Submitted to Statistical Science

Bayesian Gower Agreement for Categorical Data

John Hughes

Abstract. In this work I present two methods for measuring agreement in nominal and ordinal data. The measures, which employ Gower-type distances, are simple, intuitive, and easy to compute for any number of units and any number of coders. Missingness is easily accommodated. Influential units and/or coders are easily identified. I consider both one-way and two-way random sampling designs, and develop an approach to Bayesian inference for each. I apply the methods to simulated data and to two real datasets, the first from a one-way radiological study of congenital diaphragmatic hernia, and the second from a two-way study of psychiatric diagnosis.

Key words and phrases: agreement, Bayesian bootstrap, categorical data, Gower distance, inter-rater reliability.

1. INTRODUCTION

An inter-coder or intra-coder agreement coefficient, which takes a value in the unit interval, is a statistical measure of the extent to which two or more coders agree regarding the same units of analysis. The agreement problem has a long history and is important in many fields of inquiry, and numerous agreement statistics have been proposed.

The first agreement coefficients were S [3], π [25], and κ [5]. [3] proposed the S score as a measure of the extent to which two methods of communication provide identical information. [25] proposed the π coefficient for measuring agreement between two coders. [5] criticized π and proposed the κ coefficient as an alternative to π—although [27] noted that Francis Galton mentioned a κ-like statistic in his 1892 book, *Finger Prints*. [10] proposed multi-κ, a generalization of Scott's π for measuring agreement among three or more coders. [7] and [8] likewise generalized κ to the multi-coder setting. Other generalizations of κ, e.g., weighted κ [6], have also been proposed. The κ coefficient and its generalizations can fairly be said to dominate the field and are still widely used despite their well-known shortcomings [9, 4]. Other frequently used measures of agreement are Gwet's AC_1 and AC_2 [13] and Krippendorff's α [15]. For more comprehensive reviews of the literature on agreement, I refer the interested reader to the article by [2], the article by [1], and the book by [14].

John Hughes is Associate Professor, College of Health, Lehigh University, Bethlehem, USA (e-mail: drjphughesjr@gmail.com; URL: www.johnhughes.org).

In this article I present new means of measuring agreement for nominal and ordinal data, and develop corresponding methods for Bayesian inference for both one-way random designs (units are random, coders are fixed) and two-way random designs (both units and coders are random). In Section 2 I describe the agreement measures. In Section 3 I propose algorithms for sampling from the posterior distribution of the parameter of interest. In Section 4 I evaluate the two methodologies by applying them to simulated data. In Section 5 I apply the methods to two real datasets. In Section 6 I propose a method for obtaining a calibrated agreement scale for a given dataset, distance function, and sampling model. I make concluding remarks in Section 7.

2. GOWER-TYPE AGREEMENT MEASURES FOR NOMINAL AND ORDINAL DATA

Suppose the data X_{ij} are arranged in $n \times m$ matrix $\mathbf{X}$, where n is the number of units and m is the number of coders. Then X_{ij} is the score assigned by coder j to unit i. The building blocks of the proposed agreement measure are the row statistics

$$G_i = 1 - \frac{1}{\binom{m}{2}} \sum_{j<k} d(X_{ij}, X_{ik}),$$

where d is an appropriate distance function. For nominal data I recommend the discrete metric $d(x, y) = I\{x \neq y\}$, where I denotes the indicator function. For ordinal data I recommend the L_1 distance function given by

$$d(x, y) = \frac{|x - y|}{r},$$

where r is the range of the scores, e.g., $r = C - 1$ for scores in $\{1, 2, \ldots, C\}$.

When d is the discrete metric the distances are of course binary, and their sum is not even approximately binomial unless the intra-row dependence is very weak. This is not surprising given theoretical work regarding sums of dependent Bernoulli variables [11]. In any case, each row statistic is a Gower-type [12] measure of agreement for the row in question, and the row statistics are an identically distributed sample from a discrete distribution having its points of support in the unit interval. This distribution is determined by the marginal distribution of the scores, the dependence structure, the number of coders, and the distance function. The mean of this distribution, μ_g, say, is the proposed measure of agreement for the study. (Although Gower distance was discovered long ago, these measures do not, to my knowledge, appear in the agreement literature, nor has Bayesian inference been considered in this context.)

