Psychological Methods
Please do not copy or cite without author's permission. The final article will be available, upon publication, via its DOI: 10.1037/met0000412

Weighting Schemes and Incomplete Data: A Generalized Bayesian Framework for Chance-Corrected Interrater Agreement

Rutger van Oest[1] and Jeffrey M. Girard[2]
[1]Department of Marketing, BI Norwegian Business School
[2]Department of Psychology, University of Kansas

Van Oest (2019) developed a framework to assess interrater agreement for nominal categories and complete data. We generalize this framework to all four situations of nominal or ordinal categories and complete or incomplete data. The mathematical solution yields a chance-corrected agreement coefficient that accommodates any weighting scheme for penalizing rater disagreements and any number of raters and categories. By incorporating Bayesian estimates of the category proportions, the generalized coefficient also captures situations in which raters classify only subsets of items; that is, incomplete data. Furthermore, this coefficient encompasses existing chance-corrected agreement coefficients: the S-coefficient, Scott's pi, Fleiss' kappa, and Van Oest's uniform prior coefficient, all augmented with a weighting scheme and the option of incomplete data. We use simulation to compare these nested coefficients. The uniform prior coefficient tends to perform best, in particular, if one category has a much larger proportion than others. The gap with Scott's pi and Fleiss' kappa widens if the weighting scheme becomes more lenient to small disagreements and often if more item classifications are missing; missingness biases play a moderating role. The uniform prior coefficient usually performs much better than the S-coefficient, but the S-coefficient sometimes performs best for small samples, missing data, and lenient weighting schemes. The generalized framework implies a new interpretation of chance-corrected weighted agreement coefficients: These coefficients estimate the probability that both raters in a pair assign an item to its correct category without guessing. Whereas Van Oest showed this interpretation for unweighted agreement, we generalize to weighted agreement.

Keywords: interrater agreement, categorical data, weighting schemes, incomplete data

Introduction

In many areas of social science, education, medicine, and business, a common and critical task is to assign items (e.g., individuals, objects, or ideas) to mutually exclusive categories. Examples include the classification of feelings (e.g., positive vs. negative), behaviors (e.g., smile vs. nonsmile), expertise (e.g., expert vs. novice), and document contents (e.g., fiction vs. nonfiction). The generated classifications may be used directly to make important decisions (e.g., to treat a patient classified as sick or to censor an email classified as fraudulent) or may be subjected to further analysis (e.g., meta-analytically comparing the results of studies classified in different groups). In either case, the validity of the subsequent decisions and analyses will hinge upon the quality and reproducibility of the categorical data (Stemler & Tsai, 2008).

A key source of evidence for reproducibility of categorical data comes from the analysis of interrater agreement: the degree to which different raters assign the same items to the same categories. As such, it is important to compute and report an appropriate coefficient of interrater agreement in all work using rater-based classifications (Stemler, 2004).

The literature contains many agreement coefficients to choose from. Most use the observed proportion of agreement as their basis and then attempt to adjust or "correct" this quantity by the amount of agreement that would be expected by chance (e.g., Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Cohen, 1960). A primary difference between the various coefficients lies in how they approach the estimation of chance agreement. One popular approach is to assume that all categories have an equal probability of being selected by chance (i.e., the category-based approach; Bennett, Alpert, & Goldstein, 1954; Brennan & Prediger, 1981); another is to assume that each category's probability is equal to its relative frequency in the observed sample (i.e., the distribution-based approach; Fleiss, 1971; Scott, 1955).

Recently, Van Oest (2019) proposed a Bayesian model-based approach to correcting for chance agreement that encompasses these two approaches and offers a new "hybrid" approach and a new coefficient, called the uniform prior coefficient. This approach begins with the prior belief that all categories have equal probabilities and then updates this belief using the observed data, where the amount of updating in-

creases with the number of items in the sample (N). Thus, the uniform prior coefficient is positioned between two extremes, coinciding with the category-based approach if $N = 0$ and converging to the distribution-based approach if $N \to \infty$. Through simulation, Van Oest showed that the uniform prior coefficient performs well compared to existing coefficients; in particular, if one category is much more likely to occur than others.

Similar to Broemeling (2001), Van Oest (2019) used a (uniform) Dirichlet distribution to capture uncertainty in the prior beliefs. However, Van Oest updated category probabilities, whereas Broemeling updated the probabilities of category combinations obtained from all raters. Furthermore, Broemeling computed interrater agreement if all item classifications occur by chance, whereas Van Oest considered *chance-corrected* agreement if item classifications may be either deliberate or by chance. Two other studies using Bayesian principles to estimate chance-corrected agreement are Basu, Banerjee, and Sen (2000) and Zhang and Cutter (2009); both studies considered situations with two categories only. The two-rater model of Basu et al. incorporated beta distributions and a (conditional) uniform distribution to capture prior beliefs about the probabilities that the first rater, the second rater, and both raters assign an item to the second category; the Dirichlet distribution extends the beta distribution to situations with more than two outcomes. Zhang and Cutter incorporated prior distributions for the response coefficients and intra-class correlation coefficient in a hierarchical probit model for correlated binary outcomes.

Although Van Oest's coefficient and many other coefficients assume unordered (i.e., nominal) categories and complete data, these assumptions are often inappropriate in applied settings. First, the data are often incomplete, either by accident (i.e., item classifications get lost or raters skip items) or by design (i.e., not all raters classify all items to save costs and minimize the burden placed on raters). Second, the categories are often ordered. For example, instead of assigning students to the nominal categories of pass or fail, teachers might use ordered (i.e., ordinal) categories (e.g., insufficient, sufficient, good, very good, or excellent). In such cases, two teachers assigning the same student to similar but not identical categories should receive "partial credit" towards their interrater agreement. Weighted agreement coefficients incorporate weighting schemes to describe the amount of credit for every possible rater disagreement, sometimes with extensions to incomplete (i.e., missing) data. The most prominent weighted agreement coefficient is the weighted kappa (Cohen, 1968). Furthermore, it is easy to extend several coefficients for nominal categories with a quadratic form in the chance component, such as the S-coefficient and Fleiss' kappa, to weighted versions (Gwet, 2014). More complex agreement coefficients allowing for ordered categories and possibly incomplete data are Gwet's AC2 and Krippendorff's alpha. Alternatively, Gajewski, Hart, Bergquist-Beringer, and Dunton (2007) proposed a Bayesian approach for interrater agreement with ordinal data, incorporating a hierarchical ordinal probit model with prior distributions for the error variance and intra-class correlation coefficient.

The present study proposes a generalization of Van Oest's framework and coefficient. This generalization accommodates unordered and ordered categories (with any weighting scheme for partial credit), is suitable for complete and incomplete data, and allows for any number of raters (R) and categories (C). We make three contributions. First, we extend Van Oest's model for dichotomous (unweighted) agreement to weighted agreement. We obtain a chance-corrected *weighted* agreement coefficient that estimates the probability that both raters in a pair assign an item to its correct category without guessing. Whereas Van Oest provided a derivation for two and three raters, without the option of partial credit, we extend this derivation to any number of raters and any weighting scheme. Second, we use Bayesian updating of the category proportions to obtain a generalized *weighted* agreement coefficient that allows for incomplete data and "nests" or subsumes (i) common category-based coefficients, (ii) common distribution-based coefficients, and (iii) the hybrid uniform prior coefficient. The generalized coefficient requires only a few lines of programming code and captures nested coefficients via different values of its input parameters. Third, we run a simulation to compare the performances of these nested coefficients and extract patterns in controlled situations with different weighting schemes, missing data mechanisms, and distributions of the category proportions. We also apply the generalized coefficient to a real-world data set with ordered categories and incomplete data.

Chance-Corrected Dichotomous Agreement

Extending the original ideas of Perreault and Leigh (1989), Van Oest (2019) described a model-based approach to correct for agreement by chance. The obtained coefficient encompasses several frequently used chance-corrected agreement coefficients:

$$\hat{I}_r^2 = \frac{\hat{A} - \sum_{c=1}^{C} \hat{p}_c^2}{1 - \sum_{c=1}^{C} \hat{p}_c^2}. \tag{1}$$

In Equation 1, $\hat{A}$ is the observed proportion of agreement across all pairs of raters and all items, C is the number of categories, and $\hat{p}_c$ is the estimated proportion of category $c \in \{1, \ldots, C\}$. For example, equally large estimated category proportions (i.e., $\hat{p}_c = 1/C$, $c = 1, \ldots, C$) result in the S-coefficient (Bennett et al., 1954; Brennan & Prediger, 1981). Similarly, defining $(\hat{p}_1, \ldots, \hat{p}_C)$ by the relative frequencies of the categories across all R raters and all N items results in Scott's pi for two raters and Fleiss' kappa for more than two raters (Fleiss, 1971; Scott, 1955). The assumptions needed to obtain structure (1) are the following:

1. When classifying an item, a rater is able to make an accurate judgment of the item's correct category with probability I_r.

2. If the rater's judgment is not accurate, the rater needs to guess the item's category and uses the category proportions $(p_1, \ldots, p_C)$ as the guessing probabilities.

3. All raters work independently; both accuracies (i.e., whether raters are able to provide accurate judgments) and category guesses (if raters are not accurate) are independent across raters.

The researcher does not observe probability I_r and category proportions $(p_1, \ldots, p_C)$ and therefore needs to infer these parameters from the observed data (i.e., the item classifications); we denote model parameters without a hat and corresponding estimators with a hat.

Van Oest (2019) discussed Assumptions 1–3 that are implicit in many existing agreement coefficients and provided an overview of agreement coefficients. Assumptions 1–3 imply that the unconditional category probabilities coincide with the corresponding proportions: Any rater assigns an item to category c if either (i) the judgment is accurate and the item's correct category is c, or (ii) the judgment is not accurate and the rater guesses c for the item, with unconditional probability $I_r p_c + (1 - I_r)p_c = p_c$. Furthermore, raters' item classifications are correlated if $I_r > 0$: All raters with accurate judgments choose the same (i.e., the correct) category for the item.

Chance-Corrected Weighted Agreement

Coefficient (1) counts only full agreements, where both raters in a pair choose the *same* category for an item; it assigns weight zero to any outcome in which the two raters do not choose the same category. We extend the dichotomous framework by allowing for weights that are greater than zero in situations without full pairwise rater agreement. Let $w_{c,\tilde{c}}$ denote weighted agreement when the first rater in a pair chooses category c and the second rater chooses category $\tilde{c}$. Although full agreements should receive full weight, that is, $w_{c,\tilde{c}} = 1$ if $c = \tilde{c}$, disagreements may receive either reduced weight or no weight at all, that is, $0 \leq w_{c,\tilde{c}} \leq 1$ if $c \neq \tilde{c}$ (Fleiss, Levin, & Paik, 2003). We note that Van Oest's framework imposed more restrictive identity weights, with weight zero for all disagreements: $w_{c,\tilde{c}} = 0$ if $c \neq \tilde{c}$.

