_1 =.90	p_1 =.95	p_1 =.50	p_1 =.70	p_1 =.90	p_1 =.95
I_{true} = .50	R = 2	N = 50	.001	−.030	−.325	−.420	.001	−.098	−.428	−.483
I_{true}^2 = .25		N = 100	.000	−.047	−.367	−.463	.000	−.125	−.471	−.526
		N = 200	.000	−.064	−.399	−.501	.000	−.149	−.504	−.565
		N = 1000	.000	−.094	−.444	−.559	.000	−.181	−.549	−.623
	R = 3	N = 50	.001	−.049	−.368	−.472	.000	−.128	−.471	−.535
		N = 100	.000	−.065	−.398	−.501	−.000	−.150	−.503	−.564
		N = 200	.000	−.080	−.421	−.530	.000	−.166	−.526	−.593
		N = 1000	.000	−.102	−.453	−.572	−.000	−.189	−.559	−.635
	R = 4	N = 50	.001	−.059	−.386	−.492	−.000	−.142	−.489	−.555
		N = 100	.000	−.074	−.411	−.517	.000	−.161	−.516	−.580
		N = 200	.000	−.088	−.431	−.541	−.000	−.174	−.536	−.605
		N = 1000	.000	−.105	−.458	−.577	.000	−.192	−.563	−.641
I_{true} = .70	R = 2	N = 50	.001	−.008	−.167	−.213	.002	−.045	−.236	−.256
I_{true}^2 = .49		N = 100	.000	−.020	−.212	−.260	.001	−.067	−.282	−.302
		N = 200	.000	−.032	−.245	−.302	.000	−.087	−.316	−.346
		N = 1000	.000	−.058	−.290	−.363	.000	−.117	−.361	−.406
	R = 3	N = 50	.001	−.017	−.201	−.253	.001	−.064	−.271	−.296
		N = 100	.000	−.030	−.237	−.291	.000	−.084	−.308	−.334
		N = 200	.000	−.042	−.263	−.326	.000	−.100	−.334	−.369
		N = 1000	.000	−.064	−.298	−.374	.000	−.123	−.369	−.417
	R = 4	N = 50	.001	−.024	−.217	−.269	.001	−.075	−.287	−.312
		N = 100	.000	−.036	−.249	−.305	.000	−.092	−.320	−.348
		N = 200	.000	−.048	−.272	−.337	.000	−.107	−.343	−.380
		N = 1000	.000	−.067	−.302	−.379	−.000	−.126	−.374	−.422
I_{true} = .90	R = 2	N = 50	.001	.006	−.007	.002	.002	.001	−.032	−.014
I_{true}^2 = .81		N = 100	.000	.001	−.041	−.042	.001	−.009	−.067	−.058
		N = 200	.000	−.004	−.065	−.075	.000	−.018	−.091	−.091
		N = 1000	.000	−.015	−.096	−.119	.000	−.036	−.123	−.135
	R = 3	N = 50	.001	.003	−.026	−.020	.001	−.006	−.052	−.036
		N = 100	.000	−.002	−.055	−.060	.000	−.015	−.082	−.076
		N = 200	.000	−.007	−.076	−.089	.000	−.024	−.102	−.105
		N = 1000	.000	−.018	−.101	−.126	.000	−.039	−.128	−.142
	R = 4	N = 50	.001	.001	−.037	−.031	.001	−.010	−.063	−.047
		N = 100	.000	−.004	−.064	−.071	.000	−.019	−.090	−.087
		N = 200	.000	−.009	−.081	−.097	.000	−.028	−.108	−.113
		N = 1000	.000	−.019	−.104	−.129	.000	−.041	−.131	−.145

Note. Objective is true chance-corrected agreement, I_{true}^2; MAE = mean absolute error; I_r^2 = proposed coefficient; I_S = *S*-coefficient; I_{Gwet} = Gwet's AC1.