One can estimate μ_g as the sample mean of the G_i:

$$\hat{\mu}_g = \bar{G} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1}{\binom{m}{2}}\sum_{j<k} I\{X_{ij} \neq X_{ik}\}\right)$$

for nominal data, and

$$\hat{\mu}_g = \bar{G} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1}{\binom{m}{2}}\sum_{j<k} \frac{|X_{ij} - X_{ik}|}{r}\right)$$

for ordinal data. For a one-way design, wherein the units are random but the coders are fixed, the row-wise agreement statistics are iid, and so the ordinary central limit theorem applies: $\hat{\mu}_g \dot{\sim} \text{NORMAL}(\mu_g, \sigma_g^2/n)$, where σ_g^2 is the variance of G. But I prefer to do Bayesian inference for μ_g. In the next section I develop a Bayesian bootstrap for both the one-way design and the two-way design. These algorithms produce a sample from posterior distribution $\pi(\mu_g \mid \mathbf{X})$ so that $\mathbb{E}_\pi(\mu_g \mid \mathbf{X})$ is the agreement measure, which can be estimated as the mean of the posterior sample.

Note that this approach yields a measure of agreement for each unit (G_i) as well as a measure of agreement for the study ($\mathbb{E}_\pi(\mu_g \mid \mathbf{X})$). It is easy to accommodate any number of units and any number of coders, and missing scores can be handled by simply skipping them when computing the row statistics. Any row having just a single score is removed prior to analysis since such a row carries no information about agreement.

3. BAYESIAN INFERENCE

For a one-way study it is straightforward to adapt the Bayesian bootstrap to this setting. For two-way studies I develop a new Bayesian bootstrap based on pigeonhole resampling. These algorithms allow one to sample directly from the posterior distribution of μ_g. Here I specify the algorithms. In the next section I evaluate their performance in a Monte Carlo study.

3.1 Bayesian Bootstrap for a One-Way Random Design

For a study in which the units are random and the coders are fixed, one can employ a Bayesian bootstrap [24] to draw samples from $\pi(\mu_g \mid \mathbf{X})$ in the following way.

1. Compute the row statistics $\boldsymbol{G} = (G_1, \ldots, G_n)'$.
2. Repeat for $b = 1, 2, \ldots, B$:
   a) Draw $\boldsymbol{U} = (U_1, \ldots, U_{n-1})'$ iid $\text{UNIFORM}(0, 1)$.
   b) Sort $\boldsymbol{U}$ and form the gap sequence $\boldsymbol{W} = (U_{(1)}, U_{(2)} - U_{(1)}, U_{(3)} - U_{(2)}, \ldots, U_{(n-1)} - U_{(n-2)}, 1 - U_{(n-1)})'$.
   c) Compute $\mu_g^b = \boldsymbol{W} \cdot \boldsymbol{G}$ as the bth sample from $\pi(\mu_g \mid \mathbf{X})$.
3. Use the posterior sample of size B to do inference for μ_g.

This procedure can be carried out efficiently even for a large number of units, and typically only a small posterior sample is required to do reliable inference. In the next section I evaluate the frequentist performance of this approach.

3.2 Bayesian Bootstrap for a Two-Way Random Design

Doing posterior inference for a two-way design is more delicate. The ordinary Bayesian bootstrap is deficient for this purpose because the ordinary method accommodates only one source of random variation, the variation across units. To reflect the randomness of coders as well, one can marry the pigeonhole bootstrap [23] with the Bayesian bootstrap in the following way. To my knowledge this is a new form of Bayesian bootstrap.