To generalize agreement coefficient (1) to arbitrary (symmetric) weight matrices $W = (w_{c,\tilde{c}})$, we redefine the observed proportion of *dichotomous* agreement, $\hat{A}$, into the observed proportion of *weighted* agreement, $\hat{A}_w$; that is, the number of *weighted* pairwise agreements, divided by the corresponding maximum. For $R = 2$ raters, the maximum number of (weighted) pairwise agreements for an item is equal to one. Because the number of raters making accurate judgments about the item's correct category is two, one, or zero, we need to consider the following situations and corresponding contributions to the *expected* proportion of weighted agreement, A_w.

Situation 1. Both raters make accurate judgments, which occurs with probability I_r^2. Because this implies one (full) pairwise agreement, the contribution to A_w becomes I_r^2.

Situation 2. One rater makes an accurate judgment and the other rater needs to guess, which occurs with probability $\binom{2}{1} I_r(1 - I_r)$; the binomial coefficient captures that the two raters are interchangeable. If the item's correct category is c, the accurate rater chooses this category c and the guessing rater chooses category $\tilde{c} \in \{1, \ldots, C\}$ with probability $p_{\tilde{c}}$, which would result in weighted agreement $w_{c,\tilde{c}}$. Thus, *conditional on one accurate rater and the item's correct category being* c, expected weighted agreement is $\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_{\tilde{c}}$. By taking the expectation over the probability distribution of correct categories, $(p_1, \ldots, p_C)$, and combining with the probability of one out of two raters being accurate, we obtain the contribution to A_w:

$$\binom{2}{1} I_r(1-I_r) \sum_{c=1}^{C} p_c \left\{ \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_{\tilde{c}} \right\} = 2I_r(1-I_r) \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}.$$

Situation 3. No rater makes an accurate judgment (i.e., both raters need to guess), which occurs with probability $(1 - I_r)^2$. *Conditional on both raters guessing,* the first rater chooses category $c \in \{1, \ldots, C\}$ with probability p_c, and the second rater chooses category $\tilde{c} \in \{1, \ldots, C\}$ with probability $p_{\tilde{c}}$, resulting in expected weighted agreement $\sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}$. Thus, the contribution to A_w becomes

$$(1 - I_r)^2 \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}.$$

Combining the contributions to A_w from the three possible situations results in

$$\begin{aligned} A_w &= I_r^2 + 2I_r(1 - I_r) \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} + \\ &\quad (1 - I_r)^2 \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} \\ &= I_r^2 + (2I_r(1 - I_r) + (1 - I_r)^2) \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} \\ &= I_r^2 + (1 - I_r^2) \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}. \end{aligned} \tag{2}$$

Thus, the expected proportion of weighted agreement, A_w, equals the probability that both raters make accurate judgments, I_r^2, plus the probability that at least one rater is inaccurate, $1 - I_r^2$, times the expected weighted agreement if at least one rater is inaccurate, $\sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}$. It follows from Equation 2 that $I_r^2 \geq 0$ if and only if $A_w \geq$

$\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}$. Furthermore, Appendix A shows that identity (2) does not change when generalizing beyond two raters.

Theorem. *Under Assumptions 1–3, the expected proportion of weighted agreement across all rater pairs equals*

$$A_w = I_r^2 + (1 - I_r^2)\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}$$

for any number of raters $R \geq 2$. *For* $A_w \geq \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}$, *it holds that* $I_r^2 \geq 0$, *with*

$$I_r^2 = \frac{A_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}}.$$

By replacing A_w, $(p_1, \ldots, p_C)$, and I_r^2 in the theorem by their respective estimators $\hat{A}_w$, $(\hat{p}_1, \ldots, \hat{p}_C)$, and $\hat{I}_w^2$ (where we replaced subscript r by w to reflect the dependence on the weighting scheme), we obtain

$$\hat{A}_w = \hat{I}_w^2 + (1 - \hat{I}_w^2)\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}, \tag{3}$$

and

$$\hat{I}_w^2 = \frac{\hat{A}_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}}. \tag{4}$$

We note that Equation 4 is a chance-corrected weighted agreement coefficient that reduces to (1) if $w_{c,\tilde{c}} = 1$ for $c = \tilde{c}$ and $w_{c,\tilde{c}} = 0$ for all $c \neq \tilde{c}$. Matrix notation shows the convenient quadratic form in the chance component:

$$\hat{I}_w^2 = \frac{\hat{A}_w - \hat{p}'W\hat{p}}{1 - \hat{p}'W\hat{p}}, \tag{5}$$

where $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_C)$ is a $(C \times 1)$ vector, and $'$ denotes the transpose. Equation 5 captures weighted versions of existing coefficients with quadratic forms, described in Gwet (2014) and based on different choices of $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_C)$. Furthermore, this structure is similar to the weighted kappa coefficient that assumes two raters (Cohen, 1968; Fleiss, Cohen, & Everitt, 1969):

$$\kappa_w = \frac{\hat{A}_w - \hat{p}'W\hat{q}}{1 - \hat{p}'W\hat{q}}, \tag{6}$$

where $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_C)$ corresponds to the relative category frequencies based on the item classifications by the first rater, and $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_C)$ is analogous for the second rater.

An easy method to obtain standard errors and confidence intervals for $\hat{I}_w^2$ in (4) and (5) is bootstrapping (Efron, 1979). This method constructs the bootstrap sampling distribution of $\hat{I}_w^2$ by repeatedly sampling items with replacement from the original data set (while using the same sample size as in the original data set) and computing $\hat{I}_w^2$ for each simulated data set. Next, it uses this distribution to compute the standard errors and confidence intervals. The standard deviation provides the standard error of $\hat{I}_w^2$. Furthermore, the upper and lower quantiles provide easy percentile-based confidence intervals, although bias-corrected and accelerated (BCa) confidence intervals are preferable in both theory and practice (Efron, 1987; Efron & Tibshirani, 1993); the latter type corrects for bias and skewness in the bootstrap sampling distribution.

Because $\hat{I}_w^2$ in (4) and (5) corresponds to the square of the estimated probability of accurate judgment, we obtain an interpretation: Chance-corrected (weighted) agreement coefficients estimate the probability that both raters in a pair assign an item to its correct category without guessing. Whereas Van Oest (2019) showed this result for dichotomous agreement coefficients, we extend it to weighted agreement with arbitrary weight matrices $W = (w_{c,\tilde{c}})$.

Generalized Bayesian Coefficient

Following Van Oest (2019), we use Bayesian updating to estimate the unobserved category proportions $(p_1, \ldots, p_C)$. This approach starts from the beliefs that exist about $(p_1, \ldots, p_C)$ before observing any data outcomes. A so-called prior distribution captures these prior beliefs. For example, the mean of this distribution reflects the a priori expected category proportions, and the variance reflects the amount of uncertainty. For $N = 0$ items, the estimated category proportions would coincide with the a priori expected proportions. The next step is to obtain improved estimates by incorporating the observed data. Merging the category frequencies from the data with the prior distribution (i.e., prior beliefs) yields an updated distribution, the so-called posterior distribution. We estimate the category proportions by the mean of this distribution. As N increases, the estimated category proportions will converge to the corresponding relative frequencies in the data. Bayesian updating exploits the notion that the observed frequencies contain information about the category proportions but are unstable and potentially misleading for small samples; it relies relatively much on prior expectations for small samples and increasingly relies on observed frequencies as the sample size increases.

We use the Dirichlet distribution, with shape parameters $(\alpha_1, \ldots, \alpha_C)$, to describe the prior beliefs about the category proportions and follow similar steps as Van Oest (2019) to obtain the following Bayesian estimates of $(p_1, \ldots, p_C)$:

$$\hat{p}_c = \frac{\alpha_c + \sum_{i=1}^{N} R_{i,c}}{\sum_{\tilde{c}=1}^{C}\alpha_{\tilde{c}} + \sum_{i=1}^{N} R_i}, \quad c = 1, \ldots, C, \tag{7}$$

where $R_{i,c}$ is the number of raters who assigned item i to category c, and $R_i = \sum_{c=1}^{C} R_{i,c}$ is the number of raters who classified item i (by assigning it to any of the C categories). Bayesian updating incorporates all available item classifications but does not require that all raters classify all items. As

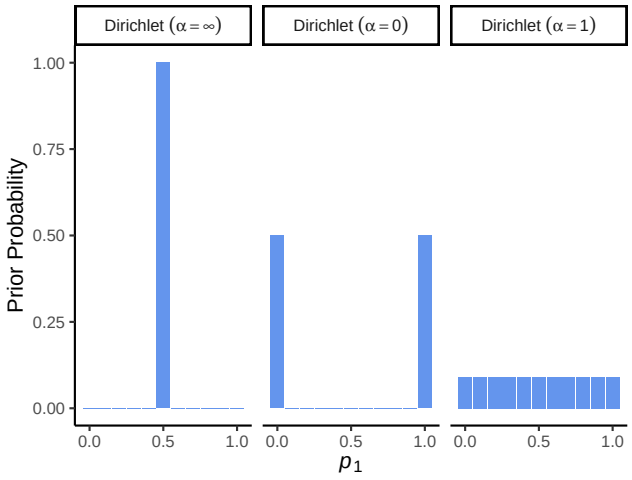

Figure 1

Discretized representation of the prior distribution for category proportion p_1 with two categories, where $p_2 = 1 - p_1$; the values of the Dirichlet parameters are $\alpha_1 = \alpha_2 = \infty$ for the S-coefficient (left), $\alpha_1 = \alpha_2 = 0$ for Fleiss' kappa (middle), and $\alpha_1 = \alpha_2 = 1$ for the uniform prior coefficient (right).

reflected by (7), the Bayesian approach allows for missing data, with $R_i \leq R$.