Table 2

Comparison of Proposed Coefficient, Fleiss' Kappa, and Krippendorff's Alpha

			MAE of I_r^2 relative to I_{Fleiss}				MAE of I_r^2 relative to I_{Kripp}			
			p_1 =.50	p_1 =.70	p_1 =.90	p_1 =.95	p_1 =.50	p_1 =.70	p_1 =.90	p_1 =.95
I_{true} = .50	R = 2	N = 50	−.000	−.002	−.018	−.056	.002	−.000	−.015	−.052
I_{true}^2 = .25		N = 100	−.000	−.000	−.006	−.018	.000	−.000	−.005	−.017
		N = 200	−.000	−.000	−.002	−.006	.000	−.000	−.002	−.006
		N = 1000	−.000	−.000	−.000	−.001	.000	.000	−.000	−.000
	R = 3	N = 50	−.000	−.001	−.012	−.035	.001	−.000	−.010	−.033
		N = 100	−.000	−.000	−.004	−.012	.000	−.000	−.003	−.011
		N = 200	−.000	−.000	−.001	−.004	.000	−.000	−.001	−.004
		N = 1000	−.000	−.000	−.000	−.000	.000	−.000	−.000	−.000
	R = 4	N = 50	−.000	−.001	−.009	−.028	.001	−.000	−.008	−.027
		N = 100	−.000	−.000	−.003	−.009	.000	−.000	−.002	−.008
		N = 200	−.000	−.000	−.001	−.003	.000	−.000	−.001	−.003
		N = 1000	−.000	−.000	−.000	−.000	.000	−.000	−.000	−.000
I_{true} = .70	R = 2	N = 50	−.000	−.001	−.019	−.070	.001	−.000	−.017	−.066
I_{true}^2 = .49		N = 100	−.000	−.000	−.006	−.021	.000	−.000	−.005	−.020
		N = 200	−.000	−.000	−.002	−.006	.000	−.000	−.002	−.006
		N = 1000	−.000	−.000	−.000	−.001	.000	−.000	−.000	−.000
	R = 3	N = 50	−.000	−.001	−.012	−.049	.000	−.000	−.011	−.048
		N = 100	−.000	−.000	−.004	−.013	.000	−.000	−.003	−.013
		N = 200	−.000	−.000	−.001	−.004	.000	−.000	−.001	−.004
		N = 1000	−.000	−.000	−.000	−.000	.000	−.000	−.000	−.000
	R = 4	N = 50	−.000	−.001	−.009	−.039	.000	−.000	−.008	−.037
		N = 100	−.000	−.000	−.003	−.010	.000	−.000	−.003	−.010
		N = 200	−.000	−.000	−.001	−.003	.000	−.000	−.001	−.003
		N = 1000	−.000	−.000	−.000	−.000	.000	−.000	−.000	−.000
I_{true} = .90	R = 2	N = 50	−.000	−.001	−.014	−.051	.001	−.000	−.013	−.049
I_{true}^2 = .81		N = 100	−.000	−.000	−.004	−.016	.000	−.000	−.003	−.015
		N = 200	−.000	−.000	−.001	−.004	.000	.000	−.001	−.004
		N = 1000	−.000	−.000	−.000	−.000	.000	.000	−.000	−.000
	R = 3	N = 50	−.000	−.000	−.009	−.039	.000	−.000	−.009	−.038
		N = 100	−.000	−.000	−.002	−.011	.000	−.000	−.002	−.010
		N = 200	−.000	−.000	−.001	−.003	.000	−.000	−.001	−.002
		N = 1000	−.000	−.000	−.000	−.000	.000	.000	−.000	−.000
	R = 4	N = 50	−.000	−.000	−.007	−.034	.000	−.000	−.007	−.033
		N = 100	−.000	−.000	−.002	−.008	.000	−.000	−.001	−.008
		N = 200	−.000	−.000	−.001	−.002	.000	−.000	−.000	−.002
		N = 1000	−.000	−.000	−.000	−.000	.000	−.000	−.000	−.000

Note. Objective is true chance-corrected agreement, I_{true}^2; MAE = mean absolute error; I_r^2 = proposed coefficient; I_{Fleiss} = Fleiss' kappa; I_{Kripp} = Krippendorff's alpha.

Table 3

Comparison of Proposed Coefficient and Cohen's Kappa for Two Raters

		MAE of $I_r{}^2$ relative to I_{Cohen}			
		p_1 =.50	p_1 =.70	p_1 =.90	p_1 =.95
I_{true} = .50	N = 50	.002	.000	−.015	−.051
$I_{\text{true}}{}^2$ = .25	N = 100	.001	.000	−.005	−.016
	N = 200	.000	.000	−.002	−.005
	N = 1000	.000	.000	−.000	−.000
I_{true} = .70	N = 50	.001	−.000	−.017	−.066
$I_{\text{true}}{}^2$ = .49	N = 100	.000	−.000	−.005	−.020
	N = 200	.000	−.000	−.002	−.006
	N = 1000	.000	−.000	−.000	−.000
I_{true} = .90	N = 50	.000	−.000	−.014	−.050
$I_{\text{true}}{}^2$ = .81	N = 100	.000	−.000	−.003	−.016
	N = 200	.000	−.000	−.001	−.004
	N = 1000	.000	−.000	−.000	−.000

Note. Objective is true chance-corrected agreement, $I_{\text{true}}{}^2$; MAE = mean absolute error; $I_r{}^2$ = proposed coefficient; I_{Cohen} = Cohen's kappa.