1. Repeat for $b = 1, 2, \ldots, B$:
   a) Resample the rows of $\mathbf{X}$ with replacement.
   b) Given the resampled rows, resample the columns of $\mathbf{X}$ with replacement. These first two steps produce $\mathbf{X}^*$.
   c) Compute the row statistics $\boldsymbol{G}^* = (G_1^*, \ldots, G_n^*)'$ from $\mathbf{X}^*$.
   d) Draw $\boldsymbol{U} = (U_1, \ldots, U_{n-1})'$ iid $\text{UNIFORM}(0, 1)$.
   e) Sort $\boldsymbol{U}$ and form the gap sequence $\boldsymbol{W} = (U_{(1)}, U_{(2)} - U_{(1)}, U_{(3)} - U_{(2)}, \ldots, U_{(n-1)} - U_{(n-2)}, 1 - U_{(n-1)})'$.
   f) Compute $\mu_g^b = \boldsymbol{W} \cdot \boldsymbol{G}^*$ as the bth sample from $\pi(\mu_g \mid \mathbf{X})$.
2. Use the posterior sample of size B to do inference for μ_g.

This approach also permits efficient computation and captures well both sources of randomness. In the next section I evaluate the frequentist performance of this method, and compare to the performance of a frequentist pigeonhole bootstrap.

4. APPLICATION TO SIMULATED DATA

For the simulation studies presented in this section I simulated data from direct Gaussian copula models with categorical marginal distributions. These are sensible proxy models since they permit one to specify appropriate correlation matrices for both the one-way design and the two-way design, and then apply those latent dependence structures to categorical outcomes. The generative form of the direct Gaussian copula model is

$$\begin{aligned} \boldsymbol{Z} &\sim \text{NORMAL}(\mathbf{0}, \boldsymbol{\Omega}) \\ U_{ij} &= \Phi(Z_{ij}) \quad (i = 1, \ldots, n)(j = 1, \ldots, m) \\ X_{ij} &= F^{-1}(U_{ij}), \end{aligned}$$

where $\boldsymbol{\Omega}$ is the copula correlation matrix, Φ is the standard Gaussian cdf, and F^{-1} is the quantile function of the desired response distribution. The random vector $\boldsymbol{U}$ is a realization of the copula, and $\mathbf{X}$ is obtained by applying the probability integral transform to the marginally standard uniform U_{ij}.

For the one-way study $\boldsymbol{\Omega}$ is block-diagonal with each block having the compound symmetry structure. I varied the intraclass correlation over the grid $\rho \in (0.01, \ldots, 0.99)$, and simulated 4,000 datasets for each value of ρ. For each simulated dataset I computed a credible interval based on a posterior sample of size 1,000.

For the two-way study the copula correlation matrix is given by the Kronecker product

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}_n(\rho_n) \otimes \boldsymbol{\Omega}_m(\rho_m),$$

where $\boldsymbol{\Omega}_n$ is a compound symmetry structure for the coders and $\boldsymbol{\Omega}_n$ is a compound symmetry structure for the units. I varied the intra-row correlation over the grid $\rho_m \in (0.01, \ldots, 0.99)$, and for each value of ρ_m I used $\rho_n = \rho_m/2$. That is, for each scenario the inter-row correlation was half the intra-row correlation. This seems like a sensible study design since this methodology is most useful when the coders do not exhibit large biases. When intra-coder dependence is stronger than intra-unit dependence, agreement is low, in which case doing inference for μ_g may be of little interest. If, for a given dataset, it seems clear that intra-coder dependence is strong, one might transpose $\mathbf{X}$ and repeat the analysis to get a sense of the strength of intra-coder agreement.

For each of 4,000 simulated datasets I computed a credible interval and a frequentist bootstrap interval, both based on a sample of size of 1,000. The frequentist bootstrap used pigeonhole resampling. It is important to note that no frequentist bootstrap for the two-way design can be exact [22], but the pigeonhole bootstrap is useful for comparison with the Bayesian bootstrap outlined above.

For all scenarios I used $p = (0.1, 0.15, 0.3, 0.4, 0.05)$ for the categorical probabilities, with $X_{ij} \in \{1, \ldots, 5\}$.