The Dirichlet prior distribution, implying (7), accounts for the logical property that category proportions need to sum to one. Furthermore, different values of α_c, $c = 1, \ldots, C$, allow for different a priori expected category proportions and different levels of uncertainty. By substituting (7) into the weighted agreement coefficient (4), we obtain a flexible structure:

$$\begin{aligned} \hat{I}^2_{w,\alpha} &= \frac{\hat{A}_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}} \\ \hat{p}_c &= \frac{\alpha_c + \sum_{i=1}^{N} R_{i,c}}{\sum_{\tilde{c}=1}^{C}\alpha_{\tilde{c}} + \sum_{i=1}^{N} R_i}. \end{aligned} \tag{8}$$

Structure (8) accommodates any weight matrix $W = (w_{c,\tilde{c}})$, including the identity matrix for dichotomous agreement, and allows for both complete and incomplete data. Furthermore, it nests (i) category-based coefficients, (ii) distribution-based coefficients, and (iii) the hybrid uniform prior coefficient, which are all obtained via different values of the Dirichlet parameters $(\alpha_1, \ldots, \alpha_C)$. To obtain "objective" agreement coefficients, we always begin with prior beliefs in which all categories are equally likely, with $\alpha_1 = \alpha_2 = \cdots = \alpha_C$, but these beliefs may be held with different degrees of uncertainty, implying different coefficients.

Category-Based Coefficients ($\alpha_c = \infty$)

For $\alpha_c \to \infty$, $c = 1, \ldots, C$, there is no uncertainty in the prior beliefs about the category proportions; the Dirichlet prior distribution becomes a zero-variance spike at point $p_c = 1/C$, $c = 1, \ldots, C$. The left panel in Figure 1 provides a discretized visualization of this distribution for p_1 with two categories, where $p_2 = 1 - p_1$; the principles extend to more than two categories but become hard to visualize. By letting all α_c in (8) tend to infinity at the same speed, we obtain a weighted version of the S-coefficient and equivalent coefficients (Bennett et al., 1954; Brennan & Prediger, 1981; Zwick, 1988):

$$\begin{aligned} \hat{I}^2_{w,\infty} &= \frac{\hat{A}_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}} \\ \hat{p}_c &= \frac{1}{C}. \end{aligned} \tag{9}$$

Because the estimated category proportions in (9) are the same for all categories and depend on only the number of categories, $\alpha_c \to \infty$ represents the class of category-based coefficients.

Distribution-Based Coefficients ($\alpha_c = 0$)

For $\alpha_c = 0$, $c = 1, \ldots, C$, there is maximum uncertainty in the prior beliefs about the category proportions; that is, the prior distribution becomes multimodal. The middle panel in Figure 1 shows this distribution for p_1 with two categories. Substituting $\alpha_c = 0$ into (8) yields a weighted version of Fleiss' kappa, extended to missing data:

$$\begin{aligned} \hat{I}^2_{w,0} &= \frac{\hat{A}_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}} \\ \hat{p}_c &= \frac{\sum_{i=1}^{N} R_{i,c}}{\sum_{i=1}^{N} R_i}, \end{aligned} \tag{10}$$

where Fleiss' kappa reduces to Scott's pi if there are only two raters, with $R = 2$ (Fleiss, 1971; Scott, 1955). Because the observed relative frequencies of the categories represent the category proportions in (10), $\alpha_c = 0$ corresponds to the class of distribution-based coefficients.

Hybrid Coefficient ($\alpha_c = 1$)

For $\alpha_c = 1$, $c = 1, \ldots, C$, all feasible combinations of category proportions (i.e., satisfying the logical property $\sum_{c=1}^{C} p_c = 1$) are a priori equally likely, a natural starting point (Broemeling, 2001; Van Oest, 2019). The right panel in Figure 1 shows the resulting "flat" uniform distribution without spike or multimodality. Substituting $\alpha_c = 1$ into (8)

yields a weighted version of Van Oest's uniform prior coefficient, extended to missing data:

$$\hat{I}^2_{w,1} = \frac{\hat{A}_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}\hat{p}_c\hat{p}_{\tilde{c}}} \qquad \hat{p}_c = \frac{1 + \sum_{i=1}^{N} R_{i,c}}{C + \sum_{i=1}^{N} R_i}. \tag{11}$$

Because (11) reduces to (9) if $N = 0$, and (11) reduces to (10) if $N \to \infty$, $\alpha_c = 1$ yields a hybrid form of category-based and distribution-based coefficients.

Observed Proportion of Weighted Agreement

Implementation of the generalized Bayesian coefficient (8), and its special cases (9) to (11), requires the observed proportion of weighted agreement, $\hat{A}_w$; that is, the number of weighted pairwise agreements, divided by the corresponding maximum. We write this ratio as

$$\hat{A}_w = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C} R_{i,c}\left(\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}R_{i,\tilde{c}} - 1\right)}{\sum_{i=1}^{N} R_i(R_i - 1)}. \tag{12}$$

The denominator in (12) captures that for each of the R_i raters who classified item $i \in \{1, \ldots, N\}$, the maximum number of agreements resulting from the choices by the other raters is $R_i - 1$. The numerator in (12) captures that for each of the $R_{i,c}$ raters who assigned item $i \in \{1, \ldots, N\}$ to category $c \in \{1, \ldots, C\}$, the number of weighted agreements resulting from the choices by the other raters is $\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}R_{i,\tilde{c}} - 1$; we deduct one point from $\sum_{c=1}^{C} w_{c,\tilde{c}}R_{i,\tilde{c}}$ to exclude the rater's self-agreement. We note that imposing identity weights and complete data (i.e., $w_{c,\tilde{c}} = 1$ if $c = \tilde{c}$, $w_{c,\tilde{c}} = 0$ if $c \neq \tilde{c}$, and $R_i = R$) reduces (12) to the expression derived by (Fleiss, 1971):

$$\hat{A} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C} R_{i,c}(R_{i,c} - 1)}{NR(R - 1)} = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c=1}^{C}\frac{R_{i,c}(R_{i,c} - 1)}{R(R - 1)}\right]. \tag{13}$$

We obtain the final coefficient by substituting (12) into (8). Appendix B provides a short R code containing one general function returning this coefficient. Using this function, we computed the S-coefficient (9), Fleiss' kappa (10), and the uniform prior coefficient (11), and we did so for dichotomous agreement and two forms of weighted agreement: linear and quadratic weights. We used a small synthetic data set in which four raters classified 30 items into one of three categories; approximately 20 percent of the item classifications were missing completely at random.

Simulation

Because the generalized Bayesian coefficient follows from a formal model of rater behavior, we use this underlying model as the data generating process to compare the performances of nested versions of (8), such as the S-coefficient (i.e., $\alpha_c \to \infty$), Fleiss' kappa (i.e., $\alpha_c = 0$), and the uniform prior coefficient (i.e., $\alpha_c = 1$). We follow similar simulation steps as Van Oest (2019) but consider two new and important dimensions: weighting schemes and missing data mechanisms. For each scenario, we simulate many data sets, compute the various chance-corrected (weighted) agreement coefficients for each data set and compare these coefficients with the true value; that is, I^2_r, where I_r is the scenario's true probability of accurate judgment. Next, we compute the mean absolute error (MAE) per scenario and coefficient (Van Oest, 2019). We provide Ox source code as supplementary material on the journal's website.

For each simulation scenario, we base the results on one million simulated data sets to obtain high precision. For the sample size (i.e., number of items), we take $N = 30$, 50, or 100, where raters classify fewer items in scenarios with incomplete data; we remove item classifications simulated as missing. We vary the number of raters from two to four. For the scenario's true chance-corrected agreement, we take either a moderate value or a high value: $I^2_r = .49$ (i.e., $I_r = .70$) or $I^2_r = .81$ (i.e., $I_r = .90$). These values for chance-corrected agreement are the same as in Van Oest (2019) and close to the values for Cohen's kappa in a simulation study by De Raadt, Warrens, Bosker, and Kiers (2019). Following Van Oest's strategy to keep the number of scenarios manageable, we consider only the smallest number of categories given the type of data. Whereas Van Oest focused on two nominal categories, we focus on three ordinal categories. Because Van Oest's results for two categories and our results for comparable scenarios with three categories are similar, we expect the obtained patterns to generalize to other numbers of categories.

We consider seven scenarios for the category proportions: one scenario with exactly equal proportions and six scenarios with unequal proportions. If the proportions are unequal, the first category has the largest proportion, either $p_1 = .50$ or $p_1 = .90$, and the third category has the smallest proportion; the second category has the same proportion as the third category (i.e., $p_2/p_3 = 1$), is three times as large (i.e., $p_2/p_3 = 3$), or is nine times as large (i.e., $p_2/p_3 = 9$). Because the weighting schemes are symmetric, changing the category proportions from descending to ascending would not affect the simulation results. Table 1 summarizes the seven scenarios and shows substantial differences in the considered category proportions, such as the number of large versus small categories and variation across categories.

Table 1

Summary of scenarios for the category proportions

Scenario		p_1	p_2	p_3	Remark on proportions
$p_1 = .\bar{3}$	$p_2/p_3 = 1$	.333	.333	.333	Exactly equal proportions
$p_1 = .5$	$p_2/p_3 = 1$	.500	.250	.250	Relatively equal proportions
$p_1 = .5$	$p_2/p_3 = 3$	.500	.375	.125	Moderate variation, one small
$p_1 = .5$	$p_2/p_3 = 9$	.500	.450	.050	Two large, one small
$p_1 = .9$	$p_2/p_3 = 1$	.900	.050	.050	One large, two equally small
$p_1 = .9$	$p_2/p_3 = 3$	.900	.075	.025	One large, unequal small
$p_1 = .9$	$p_2/p_3 = 9$	.900	.090	.010	One large, highly unequal small

Weighting Schemes

We consider dichotomous agreement (i.e., identity weights) and the two most common forms of weighted agreement: linear and quadratic weights. Linear weights start from weight one for full agreement and then deduct a penalty factor equal to the relative distance of disagreement (Cicchetti & Allison, 1971); quadratic weights use the squared relative distance as the penalty factor (Fleiss & Cohen, 1973). For three categories, the weight matrices become

$$W_{\text{identity}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$W_{\text{linear}} = \begin{pmatrix} 1.0 & 0.5 & 0.0 \\ 0.5 & 1.0 & 0.5 \\ 0.0 & 0.5 & 1.0 \end{pmatrix},$$

$$W_{\text{quadratic}} = \begin{pmatrix} 1.00 & 0.75 & 0.00 \\ 0.75 & 1.00 & 0.75 \\ 0.00 & 0.75 & 1.00 \end{pmatrix}.$$

Studies involving ordered categories may report coefficient values for both linear and quadratic weights, containing complementary information about the first and second moments of the distance of rater disagreement (Vanbelle, 2016). The literature also provides other interpretations of the weighted kappa coefficient with linear or quadratic weights (Cohen, 1968; Kvålseth, 2018; Schuster, 2004; Vanbelle & Albert, 2009; Warrens, 2011).