Table 4

Contingency Tables in the Examples

	Contingency table			Remark
Example 1:	81	9	90	Independence, unequal proportions
Perreault and Leigh (1989)	9	1	10	
	90	10		
Example 2a:	118	5	123	Disagreement on small category
Gwet (2008)	2	0	2	
	120	5		
Example 2b:	472	20	492	Multiplied all cells by four
Gwet (2008) - adjusted	8	0	8	
	480	20		

Table 5

Agreement Coefficients in the Examples

	Example 1	Example 2a	Example 2b
Hit rate H (percentage agreement)	82%	94.4%	94.4%
Proposed agreement coefficient (${I_r}^2$ in (20))	.034	.089	.004
Scott's pi & Fleiss' kappa (I_{Scott} & I_{Fleiss} in (16))	.000	−.029	−.029
Krippendorff's alpha (I_{Kripp} in (16))	.005	−.025	−.028
Cohen's kappa (I_{Cohen} in (16))	.000	−.023	−.023
S-coefficient (I_S in (16))	.640	.888	.888
Gwet's AC1 (I_{Gwet} in (16))	.780	.941	.941

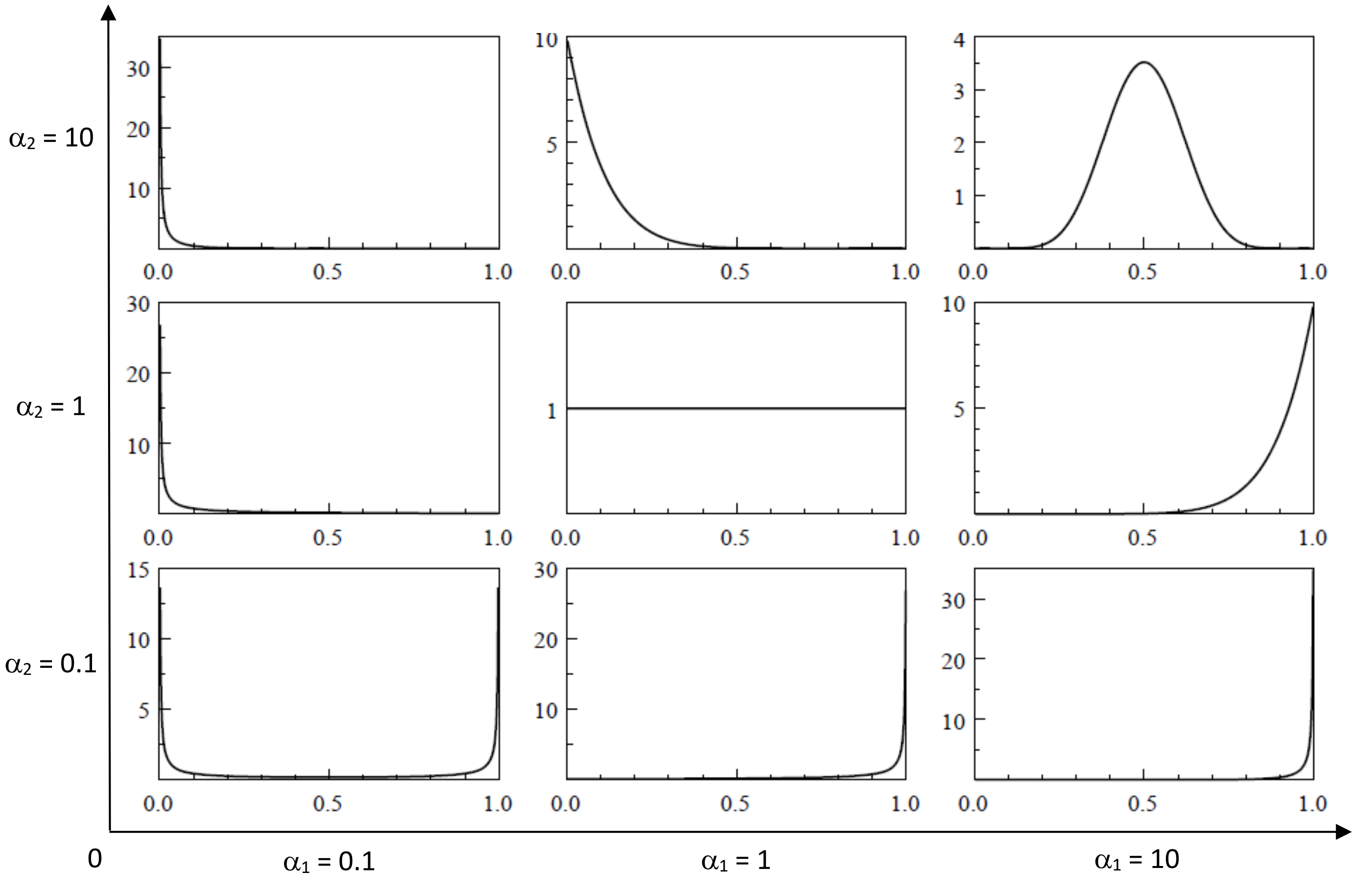

Figure 1. Shapes of the beta density for different values of its two parameters, α_1 and α_2.

Appendix

We consider $R = 3$ raters. The maximum number of pairwise agreements is $3N$, which is the number of rater pairs $\binom{3}{2}$ times the number of items N. Consistent with Fleiss (1971), we define the hit rate H as the number of agreements across all pairs of raters and all items, divided by the maximum, $3N$.

Situations in which all three raters agree on an item's classification correspond to three pairwise agreements. These situations are the following:

1. All three raters make accurate judgments for the item, which occurs with probability ${I_r}^3$.
2. Two raters make accurate judgments and the third rater is correct by chance. If the correct category is c, this probability is $\binom{3}{2}{I_r}^2(1 - I_r)p_c$; the binomial coefficient is the number of combinations in which two raters make accurate judgments and the third rater is correct by chance. Summing over all possible correct categories and their probabilities of occurrence yields the probability *not* conditional on the correct category: $\sum_{c=1}^{C} p_c[3{I_r}^2(1 - I_r)p_c]$.
3. One rater makes an accurate judgment and the other two raters are correct by chance. If the correct category is c, this probability is $\binom{3}{1} I_r(1 - I_r)^2{p_c}^2$; the binomial coefficient is again the number of valid rater combinations. Summing over all possible correct categories and their probabilities of occurrence yields the probability *not* conditional on the correct category: $\sum_{c=1}^{C} p_c[3I_r(1 - I_r)^2{p_c}^2]$.