4.1 Nominal Data

The coverage profile for the one-way design with 16 units and four coders is shown in Figure 1. We see that the 95% credible interval offers nearly nominal frequentist coverage across the range of latent correlation ρ. This is a small sample geometry. The coverage profile improves as the number of units and/or coders increases.

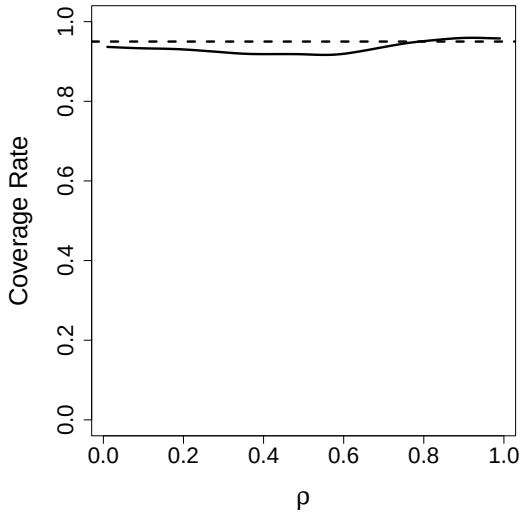

FIG 1. *The coverage profile for the 95% credible interval for nominal data with 16 units and four coders, one-way sampling design.*

The coverage profile for the two-way design with 16 units and four coders is shown in Figure 2. We see that the 95% credible interval offers slightly better than nominal frequentist coverage (solid line) across the range of latent correlation ρ_m. The coverage profile for the pigeonhole bootstrap is shown as a dotted line. Doing Bayesian inference clearly offers a substantial advantage here.

4.2 Ordinal Data

The coverage profile for the one-way design with 16 units and four coders is shown in Figure 3. We see that the 95% credible interval offers nearly nominal frequentist coverage across the range of latent correlation ρ. The coverage profile improves as the number of units and/or coders increases.

The coverage profile for the two-way design with 16 units and four coders is shown in Figure 4. We see that the two-way design presents more of a challenge to the methodology, with the coverage rate dipping as low as

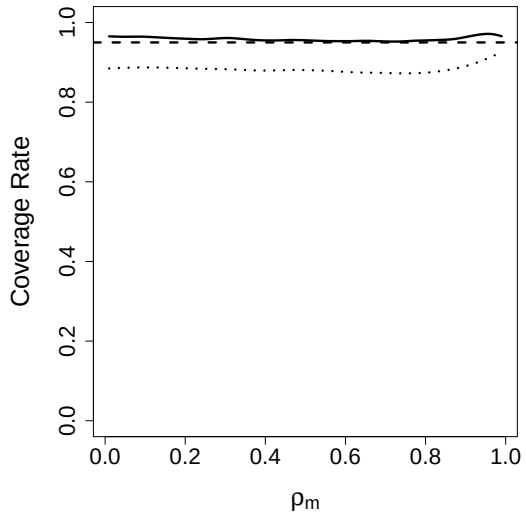

FIG 2. *The coverage profile for the 95% credible interval for nominal data with 16 units and four coders, two-way sampling design.*

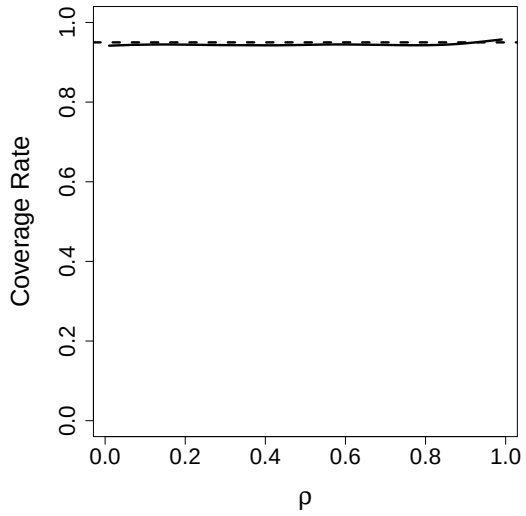

FIG 3. *The coverage profile for the 95% credible interval for ordinal data with 16 units and four coders, one-way sampling design.*

90% for some combinations of ρ_m and ρ_n. But the performance of the credible interval is still very much better than that of the frequentist pigeonhole bootstrap.