Missing Data Mechanisms

In addition to scenarios with complete data, we consider three mechanisms to generate incomplete data:

1. Missing classifications occur completely at random.
2. Missing classifications occur in the large category.
3. Missing classifications occur in the small categories.

De Raadt et al. (2019) considered the first two missingness mechanisms in their simulation study for the kappa coefficient. However, they did not include weighting schemes, comparison to other coefficients, and situations with more than two raters. The first mechanism implies that the probability of an item classification being missing is independent of the item's correct category; that is, item classifications are missing completely at random (MCAR). Because there are no systematic patterns for missing data, the relative category frequencies remain unbiased. The second and third mechanisms imply that the likelihood of missing depends on which category the item belongs to; that is, items are missing not at random (MNAR). In the second mechanism, raters always classify items belonging to the two small categories but may not classify items belonging to the large category. Thus, raters do not choose the large category often enough compared to the two small categories, resulting in a bias. In the third mechanism, missing data occur in the two small categories, resulting in an opposite bias. For all three mechanisms, we hold the overall percentage of missing data constant at 18%, approximately halfway the range considered by De Raadt et al. (2019). Table 2 summarizes the percentages of missing data per mechanism and category. We implement all three missing data mechanisms for scenarios with unequal category proportions but exclude the MNAR mechanisms if the category proportions are equal (i.e., if large versus small categories is not meaningful).

Results

We consider the differences in MAE between Fleiss' kappa (i.e., $\alpha_c = 0$) and the S-coefficient (i.e., $\alpha_c \to \infty$) on one side and the uniform prior coefficient (i.e., $\alpha_c = 1$) on the other side; the uniform prior coefficient serves as the benchmark. Positive differences in MAE imply that the uniform prior coefficient performs better (i.e., has lower MAE) than the other coefficients, whereas negative differences imply the opposite.

Equal category proportions

The simulation results for equal category proportions mirror those by Van Oest (2019): The three (weighted) coefficients have almost identical performances. Although the

Table 2

Fraction of missing data (m) for each of the categories and each of the three mechanisms

Missing Data Mechanism	Scenario		m_1	m_2	m_3
MCAR	$p_1 = .5$	$p_2/p_3 = 1$	.18	.18	.18
MCAR	$p_1 = .5$	$p_2/p_3 = 3$	.18	.18	.18
MCAR	$p_1 = .5$	$p_2/p_3 = 9$	.18	.18	.18
MCAR	$p_1 = .9$	$p_2/p_3 = 1$	.18	.18	.18
MCAR	$p_1 = .9$	$p_2/p_3 = 3$	.18	.18	.18
MCAR	$p_1 = .9$	$p_2/p_3 = 9$	.18	.18	.18
MNAR: large category	$p_1 = .5$	$p_2/p_3 = 1$	.36	.00	.00
MNAR: large category	$p_1 = .5$	$p_2/p_3 = 3$	.36	.00	.00
MNAR: large category	$p_1 = .5$	$p_2/p_3 = 9$	.36	.00	.00
MNAR: large category	$p_1 = .9$	$p_2/p_3 = 1$	.20	.00	.00
MNAR: large category	$p_1 = .9$	$p_2/p_3 = 3$	.20	.00	.00
MNAR: large category	$p_1 = .9$	$p_2/p_3 = 9$	.20	.00	.00
MNAR: small categories	$p_1 = .5$	$p_2/p_3 = 1$	.00	.36	.36
MNAR: small categories	$p_1 = .5$	$p_2/p_3 = 3$	.00	.36	.36
MNAR: small categories	$p_1 = .5$	$p_2/p_3 = 9$	.00	.36	.36
MNAR: small categories	$p_1 = .9$	$p_2/p_3 = 1$	.16	.36	.36
MNAR: small categories	$p_1 = .9$	$p_2/p_3 = 3$	.16	.36	.36
MNAR: small categories	$p_1 = .9$	$p_2/p_3 = 9$	.16	.36	.36

Note. Unequal category proportions. MCAR = Missing Completely at Random; MNAR = Missing Not at Random. The overall fraction missing across categories is $\sum_{c=1}^{3} p_c m_c = .18$ for all missing data mechanisms and scenarios. If $p_1 = .90$ and the missing data mechanism is MNAR in the small categories, we limit the fraction of missing data in the two small categories to .36 to avoid that these categories become too rare.

S-coefficient tends to perform best, the differences in MAE with Fleiss' kappa and the uniform prior coefficient are small and do not exceed .003 for the considered sample sizes. Thus, although the S-coefficient holds its prior belief of equal category proportions with certainty, consistent with the scenarios, this strong prior does not result in a substantially better performance. Table 3 shows the differences in MAE for all simulation scenarios with equal category proportions.

Unequal category proportions

We start the analysis for unequal category proportions from 1296 scenarios, based on all combinations in the simulation design. We reiterate the simulation parameters: sample size (N = 30, 50, or 100), proportion of the large category (p_1 = .50 or .90), proportion of the middle category relative to the small category (p_2/p_3 = 1, 3, or 9), weighting scheme (identity, linear, or quadratic), missingness (none, MCAR, large category, or small categories), true chance-corrected agreement (I_r = .70 or .90), and number of raters (R = 2, 3, or 4).

Sample Size and Category Proportions

First, we vary the sample size (N) and the proportion of the large category (p_1), which are the two simulation param-

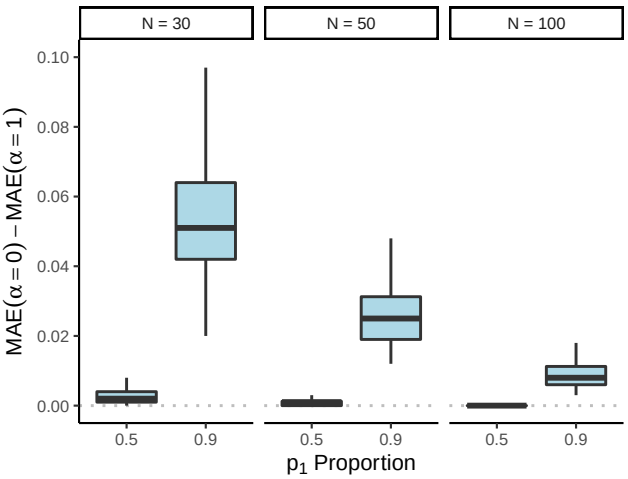

Figure 2

Distributions of marginalized simulation results comparing Fleiss' kappa ($\alpha_c = 0$) and the uniform prior coefficient ($\alpha_c = 1$) at different sample sizes and large category proportions.

Table 3

Differences in mean absolute error (MAE) for Fleiss' kappa ($\alpha_c = 0$) and the S-coefficient ($\alpha_c = \infty$) compared with the uniform prior coefficient ($\alpha_c = 1$); equal category proportions

				$R = 2$		$R = 3$		$R = 4$	
Compare	N	Weighting	Missing	$I_r = .70$	$I_r = .90$	$I_r = .70$	$I_r = .90$	$I_r = .70$	$I_r = .90$
$\alpha_c = 0$	30	Identity	None	.000	.000	.000	.000	.000	.000
	30	Linear	None	.000	.000	.000	.000	.000	.000
	30	Quadratic	None	.001	.000	.000	.000	.000	.000
	30	Identity	MCAR	.001	.000	.000	.000	.000	.000
	30	Linear	MCAR	.001	.000	.000	.000	.000	.000
	30	Quadratic	MCAR	.001	.001	.001	.000	.000	.000
	50	Identity	None	.000	.000	.000	.000	.000	.000
	50	Linear	None	.000	.000	.000	.000	.000	.000
	50	Quadratic	None	.000	.000	.000	.000	.000	.000
	50	Identity	MCAR	.000	.000	.000	.000	.000	.000
	50	Linear	MCAR	.000	.000	.000	.000	.000	.000
	50	Quadratic	MCAR	.000	.000	.000	.000	.000	.000
	100	Identity	None	.000	.000	.000	.000	.000	.000
	100	Linear	None	.000	.000	.000	.000	.000	.000
	100	Quadratic	None	.000	.000	.000	.000	.000	.000
	100	Identity	MCAR	.000	.000	.000	.000	.000	.000
	100	Linear	MCAR	.000	.000	.000	.000	.000	.000
	100	Quadratic	MCAR	.000	.000	.000	.000	.000	.000
$\alpha_c = \infty$	30	Identity	None	−.002	−.002	−.002	−.002	−.002	−.001
	30	Linear	None	−.001	−.002	−.001	−.001	−.001	−.001
	30	Quadratic	None	.000	−.002	.001	−.002	.001	−.002
	30	Identity	MCAR	−.003	−.002	−.002	−.002	−.002	−.002
	30	Linear	MCAR	−.002	−.002	−.001	−.002	−.001	−.002
	30	Quadratic	MCAR	−.001	−.003	.001	−.002	.001	−.002
	50	Identity	None	−.002	−.001	−.001	−.001	−.001	−.001
	50	Linear	None	.000	−.001	.000	−.001	.000	−.001
	50	Quadratic	None	.001	−.001	.001	−.001	.001	−.001
	50	Identity	MCAR	−.001	−.001	−.001	−.001	−.001	−.001
	50	Linear	MCAR	−.001	−.001	.000	−.001	.000	−.001
	50	Quadratic	MCAR	.001	−.001	.001	−.001	.001	−.001
	100	Identity	None	−.001	.000	.000	.000	.000	.000
	100	Linear	None	.000	.000	.000	.000	.000	.000
	100	Quadratic	None	.001	.000	.001	.000	.001	.000
	100	Identity	MCAR	−.001	.000	.000	.000	.000	.000
	100	Linear	MCAR	.000	.000	.000	.000	.000	.000
	100	Quadratic	MCAR	.001	.000	.001	.000	.001	.000

Note. Positive scores indicate that the uniform prior coefficient performs better; negative scores indicate that the alternative coefficient performs better.