4. No rater makes an accurate judgment, but all raters agree by chance. The probability that all three raters pick category c by chance is $(1 - I_r)^3{p_c}^3$. Summing over all possible categories $c = 1, \dots, C$ yields $\sum_{c=1}^{C}(1 - I_r)^3{p_c}^3$.

Taking everything together, the number of pairwise agreements due to items with unanimous agreement equals

$$3N\left({I_r}^3+3{I_r}^2(1-I_r)\sum_{c=1}^{C}{p_c}^2+3I_r(1-I_r)^2\sum_{c=1}^{C}{p_c}^3+(1-I_r)^3\sum_{c=1}^{C}{p_c}^3\right), \qquad \text{(A1)}$$

where we multiply by $3N$ because there are N items and each item with unanimous agreement contributes three pairwise agreements.

Situations in which two raters agree on an item's classification and the third rater disagrees correspond to one pairwise agreement:

5. Two raters make accurate judgments and the third rater is incorrect by chance. If the correct category is *c*, this probability is $\binom{3}{2}{I_r}^2(1-I_r)(1-p_c)$. Summing over all possible correct categories and their probabilities of occurrence yields the probability *not* conditional on the correct category: $\sum_{c=1}^{C}p_c[3{I_r}^2(1-I_r)(1-p_c)]$.
6. One rater makes an accurate judgment, one rater is correct by chance and one rater is incorrect by chance. If the correct category is *c*, this probability is $(3!)I_r(1-I_r)^2p_c(1-p_c)$. Summing over all possible correct categories and their probabilities of occurrence yields the probability *not* conditional on the correct category: $\sum_{c=1}^{C}p_c[6I_r(1-I_r)^2p_c(1-p_c)]$.
7. No rater makes an accurate judgment, but two raters agree by chance. For agreed category *c*, this probability is $\binom{3}{2}(1-I_r)^3{p_c}^2(1-p_c)$. Summing over all possible categories yields $\sum_{c=1}^{C}3(1-I_r)^3{p_c}^2(1-p_c)$.
8. One rater makes an accurate judgment, whereas the other two raters agree on an incorrect category by chance. If the correct category is *c*, the probability is $\binom{3}{1}I_r(1-I_r)^2\sum_{k\neq c}{p_k}^2$. Summing over all possible correct categories and their

probabilities of occurrence yields the probability *not* conditional on the correct category: $\sum_{c=1}^{C} p_c [3I_r(1-I_r)^2 \sum_{k\neq c} {p_k}^2]$, which we can rewrite as

$$3I_r(1-I_r)^2 \sum_{c=1}^{C} p_c \left[\sum_{k=1}^{C} {p_k}^2 - {p_c}^2\right] = 3I_r(1-I_r)^2 \left[\left(\sum_{c=1}^{C} p_c\right)\left(\sum_{k=1}^{C} {p_k}^2\right) - \sum_{c=1}^{C} {p_c}^3\right]$$

$$= 3I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^2 - 3I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^3.$$