5. APPLICATION TO REAL DATA

In this section I apply the proposed methods to two real datasets. The first is from a one-way magnetic resonance imaging study of congenital diaphragmatic hernia. The scores are ordinal. The second dataset is from a two-way study of psychiatric diagnosis. The scores are nominal.

I will interpret results according to the agreement scale given in Table 1 [20]. Although this scale is well-established, agreement scales remain a subject of debate

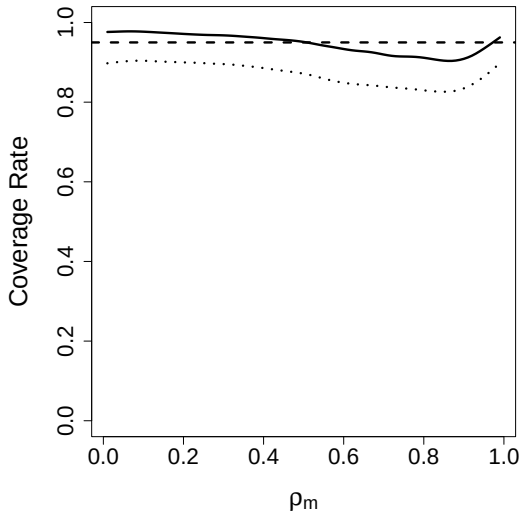

FIG 4. *The coverage profile for the 95% credible interval for ordinal data with 16 units and four coders, two-way sampling design.*

[28], and so the following scale (or any agreement scale) should be applied with caution.

TABLE 1
Guidelines for interpreting values of an agreement coefficient.

Range of Agreement	Interpretation
$\mu_g \leq 0.2$	Slight Agreement
$0.2 < \mu_g \leq 0.4$	Fair Agreement
$0.4 < \mu_g \leq 0.6$	Moderate Agreement
$0.6 < \mu_g \leq 0.8$	Substantial Agreement
$\mu_g > 0.8$	Near-Perfect Agreement

5.1 Ordinal Data from a One-Way Radiological Study of Congenital Diaphragmatic Hernia

The data for this example are liver-herniation scores (in $\{1, \ldots, 5\}$) assigned by two coders (radiologists) to magnetic resonance images of the liver in a study pertaining to congenital diaphragmatic hernia (CDH) [21], in which a hole in the diaphragm permits abdominal organs to enter the chest. The five grades are described in Table 2.

Each radiologist scored each of the 47 images twice, and so we are interested in assessing both intra-coder and inter-coder agreement. This is a one-way study, which is to say we are interested in measuring agreement for these two radiologists, as opposed to considering the radiologists as having been drawn from a larger population. The results are shown in Table 3. We see that both intra-coder and inter-coder agreement are very nearly perfect. Note that each of the execution times was shorter than one second despite my having drawn 10,000 posterior samples for each. The posterior sample for the second radiologist is shown in Figure 5. Superimposed are a kernel density

TABLE 2
Liver herniation grades for the CDH study.

Grade	Description
1	No herniation of liver into the fetal chest
2	Less than half of the ipsilateral thorax is occupied by the fetal liver
3	Greater than half of the thorax is occupied by the fetal liver
4	The liver dome reaches the thoracic apex
5	The liver dome not only reaches the thoracic apex but also extends across the thoracic midline

estimate and the limits of the 95% credible interval (sample quantiles). Since the distribution is markedly skewed to the left, I should report the estimated posterior median: 0.991.