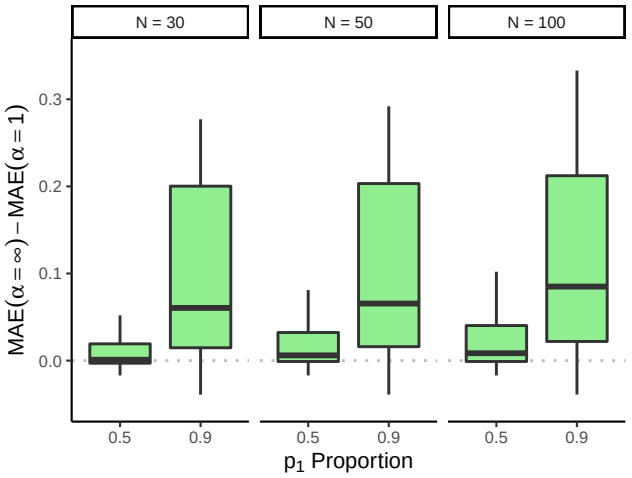

Figure 3

Distributions of marginalized simulation results comparing the S-coefficient ($\alpha_c = \infty$) and the uniform prior coefficient ($\alpha_c = 1$) at different sample sizes and large category proportions.

eters that affect the results most; we aggregate over all other parameters. The box plots in Figure 2 visualize the distributions of the differences in MAE between Fleiss' kappa and the uniform prior coefficient for different values of N and p_1. Similarly, Figure 3 compares the S-coefficient and the uniform prior coefficient for different values of N and p_1.

Figures 2 and 3 resemble the results reported by Van Oest (2019). Figure 2 shows that the uniform prior coefficient performs better than Fleiss' kappa in all scenarios. The differences are small if no category dominates in terms of proportions (i.e., if $p_1 = .50$) but quite substantial if a dominating category is present (i.e., if $p_1 = .90$) and the sample is small (i.e., $N = 30$ or 50). Furthermore, Figure 3 shows that the uniform prior coefficient often performs much better than the S-coefficient if a dominating category is present (i.e., if $p_1 = .90$), although the S-coefficient sometimes performs best.

Because sample size $N = 50$ is the middle option in the simulation design, and large differences between the three coefficients tend to occur for $p_1 = .90$, we narrow down to scenarios with $N = 50$ and $p_1 = .90$. An advantage is that the more subtle effects of the other simulation parameters become more visible without being polluted by the stronger effects of N and p_1. Another advantage is that the number of scenarios reduces by factor six, from 1296 to 216, making it feasible to show detailed results for all these scenarios. Figure 4 shows the distributions of the differences in MAE between Fleiss' kappa and the uniform prior coefficient, based on the 216 scenarios; it varies one simulation parameter at a time to obtain the corresponding marginalized simulation results; Table 4 provides the underlying numbers for each of the 216 scenarios. Similarly, Figure 5 and Table 5 compare the S-coefficient and the uniform prior coefficient.

A first result, shown in Figures 4 and 5, is that an increase in the proportion of the middle category relative to the small category (p_2/p_3) decreases the relative performance of the S-coefficient but has no meaningful effect on Fleiss' kappa. Thus, the S-coefficient becomes less suitable if one category is relatively rare. In line with our finding, Krippendorff (2004, p. 418) noted that the S-coefficient "becomes inflated by unused categories."

Weighting

Figure 4 shows that the uniform prior coefficient becomes relatively more preferred over Fleiss' kappa when moving from dichotomous agreement to forms of weighted agreement, with the largest changes being when moving from linear to quadratic weights. Thus, compared to Fleiss' kappa, the uniform prior coefficient benefits from generous weighting schemes that award relatively much credit to disagreements. However, Figure 5 shows that generous weighting schemes benefit the S-coefficient most, where again the transition from linear to quadratic weights triggers the most substantial changes in relative performance.

Missing

Figure 4 shows that incomplete data improve the performance of the uniform prior coefficient relative to Fleiss' kappa for two of the three missingness mechanisms, where missing data occur either completely at random (i.e., the first mechanism) or in the two small categories (i.e., the third mechanism). For the remaining (second) mechanism in which missing data occur in the large category, the relative performance remains essentially unaffected compared to the situation of complete data.

Intuitively, missing item classifications decrease the precision of the relative category frequencies observed in the data, which makes discounting via a somewhat informative (e.g., uniform) prior more important. However, Fleiss' kappa uses a prior with the lowest possible information (i.e., maximum variance) and therefore relies completely on these relative frequencies; missing data have a particularly undesirable effect on the performance of Fleiss' kappa. On the other hand, missing data in the large category (i.e., the second mechanism) helps Fleiss' kappa via a counter mechanism: The bias makes the relative frequencies in the data more balanced than the corresponding category proportions, which benefits Fleiss' kappa by triggering a similar effect as the prior expectation of equal category proportions in the uniform prior coefficient.

Figure 5 shows that incomplete data improve the performance of the S-coefficient relative to the uniform prior coefficient. Although missing item classifications reduce the

Table 4

Differences in mean absolute error (MAE) between Fleiss' kappa ($\alpha_c = 0$) and the uniform prior coefficient ($\alpha_c = 1$); unequal category proportions

			$R = 2$		$R = 3$		$R = 4$	
Weighting	Missing	p_2/p_3	$I_r = .70$	$I_r = .90$	$I_r = .70$	$I_r = .90$	$I_r = .70$	$I_r = .90$
Identity	None	1	.028	.024	.019	.016	.015	.012
		3	.027	.023	.019	.016	.015	.012
		9	.027	.024	.019	.016	.015	.012
Linear	None	1	.030	.025	.021	.017	.017	.013
		3	.031	.028	.023	.019	.019	.015
		9	.029	.029	.022	.020	.018	.016
Quadratic	None	1	.037	.029	.026	.020	.021	.016
		3	.041	.035	.031	.025	.027	.021
		9	.029	.035	.024	.026	.022	.022
Identity	MCAR	1	.037	.032	.023	.020	.018	.015
		3	.037	.032	.023	.020	.018	.015
		9	.037	.032	.023	.020	.017	.015
Linear	MCAR	1	.040	.033	.025	.021	.019	.016
		3	.042	.037	.026	.023	.021	.018
		9	.041	.039	.025	.025	.020	.019
Quadratic	MCAR	1	.048	.037	.030	.024	.024	.019
		3	.052	.043	.033	.030	.028	.024
		9	.041	.046	.024	.031	.022	.025
Identity	Large	1	.031	.025	.019	.016	.015	.012
		3	.031	.025	.019	.016	.015	.012
		9	.031	.025	.020	.016	.015	.012
Linear	Large	1	.033	.026	.021	.017	.017	.013
		3	.035	.030	.023	.020	.019	.015
		9	.035	.031	.023	.021	.018	.016
Quadratic	Large	1	.040	.031	.026	.020	.020	.016
		3	.046	.037	.032	.026	.027	.021
		9	.039	.039	.027	.027	.023	.022
Identity	Small	1	.046	.043	.026	.025	.020	.018
		3	.047	.043	.026	.025	.020	.018
		9	.047	.043	.026	.025	.020	.018
Linear	Small	1	.050	.043	.028	.026	.022	.020
		3	.050	.048	.027	.029	.021	.022
		9	.047	.051	.024	.030	.019	.023
Quadratic	Small	1	.059	.045	.034	.030	.026	.023
		3	.058	.052	.032	.035	.026	.027
		9	.041	.056	.017	.035	.015	.028

Note. Positive scores indicate that the uniform prior coefficient performs better; negative scores would indicate that Fleiss' kappa performs better; $N = 50$ and $p_1 = .90$ are fixed.

Table 5

Differences in mean absolute error (MAE) between the S-coefficient ($\alpha_c = \infty$) and the uniform prior coefficient ($\alpha_c = 1$); unequal category proportions

			$R = 2$		$R = 3$		$R = 4$	
Weighting	Missing	p_2/p_3	$I_r = .70$	$I_r = .90$	$I_r = .70$	$I_r = .90$	$I_r = .70$	$I_r = .90$
Identity	None	1	.105	−.019	.152	.002	.173	.013
		3	.163	.001	.213	.023	.235	.036
		9	.207	.027	.256	.047	.279	.058
Linear	None	1	.202	.015	.253	.043	.276	.056
		3	.202	.015	.253	.042	.276	.056
		9	.201	.015	.253	.043	.276	.056
Quadratic	None	1	.180	.003	.233	.032	.257	.047
		3	.201	.016	.254	.043	.278	.057
		9	.215	.026	.266	.052	.292	.065
Identity	MCAR	1	.146	−.016	.205	.016	.232	.031
		3	.181	.006	.241	.034	.269	.048
		9	.205	.030	.262	.053	.292	.066
Linear	MCAR	1	.119	−.024	.177	.002	.204	.015
		3	.147	−.010	.204	.016	.232	.029
		9	.165	.002	.221	.027	.249	.040
Quadratic	MCAR	1	.082	−.038	.144	−.013	.174	−.001
		3	.127	−.012	.191	.010	.223	.023
		9	.159	.014	.219	.034	.252	.046
Identity	Large	1	.130	−.016	.180	.006	.203	.018
		3	.130	−.016	.181	.006	.203	.018
		9	.130	−.016	.181	.007	.204	.019
Linear	Large	1	.101	−.025	.154	−.004	.178	.007
		3	.135	−.012	.188	.010	.212	.022
		9	.157	.000	.209	.022	.234	.033
Quadratic	Large	1	.065	−.039	.120	−.019	.145	−.009
		3	.119	−.015	.180	.004	.209	.016
		9	.161	.011	.219	.030	.249	.042
Identity	Small	1	.156	−.013	.216	.019	.245	.034
		3	.155	−.013	.216	.019	.245	.034
		9	.155	−.014	.216	.019	.245	.034
Linear	Small	1	.133	−.021	.195	.009	.226	.024
		3	.153	−.007	.214	.022	.245	.036
		9	.166	.004	.225	.031	.256	.045
Quadratic	Small	1	.097	−.034	.165	−.006	.198	.009
		3	.129	−.007	.195	.016	.230	.030
		9	.151	.016	.211	.037	.247	.049

Note. Positive scores indicate that the uniform prior coefficient performs better; negative scores indicate that the S-coefficient performs better; $N = 50$ and $p_1 = .90$ are fixed.