Taking everything together, the number of pairwise agreements due to items with non-unanimous agreement equals

$$N\left(3{I_r}^2(1-I_r)\sum_{c=1}^{C} p_c(1-p_c) + 6I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^2(1-p_c) + 3(1-I_r)^3 \sum_{c=1}^{C} {p_c}^2(1-p_c)\right.$$
$$\left. + 3I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^2 - 3I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^3\right)$$
$$= 3N\left({I_r}^2(1-I_r)\sum_{c=1}^{C} p_c(1-p_c) + 2I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^2(1-p_c)\right.$$
$$\left. + (1-I_r)^3 \sum_{c=1}^{C} {p_c}^2(1-p_c) + I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^2 - I_r(1-I_r)^2 \sum_{c=1}^{C} {p_c}^3\right),$$

(A2)

where we multiply by N because there are N items and each item with non-unanimous agreement contributes one pairwise agreement.

The hit rate H is the total number of pairwise agreements, that is, (A1) plus (A2), divided by the maximum number of agreements, $3N$:

$$H = \left({I_r}^3 + 3{I_r}^2(1-I_r)\sum_{c=1}^{C} {p_c}^2 + 3I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^3 + (1-I_r)^3\sum_{c=1}^{C} {p_c}^3 \right)$$

$$+ \left({I_r}^2(1-I_r)\sum_{c=1}^{C} p_c(1-p_c) + 2I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^2(1-p_c) \right.$$

$$\left. + (1-I_r)^3\sum_{c=1}^{C} {p_c}^2(1-p_c) + I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^2 - I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^3 \right).$$

By combining the second term in the first set of brackets with the first term in the second set of brackets and doing the same for the next two terms, we obtain

$$H = {I_r}^3 + \left({I_r}^2(1-I_r)\sum_{c=1}^{C} p_c\,(p_c + (1-p_c)) + 2{I_r}^2(1-I_r)\sum_{c=1}^{C} {p_c}^2 \right)$$

$$+ \left(2I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^2(p_c + (1-p_c)) + I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^3 \right)$$

$$+ \left((1-I_r)^3\sum_{c=1}^{C} {p_c}^2(p_c + (1-p_c)) \right) + I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^2$$

$$- I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^3,$$

which we can rewrite as

$$H = {I_r}^3 + {I_r}^2(1-I_r) + 2{I_r}^2(1-I_r)\sum_{c=1}^{C} {p_c}^2 + 2I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^2 + I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^3$$

$$+ (1-I_r)^3\sum_{c=1}^{C} {p_c}^2 + I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^2 - I_r(1-I_r)^2\sum_{c=1}^{C} {p_c}^3. \tag{A3}$$

By combining the first two terms in the right-hand side of (A3), combining all terms containing $\sum_{c=1}^{C} {p_c}^2$, and noticing that the two terms containing $\sum_{c=1}^{C} {p_c}^3$ cancel out, we simplify (A3) into

$$H = {I_r}^2 + \left[2{I_r}^2(1 - I_r) + 2I_r(1 - I_r)^2 + (1 - I_r)^3 + I_r(1 - I_r)^2\right] \sum_{c=1}^{C} {p_c}^2$$

$$= {I_r}^2 + (1 - I_r)\left[2{I_r}^2 + 2I_r(1 - I_r) + (1 - I_r)^2 + I_r(1 - I_r)\right] \sum_{c=1}^{C} {p_c}^2$$

$$= {I_r}^2 + (1 - I_r)\left[2I_r + (1 - I_r)^2 + I_r - {I_r}^2\right] \sum_{c=1}^{C} {p_c}^2$$

$$= {I_r}^2 + (1 - I_r)(1 + I_r) \sum_{c=1}^{C} {p_c}^2 = {I_r}^2 + (1 - {I_r}^2) \sum_{c=1}^{C} {p_c}^2. \tag{A4}$$

Solving (A4) with respect to I_r yields

$$I_r = \sqrt{\frac{H - \sum_{c=1}^{C} {p_c}^2}{1 - \sum_{c=1}^{C} {p_c}^2}}. \tag{A5}$$