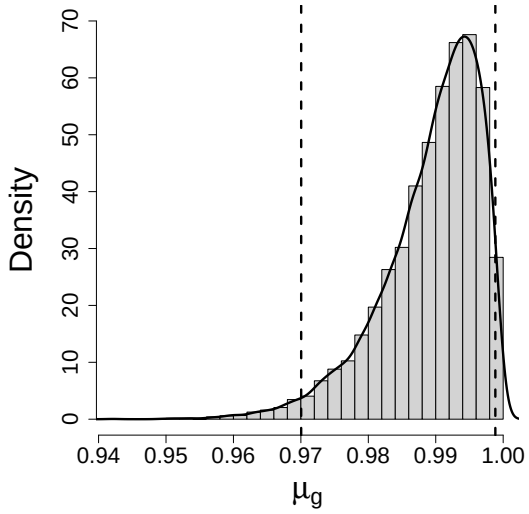

FIG 5. *A histogram of the posterior sample for the second radiologist in the CDH study.*

5.2 Nominal Data from a Two-Way Study of Psychiatric Diagnosis

The data from this study are psychiatric diagnoses (depression, personality disorder, schizophrenia, neurosis, and other) assigned to 30 patients by six raters [10]. I apply the two-way nominal methodology to these data, wherein both patients and raters are assumed to have been sampled from larger populations. The estimated posterior mean is 0.556, and the 95% credible interval is $(0.474, 0.650)$. This is perhaps alarmingly poor agreement (only moderate according to the agreement scale given above) considering the stakes, but, to be fair, these are old data and so do not reflect recent advances in psychiatric diagnosis.

Note that $\pi(\mu_g \mid \mathbf{X})$ is approximately Gaussian for these data (Figure 6), although we see slight asymmetry—the Shapiro–Wilk test [26] rejects the null hypothesis of normality. Also note that Fleiss reported a κ value of 0.430, which is not contained in the credible interval for μ_g. This is close to the Krippendorff's α value of 0.440, but α is inappropriate for these data because that methodology is for one-way designs [19, 18].

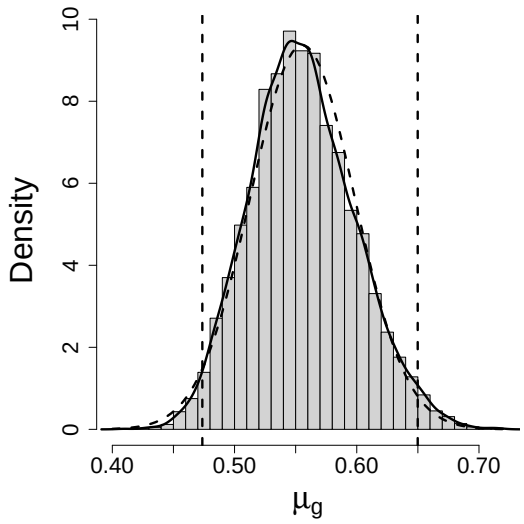

FIG 6. *A histogram of the posterior sample for the psychiatric diagnosis study. A Gaussian density is shown dashed.*

6. AGREEMENT SCALE CALIBRATION

As I carried out the simulation studies for this paper I was able to see how, exactly, the range of possible values of μ_g is constrained by the categorical marginal distribution and the distance function. Knowledge of said range can help one choose an appropriate scale for a given study, if one is willing to posit a direct Gaussian copula model with categorical margins as the data-generating mechanism.

For example, consider the plot in Figure 7, which shows μ_g as a function of latent correlation ρ for the one-way nominal study with four coders. We see that μ_g is constrained to the range $[0.29, 0.89]$. One might use this relationship to devise a linear agreement scale such that $\mu_g \leq 0.41$ represents slight agreement, $0.41 < \mu_g \leq 0.53$

TABLE 3
Results from applying the proposed methodology to the liver data.

	Estimated Posterior Mean	95% Credible Interval
Radiologist 1	0.984	$(0.962, 0.997)$
Radiologist 2	0.989	$(0.970, 0.999)$
Overall	0.972	$(0.954, 0.987)$

represents fair agreement, $0.53 < \mu_g \leq 0.65$ represents moderate agreement, $0.65 < \mu_g \leq 0.77$ represents substantial agreement, and $\mu_g > 0.77$ represents near-perfect agreement. Or one might consider μ_g against Gaussian mutual information, $I(\rho) = -0.5\log(1-\rho^2)$ (see Figure 8). This provides what is perhaps the most sensible scale since mutual information is arguably superior to ρ as a measure of redundancy [29].