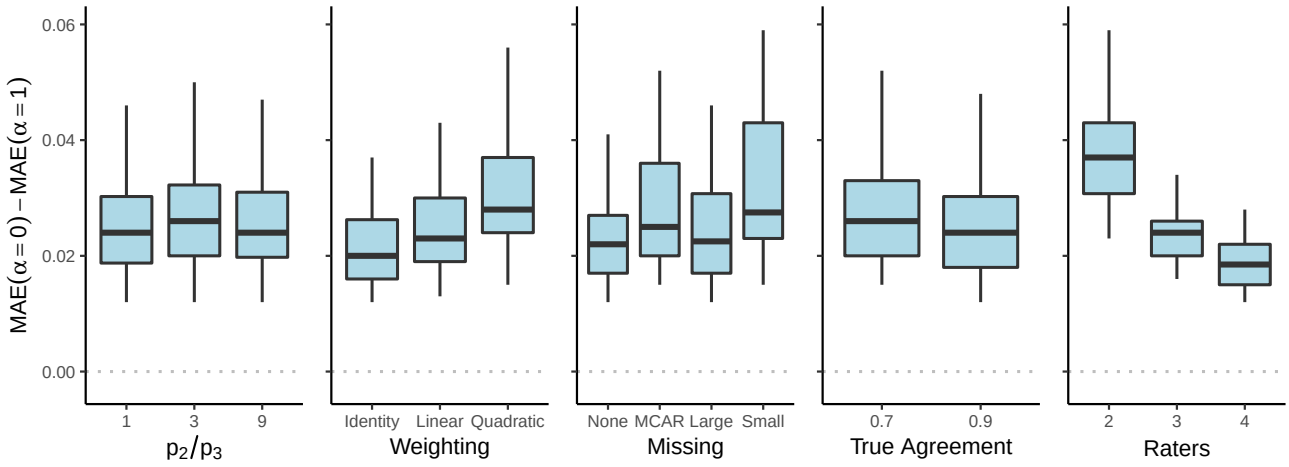

Figure 4

Distributions of marginalized simulation results comparing Fleiss' kappa ($\alpha_c = 0$) *and the uniform prior coefficient* ($\alpha = 1$) *at different small category proportions, weighting schemes, missing data mechanisms, true chance-corrected agreement levels, and number of raters.* $N = 50$, $p_1 = .90$.

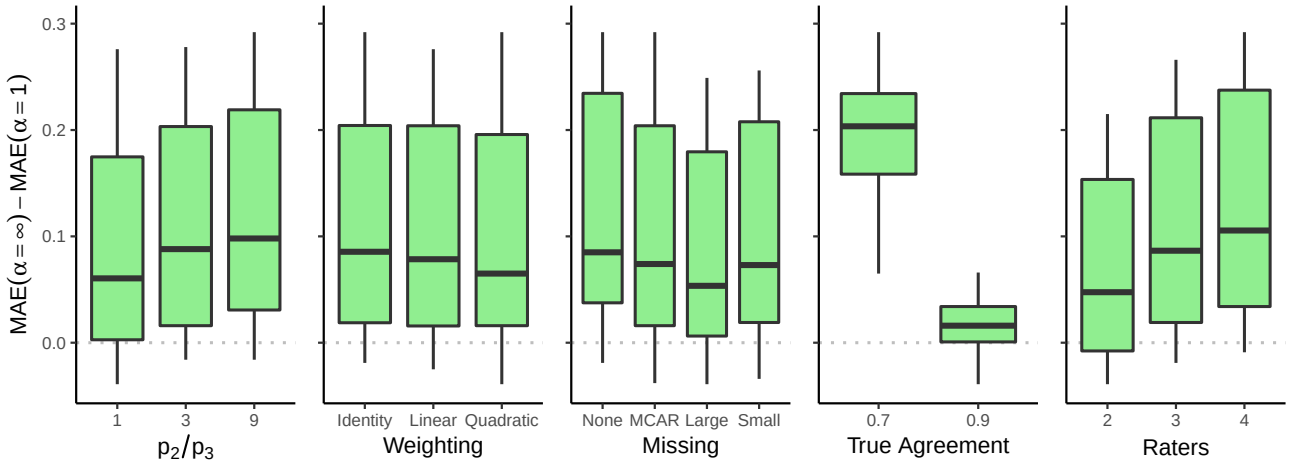

Figure 5

Distributions of marginalized simulation results comparing the S-coefficient ($\alpha_c = \infty$) *and the uniform prior coefficient* ($\alpha = 1$) *at different small category proportions, weighting schemes, missing data mechanisms, true chance-corrected agreement levels, and number of raters.* $N = 50$, $p_1 = .90$.

sample size and make the relative category frequencies less precise, the S-coefficient is "immune" because it does not rely extensively on these relative frequencies. In particular, the second missingness mechanism reduces the gap between the S-coefficient and the uniform prior coefficient because missing data in the large category make the observed category frequencies more balanced and the coefficients more similar.

True Chance-Corrected Agreement

Figures 4 and 5 show that the performance of the uniform prior coefficient improves relative to both alternative coefficients as the true chance-corrected agreement decreases from high (i.e., $I_r = .90$ or $I_r^2 = .81$) to moderate (i.e., $I_r = .70$ or $I_r^2 = .49$). Thus, the uniform prior coefficient appears particularly useful when it is unclear whether the level of chance-corrected agreement is high enough to trust the

rater-based data. Although the true chance-corrected agreement is relatively unimportant when comparing Fleiss' kappa and the uniform prior coefficient, moderate levels of chance-corrected agreement affect the S-coefficient strongly.

Number of Raters

Figures 4 and 5 show that the uniform prior coefficient improves relative to Fleiss' kappa as the number of raters decreases and improves relative to the S-coefficient as the number of raters increases. The largest changes occur between two and three raters. Because both the number of raters (R) and items (N) increase the number of available item classifications, their effects are directionally similar.

Real-World Data Example

We illustrate our approach in an application with ordered categories and incomplete data. The data came from a project studying the expression of positive and negative emotions in online photographs of celebrities from different countries. Furthermore, the study aimed to validate a computer vision algorithm designed to estimate the positivity and negativity of facial expressions (McDuff & Girard, 2019).

The study used five participants ($R = 5$) to rate over one hundred images ($N = 110$) on two scales: (i) how positive is the expression in this image, and (ii) how negative is the expression in this image. Both scales used six ordered categorical response options ($C = 6$), ranging from 0 (*very little or not at all*) to 5 (*extremely*). The study implemented a planned missing design, where all five participants rated 10 random images, and two participants rated the other 100 images. Because the design was balanced, each participant rated 50 images.

Figure 6 shows that the category frequencies differ greatly between the two rating scales: These frequencies are approximately balanced for the positive emotion rating scale but quite unbalanced for the negative emotion rating scale. We use the generalized Bayesian coefficient to compute chance-corrected agreement for different combinations of weighting schemes (i.e., identity, linear, and quadratic) and α_c values (i.e., 0 for Fleiss' kappa, 1 for the uniform prior coefficient, and α for the S-coefficient).

Table 6 reports the computed coefficient values, bootstrapped standard errors, and bias-corrected and accelerated (BCa) confidence intervals, based on 100,000 resamples; Figure 7 visualizes the coefficient values. The estimated chance-corrected agreement increases from identity to linear to quadratic weights and from Fleiss' kappa to the uniform prior coefficient to the S-coefficient. Furthermore, the differences between the three coefficients become more pronounced as the weighting scheme rewards rater disagreements more by moving from identity to linear to quadratic weights. The effects are particularly prominent for the unbalanced negative emotion rating scale, whereas they are weaker

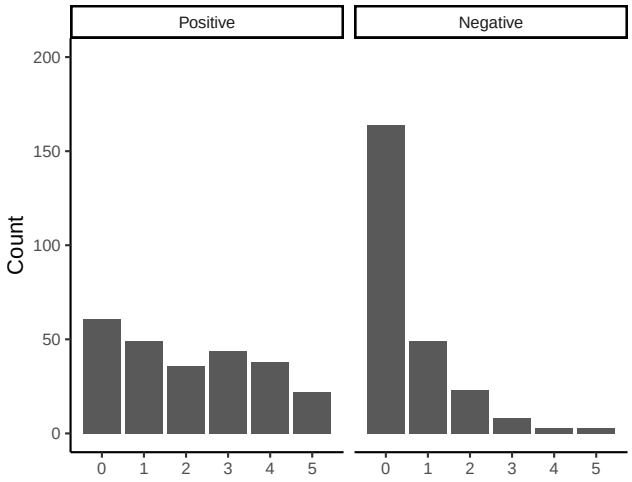

Figure 6

Category frequencies for positive and negative emotion rating scales in real-world data example.

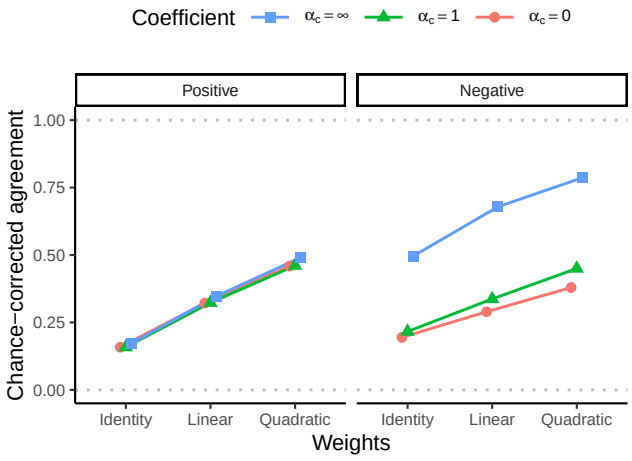

Figure 7

Chance-corrected agreement for positive and negative emotion rating scales, computed from the generalized Bayesian coefficient for different combinations of weighting schemes (i.e., identity, linear, and quadratic) and α_c values (i.e., 0 for Fleiss' kappa, 1 for the uniform prior coefficient, and ∞ for the S-coefficient).

for the almost balanced positive emotion rating scale. The S-coefficient is much higher than the other two coefficients in the negative emotion rating scale due to the violation of the coefficient's assumption that the categories are equally likely. The differences between the values of Fleiss' kappa and the uniform prior coefficient in the negative scale are .021 for identity weights, .047 for linear weights, and .071

Table 6

Bootstrapped standard errors and confidence intervals in real-world data example, based on 100,000 resamples

Variable	Weighting	Coefficient	Value	SE	95% CI
Positive	Identity	$\alpha_c = 0$	0.158	0.059	[0.064, 0.308]
	Identity	$\alpha_c = 1$	0.159	0.059	[0.065, 0.307]
	Identity	$\alpha_c = \infty$	0.172	0.060	[0.077, 0.323]
	Linear	$\alpha_c = 0$	0.322	0.075	[0.184, 0.478]
	Linear	$\alpha_c = 1$	0.323	0.075	[0.187, 0.478]
	Linear	$\alpha_c = \infty$	0.347	0.064	[0.227, 0.479]
	Quadratic	$\alpha_c = 0$	0.459	0.087	[0.286, 0.623]
	Quadratic	$\alpha_c = 1$	0.461	0.087	[0.287, 0.623]
	Quadratic	$\alpha_c = \infty$	0.491	0.070	[0.343, 0.618]
Negative	Identity	$\alpha_c = 0$	0.195	0.081	[0.050, 0.371]
	Identity	$\alpha_c = 1$	0.216	0.080	[0.075, 0.389]
	Identity	$\alpha_c = \infty$	0.496	0.069	[0.366, 0.635]
	Linear	$\alpha_c = 0$	0.290	0.081	[0.138, 0.454]
	Linear	$\alpha_c = 1$	0.337	0.076	[0.191, 0.486]
	Linear	$\alpha_c = \infty$	0.679	0.052	[0.559, 0.767]
	Quadratic	$\alpha_c = 0$	0.379	0.097	[0.195, 0.567]
	Quadratic	$\alpha_c = 1$	0.450	0.084	[0.278, 0.604]
	Quadratic	$\alpha_c = \infty$	0.787	0.050	[0.661, 0.863]

Note. Fleiss' kappa corresponds to $\alpha_c = 0$; the uniform prior coefficient corresponds to $\alpha_c = 1$; the S-coefficient corresponds to $\alpha_c = \infty$.

for quadratic weights. In this unbalanced scale, the uniform prior coefficient has a lower standard error than Fleiss' kappa, in particular, for linear and quadratic weights. However, the S-coefficient tends to have the lowest standard error of all three coefficients, which we conjecture is because its chance correction does not depend on the observed data.