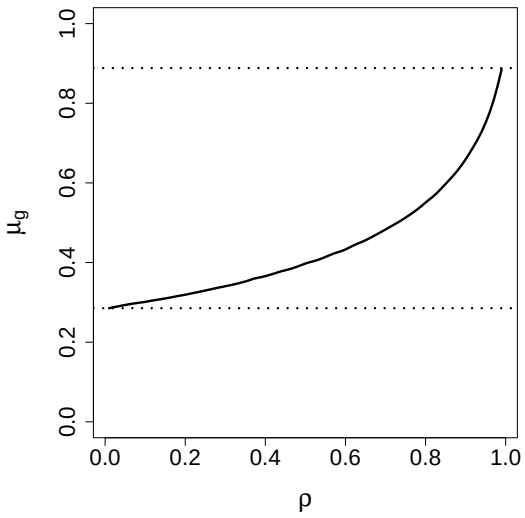

FIG 7. *μ_g as a function of ρ for the one-way nominal simulation study with 16×4 data matrix.*

In any case, the function $\mu_g\{f(\rho)\}$ can be revealed by doing a simulation study wherein the empirical categorical probabilities of the sample are used to generate the outcomes in the Gaussian copula model described earlier. Since $\mu_g\{f(\rho)\}$ is the same for the one-way and two-way designs, a simple one-way simulation can be used for scale calibration in either design.

7. DISCUSSION

Although the discrete metric appears to be an obvious choice of distance function for nominal data, the L_1 distance function is perhaps a less obvious choice for ordinal data. Hence other distance functions might be used for measuring agreement for ordinal scores. One might

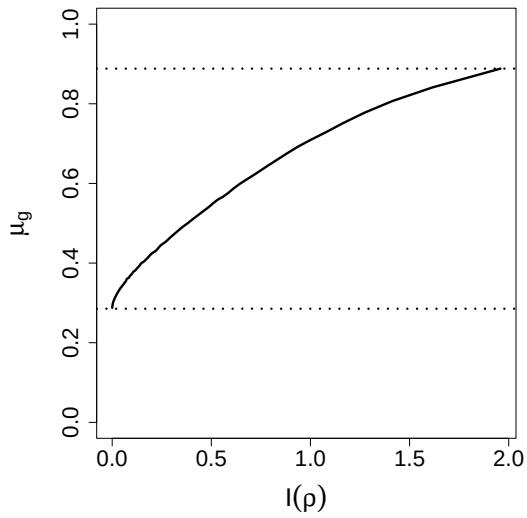

FIG 8. *μ_g as a function of $I(\rho)$ for the one-way nominal simulation study with 16×4 data matrix.*

use the discrete metric for ordinal outcomes, or one might use, for example,

$$G_i = 1 - \frac{\max_{j<k}\{|X_{ij} - X_{ik})|\}}{r}$$

as the measure of agreement for a given row of $\mathbf{X}$. Applying this latter distance measure to the full dataset from the CDH study yields the posterior distribution shown in Figure 9. This distribution has a smaller center and is more symmetric and more dispersed than the posterior obtained by applying the L_1 distance function.

Some readers may wonder, considering that I used the direct Gaussian copula model with discrete margins as a data-generating mechanism, why I developed the methodology presented in this article. The problem with the Gaussian copula model is that the likelihood is intractable for more than a few coders. This makes fully Bayesian analysis impractical for the copula model, whereas fully Bayesian analysis for Gower agreement is straightforward and computationally efficient. Methods for approximate Bayesian analysis have been developed (see, e.g., [17, 16]) for Gaussian copula models with discrete marginals, but the methodology presented here is, in my opinion, at least as compelling. Also, the methods in this article clearly do not assume any particular data-

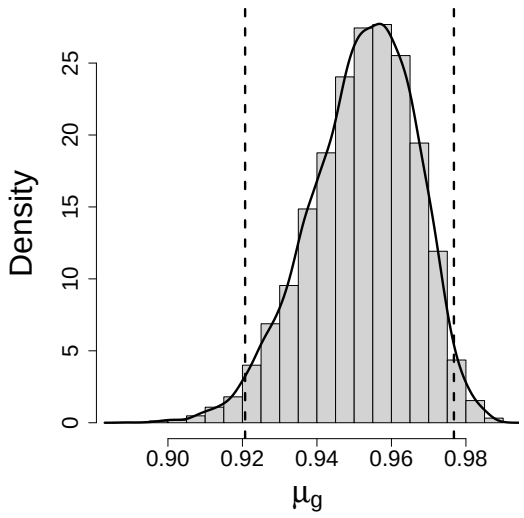

FIG 9. *A histogram of the posterior sample for the overall measure of agreement in the CDH study and using the max norm.*

generating mechanism, only a one-way or two-way study design.

The methodology developed in this paper is supported by R package `goweragreement`, which is freely available on the Comprehensive R Archive Network. The package supports user-supplied distance functions and means (leave-one-out analyses) of identifying influential units and/or coders.

REFERENCES

[1] ARTSTEIN, R. and POESIO, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics* **34** 555–596.

[2] BANERJEE, M., CAPOZZOLI, M., MCSWEENEY, L. and SINHA, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics* **27** 3–23.

[3] BENNETT, E. M., ALPERT, R. and GOLDSTEIN, A. C. (1954). Communications through limited-response questioning. *Public Opinion Quarterly* **18** 303-308.

[4] CICCHETTI, D. V. and FEINSTEIN, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* **43** 551–558.

[5] COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20** 37–46.

[6] COHEN, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70** 213–220.

[7] CONGER, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin* **88** 322.

[8] DAVIES, M. and FLEISS, J. L. (1982). Measuring agreement for multinomial data. *Biometrics* 1047–1051.

[9] FEINSTEIN, A. R. and CICCHETTI, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* **43** 543–549.

[10] FLEISS, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76** 378.

[11] GONZALEZ-BARRIOS, J. M. (1998). Sums of nonindependent Bernoulli random variables. *Brazilian Journal of Probability and Statistics* 55–64.

[12] GOWER, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27** 857–871.

[13] 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.

[14] GWET, K. L. (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 4th ed. Advanced Analytics, LLC, Gaithersburg, MD.

[15] HAYES, A. F. and KRIPPENDORFF, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* **1** 77–89.

[16] HENN, L. L. (2021). Limitations and performance of three approaches to Bayesian inference for Gaussian copula regression models of discrete data. *Computational Statistics* 1–38.

[17] HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics* **1** 265–283.

[18] HUGHES, J. (2021). krippendorffsalpha: An R package for measuring agreement using Krippendorff's Alpha coefficient. *The R Journal* **13** 413–425.

[19] KRIPPENDORFF, K. (2012). *Content Analysis: An Introduction to Its Methodology*. Sage.

[20] LANDIS, J. R. and KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 159–174.

[21] LONGONI, M., POBER, B. R. and HIGH, F. A. (2020). Congenital diaphragmatic hernia overview. *GeneReviews®[Internet]*.

[22] MCCULLAGH, P. (2000). Resampling and exchangeable arrays. *Bernoulli* 285–301.

[23] OWEN, A. B. (2007). The pigeonhole bootstrap. *The Annals of Applied Statistics* **1** 386 – 411.

[24] RUBIN, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* 130–134.

[25] SCOTT, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* **19** 321–325.

[26] SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52** 591–611.

[27] SMEETON, N. C. (1985). Early history of the kappa statistic. *Biometrics* **41** 795–795.

[28] TABER, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education* **48** 1273–1296.

[29] TALEB, N. N. (2019). Fooled by correlation: Common misinterpretations in social science. *Academia Online*.