The reported coefficient values are often low for both rating scales. Furthermore, the 95% confidence intervals for Fleiss' kappa and the uniform prior coefficient do not contain values greater than .39 for identity weights, values greater than .49 for linear weights, and values greater than .63 for quadratic weights. We conclude that the ratings of a single randomly-selected participant are not reliable enough to be used interchangeably for either rating scale. Potential next steps include better training of participants or studying potential sources of heterogeneity in the ratings (e.g., to see whether aspects of the participants' backgrounds explain this variability).

Discussion

Van Oest (2019) presented a model-based framework for estimating chance-corrected interrater agreement that encompasses existing coefficients and developed a new coefficient based on Bayesian estimation of the category proportions. Whereas this framework focused on unordered (i.e., nominal) categories and complete data, the present framework accommodates all four combinations of unordered or ordered categories and complete or incomplete data. This extension greatly enhances applicability in real-world settings. Besides incorporating weighting schemes to allow for partial credit when raters choose different categories, the generalized framework enables item classifications to be missing either by circumstance (e.g., technical error or rater attrition) or by design (i.e., planned missing data; Graham, Taylor, Olchowski, & Cumsille, 2006). In the latter situation, the item-by-category matrix may have intentional (and random) holes such that raters classify only subsets of items. Such a design reduces the rater burden and increases the cost efficiency of the study. It would flexibly allow some items to be classified by all raters (for comparison) and other items by different subsets of raters (for efficiency).

The extended framework resulted in a generalized Bayesian coefficient that nests both category-based and distribution-based coefficients, including weighted versions with possibly incomplete data. Furthermore, it resulted in the generalized uniform prior coefficient. A simulation showed that this hybrid coefficient is as good as its "pure" counterparts if no dominating category exists and tends to perform better than these alternatives in the presence of a dominating category. Compared to Scott's pi and Fleiss' kappa (i.e.,

distribution-based coefficients), the benefit was particularly prominent for generous (e.g., quadratic) weighting schemes and missing data; missingness biases either reinforced or attenuated this benefit. However, generous weighting schemes and missing data reduced the gap with the category-based S-coefficient. Although the uniform prior coefficient usually performed much better than the S-coefficient, the S-coefficient sometimes performed better for small sample sizes, incomplete data, and generous weighting schemes.

The literature has produced extremely diverse views on which chance-corrected agreement coefficient is best, and our study does not pretend to settle the issue. Although the simulation results provided support for the hybrid uniform prior coefficient, researchers can also use our framework if they believe that a pure category-based or distribution-based coefficient is more appropriate. The function to compute the generalized Bayesian coefficient consists of only a few lines of code (see Appendix B for this code in R) and encompasses the S-coefficient and equivalent coefficients, Scott's pi, and Fleiss' kappa, all augmented with a weighting scheme and the option of incomplete data. The S-coefficient is strongly connected with the proportion of observed agreement (via a linear transformation), and the values of Scott's pi are usually similar to Cohen's kappa, the most common chance-corrected agreement coefficient (Lombard, Snyder-Duch, & Bracken, 2002; Stemler & Tsai, 2008), with Fleiss' kappa acting as an extension of Cohen's kappa to more than two raters (Fleiss, 1971). Additionally, the *agreement* software package by Girard (2020) provides a comprehensive suite of R functions for working with these and other agreement coefficients (e.g., tidying data for analysis, generating weight matrices, constructing bootstrapped confidence intervals, and visualizing the results).

Besides offering a widely applicable coefficient, our framework implies a new interpretation of chance-corrected weighted agreement coefficients. We showed that these coefficients estimate the probability that both raters in a pair assign an item to its correct category without guessing. Because this result does not depend on the specific weighting scheme, weighted and unweighted agreement coefficients share the same interpretation. This provides conceptual support for the idea that "the interpretation of the magnitude of weighted kappa is like that of unweighted kappa" (Fleiss et al., 2003, p. 609), implying that researchers can use the same standard reference tables as a starting point (e.g., Landis & Koch, 1977). However, usual caution applies. Quadratic weights are most common to assign partial credit in chance-corrected agreement (Vanbelle, 2016) but often result in relatively high coefficient values that are sensitive to the number of categories (Brenner & Kliebsch, 1996; Warrens, 2012, 2013). Because the choice of the weighting scheme is arbitrary, we recommend reporting coefficient values for multiple weighting schemes. The proposed framework accommodates any weighting scheme for category-based, distribution-based, and hybrid coefficients.

References

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23. doi: 10/cvrgb2

Basu, S., Banerjee, M., & Sen, A. (2000). Bayesian Inference for Kappa from Single and Multiple Studies. *Biometrics*, *56*(2), 577–582. doi: 10/fsfnw8

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communication through limited response questioning. *The Public Opinion Quarterly*, *18*(3), 303–308. doi: 10/brfm77

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. doi: 10/d22q4b

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*(2), 199–202. doi: 10/b76mgk

Broemeling, L. D. (2001). A Bayesian analysis for inter-rater agreement. *Communications in Statistics - Simulation and Computation*, *30*(3), 437–446. doi: 10/ftk2sw

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*(3), 101–110. doi: 10/ggw9hx

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi: 10/dghsrr

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. doi: 10/dpbw5f

De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2019). Kappa coefficients for missing data. *Educational and Psychological Measurement*, *79*(3), 558–576. doi: 10/ggdf54

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26. doi: 10/dj84pt

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185. doi: 10/db5gw5

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. doi: 10/bzhdfc

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measuremen*, *33*, 613–619. doi: 10/c29gt3

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 323–327. doi: 10/cc5hq2

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Gajewski, B. J., Hart, S., Bergquist-Beringer, S., & Dunton, N. (2007). Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine*, *26*(25), 4602–4618. doi: 10/bkrmmj

Girard, J. M. (2020). *Agreement: An R package for the tidy analysis of agreement and reliability.*

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. doi: 10/d7v7nv

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (Fourth ed.). Gaithersburg, MD: Advanced Analytics.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433. doi: 10/frt2c6

Kvålseth, T. O. (2018). An alternative interpretation of the linearly weighted kappa coefficients for ordinal data. *Psychometrika*, *83*(3), 618–627. doi: 10/ggw9nf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi: 10/dtzfj3

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. doi: 10/fjbch6

McDuff, D., & Girard, J. M. (2019). Democratizing psychological insights from analysis of nonverbal behavior. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 220–226). Cambridge, UK: IEEE. doi: 10/gjr2v4

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, *26*(2), 135–135. doi: 10/fj4cb3

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*(2), 243–253. doi: 10/btkb3g

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scaling. *Public Opinion Quarterly*, *19*(3), 321–325. doi: 10/bzw9xp

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1–11. doi: 10/ggw9n7

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.

Van Oest, R. (2019). A new coefficient of interrater agreement: The challenge of highly unequal category proportions. *Psychological Methods*, *24*(4), 439–451. doi: 10/ggbk3f

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, *81*(2), 399–410. doi: 10/f8rfdt

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*(2), 157–163. doi: 10/drp7vh

Warrens, M. J. (2011). Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. *Psychometrika*, *76*(3), 471–486. doi: 10/dwdcx9

Warrens, M. J. (2012). Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology*, *9*(3), 440–444. doi: 10/cbq59k

Warrens, M. J. (2013). Conditional inequalities between Cohen's kappa and weighted kappas. *Statistical Methodology*, *10*(1), 14–22. doi: 10/gjr2pb

Zhang, X., & Cutter, G. (2009). A Bayesian Method of Estimating Kappa Coefficient with Application to a Rheumatoid Arthritis Study. *Communications in Statistics - Theory and Methods*, *38*(18), 3432–3444. doi: 10/bpvr4w

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*(3), 374–378. doi: 10/cjgw9s

Appendix A

We consider $R \geq 2$ raters who assign an item to one of C mutually exclusive categories. The expected proportion of weighted agreement, A_w, is the expected number of weighted agreements across all rater pairs, divided by the corresponding maximum, $\binom{R}{2}$. The number of raters making accurate judgments about the item's correct category may vary from zero to R. The probability that $j \in \{0, \ldots, R\}$ raters make accurate judgments (and thus $R - j$ raters need to guess) immediately follows from the binomial distribution: $\binom{R}{j} I_r^j (1 - I_r)^{R-j}$.

If j raters make accurate judgments and the item's correct category is c, the j accurate raters contribute $\binom{j}{2}$ pairwise agreements, each of j accurate raters (choosing c) paired with each of the $R - j$ inaccurate raters (with guessing probabilities $p_1, \ldots, p_C$) contributes $\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_{\tilde{c}}$, and each of the $\binom{R-j}{2}$ pairs of guessing raters contributes $\sum_{\bar{c}=1}^{C} \sum_{\tilde{c}=1}^{C} w_{\bar{c},\tilde{c}} p_{\bar{c}} p_{\tilde{c}}$ (not depending on the correct category c). Thus, conditional on j accurate raters and correct category c, the expected number of weighted agreements across all rater pairs is $\binom{j}{2} + j(R-j) \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_{\tilde{c}} + \binom{R-j}{2} \sum_{\bar{c}=1}^{C} \sum_{\tilde{c}=1}^{C} w_{\bar{c},\tilde{c}} p_{\bar{c}} p_{\tilde{c}}$. Dividing this number by the corresponding maximum, $\binom{R}{2}$, yields A_w conditional on j accurate raters and correct category c:

$$A_w|j, c = \frac{\binom{j}{2} + j(R-j) \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_{\tilde{c}} + \binom{R-j}{2} \sum_{\bar{c}=1}^{C} \sum_{\tilde{c}=1}^{C} w_{\bar{c},\tilde{c}} p_{\bar{c}} p_{\tilde{c}}}{\binom{R}{2}},$$

where we follow the standard convention for binomial coefficients that $\binom{n}{k} = 0$ if $n < k$, where n and k are nonnegative integers.

By taking the expectation over the probability distribution of correct categories, $(p_1, \ldots, p_C)$, and using that $\sum_{c=1}^{C} p_c = 1$, we obtain

$$\begin{aligned} A_w|j &= \sum_{c=1}^{c} p_c \left[\frac{\binom{j}{2} + j(R-j) \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_{\tilde{c}} + \binom{R-j}{2} \sum_{\bar{c}=1}^{C} \sum_{\tilde{c}=1}^{C} w_{\bar{c},\tilde{c}} p_{\bar{c}} p_{\tilde{c}}}{\binom{R}{2}} \right] \\ &= \frac{\binom{j}{2} + j(R-j) \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} + \binom{R-j}{2} \sum_{\bar{c}=1}^{C} \sum_{\tilde{c}=1}^{C} w_{\bar{c},\tilde{c}} p_{\bar{c}} p_{\tilde{c}}}{\binom{R}{2}} \\ &= \frac{\binom{j}{2}}{\binom{R}{2}} + \left\{ \frac{j(R-j)}{\binom{R}{2}} + \frac{\binom{R-j}{2}}{\binom{R}{2}} \right\} \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}. \end{aligned}$$

By taking the expectation over the (binomial) probability distribution of j, we obtain

$$\begin{aligned} A_w &= \sum_{j=1}^{R} \binom{R}{j} I_r^j (1 - I_r)^{R-j} \left[\frac{\binom{j}{2}}{\binom{R}{2}} + \left\{ \frac{j(R-j)}{\binom{R}{2}} + \frac{\binom{R-j}{2}}{\binom{R}{2}} \right\} \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} \right] \\ &= \sum_{j=0}^{R} \left[\frac{\binom{R}{j}\binom{j}{2}}{\binom{R}{2}} I_r^j (1 - I_r)^{R-j} + \left\{ \frac{\binom{R}{j} j(R-j)}{\binom{R}{2}} + \frac{\binom{R}{j}\binom{R-j}{2}}{\binom{R}{2}} \right\} I_r^j (1 - I_r)^{R-j} \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} \right]. \end{aligned}$$

Breaking up the expression within square brackets, using that $j(R - j) = 0$ if either $j = 0$ or $j = R$, and using that $\binom{n}{k} = 0$ if $n < k$ reduces the number of terms in the summations:

$$\begin{aligned} A_W = \sum_{j=2}^{R} \left[\frac{\binom{R}{j}\binom{j}{2}}{\binom{R}{2}} I_r^j (1 - I_r)^{R-j} \right] + \sum_{j=1}^{R-1} \left[\frac{\binom{R}{j} j(R-j)}{\binom{R}{2}} I_r^j (1 - I_r)^{R-j} \right] \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}} + \\ \sum_{j=0}^{R-2} \left[\frac{\binom{R}{j}\binom{R-j}{2}}{\binom{R}{2}} I_r^j (1 - I_r)^{R-j} \right] \sum_{c=1}^{C} \sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}} p_c p_{\tilde{c}}. \end{aligned}$$

Because the fractions containing binomial coefficients simplify into

$$\frac{\binom{R}{j}\binom{j}{2}}{\binom{R}{2}} = \frac{\frac{R!}{j!(R-j)!}\frac{j!}{2!(j-2)!}}{\frac{R!}{2!(R-2)!}} = \frac{(R-2)!}{(R-j)!(j-2)!} = \binom{R-2}{j-2},$$
$$\frac{\binom{R}{j}j(R-j)}{\binom{R}{2}} = \frac{\frac{R!\,j(R-j)}{j!(R-j)!}}{\frac{R!}{2!(R-2)!}} = \frac{2(R-2)!}{(j-1)!(R-j-1)!} = 2\binom{R-2}{j-1},$$
$$\frac{\binom{R}{j}\binom{R-j}{2}}{\binom{R}{2}} = \frac{\frac{R!}{j!(R-j)!}\frac{(R-j)!}{2!(R-j-2)!}}{\frac{R!}{2!(R-2)!}} = \frac{(R-2)!}{j!(R-j-2)!} = \binom{R-2}{j},$$

we can rewrite A_w as

$$A_w = \sum_{j=2}^{R}\left[\binom{R-2}{j-2}I_r^j(1-I_r)^{R-j}\right] + \sum_{j=1}^{R-1}\left[2\binom{R-2}{j-1}I_r^j(1-I_r)^{R-j}\right]\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}} + \sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^j(1-I_r)^{R-j}\right]\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}.$$

Manipulating the indices j in the first and second summations (without altering the summations themselves) yields

$$A_w = \sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^{j+2}(1-I_r)^{R-j-2}\right] + \sum_{j=0}^{R-2}\left[2\binom{R-2}{j}I_r^{j+1}(1-I_r)^{R-j-1}\right]\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}} + \sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^{j}(1-I_r)^{R-j}\right]\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}},$$

which we further rewrite as

$$A_w = I_r^2\sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^{j}(1-I_r)^{R-2-j}\right] + 2I_r(1-I_r)\sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^{j}(1-I_r)^{R-2-j}\right]\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}} + (1-I_r)^2\sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^{j}(1-I_r)^{R-2-j}\right]\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}.$$

By recognizing the sum of all probabilities for the binomial distribution with parameters $R-2$ and I_r, that is, $\sum_{j=0}^{R-2}\left[\binom{R-2}{j}I_r^j(1-I_r)^{R-2-j}\right] = 1$, we obtain

$$\begin{aligned} A_w &= I_r^2 + 2I_r(1-I_r)\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}} + (1-I_r)^2\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}} \\ &= I_r^2 + (1-I_r^2)\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}. \end{aligned}$$

For $A_w \geq \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}$, there is a solution for I_r^2 that satisfies $I_r^2 \geq 0$:

$$I_r^2 = \frac{A_w - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}}{1 - \sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} w_{c,\tilde{c}}p_c p_{\tilde{c}}}.$$

Appendix B

The function to compute the generalized Bayesian agreement coefficient, defined by (8) and (12), requires three inputs: the $(N \times C)$ data matrix `Rmat`, where element (i, c) is the number of raters who assigned item $i \in \{1, \ldots, N\}$ to category $c \in \{1, \ldots, C\}$, the $(C \times C)$ symmetric weight matrix `W`, and the $(1 \times C)$ vector `alpha` containing the Dirichlet parameters describing the prior distribution of the category proportions.

```
Coefficient <- function( Rmat, W, alpha )
{
  Ri <-         rowSums( Rmat )
  Aw_top <-     sum( rowSums( Rmat * ( Rmat %*% W - 1 ) ) )
  Aw_bottom <-  sum( Ri * ( Ri - 1 ) )
  Aw <-         Aw_top / Aw_bottom
  p <-          ( alpha + colSums( Rmat ) ) / ( sum( alpha ) + sum( Ri ) )
  pWp <-        p %*% W %*% t(p)
  coeff <-      ( Aw - pWp ) / ( 1 - pWp )
  coeff
}
```

Next, we enter the function's inputs: (i) the data matrix, (ii) the various weight matrices (i.e., identity, linear, and quadratic), and (iii) the alpha coefficients corresponding to Fleiss' kappa, the uniform prior coefficient, and the S-coefficient.

```
Rmat <- matrix( c( 3, 2, 4, 2, 3, 0, 4, 2, 4, 3, 1, 3, 0, 4, 1,
                   1, 1, 3, 0, 1, 3, 3, 3, 4, 1, 4, 0, 2, 3, 1,
                   0, 1, 0, 0, 1, 3, 0, 0, 0, 0, 0, 0, 1, 0, 2,
                   1, 0, 0, 0, 0, 0, 1, 0, 0, 3, 0, 0, 1, 0, 2,
                   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,
                   1, 3, 0, 4, 2, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0 ), ncol=3 )
Rmat
W_identity <-       matrix( c( 1.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 0.00, 1.00 ), ncol=3 )
W_linear <-         matrix( c( 1.00, 0.50, 0.00, 0.50, 1.00, 0.50, 0.00, 0.50, 1.00 ), ncol=3 )
W_quadratic <-      matrix( c( 1.00, 0.75, 0.00, 0.75, 1.00, 0.75, 0.00, 0.75, 1.00 ), ncol=3 )
alpha_FleissK <-    t( c( 0, 0, 0 ) )
alpha_UniPrior <-   t( c( 1, 1, 1 ) )
alpha_SCoeff <-     t( c( 1000000, 1000000, 1000000 ) )
```

Finally, we compute and print Fleiss' kappa, the uniform prior coefficient, and the S-coefficient for the three different weight matrices (i.e., identity, linear, and quadratic).

```
( FleissK_identity <-   Coefficient( Rmat, W_identity, alpha_FleissK ) )    # 0.4677686
( UniPrior_identity <-  Coefficient( Rmat, W_identity, alpha_UniPrior ) )   # 0.4792173
( SCoeff_identity <-    Coefficient( Rmat, W_identity, alpha_SCoeff ) )     # 0.6120690
( FleissK_linear <-     Coefficient( Rmat, W_linear, alpha_FleissK ) )      # 0.5048103
( UniPrior_linear <-    Coefficient( Rmat, W_linear, alpha_UniPrior ) )     # 0.5150104
( SCoeff_linear <-      Coefficient( Rmat, W_linear, alpha_SCoeff ) )       # 0.6120705
( FleissK_quadratic <-  Coefficient( Rmat, W_quadratic, alpha_FleissK ) )   # 0.5370316
( UniPrior_quadratic <- Coefficient( Rmat, W_quadratic, alpha_UniPrior ) )  # 0.5461999
( SCoeff_quadratic <-   Coefficient( Rmat, W_quadratic, alpha_SCoeff ) )    # 0.6120721
```