Manuscript APM-13-09-131

A Monte Carlo-based Bayesian approach for measuring agreement in a qualitative scale.

We have modified the paper attending to the comments and suggestions raised by the Reviewers. An itemized reply is presented here. We thank all the Reviewers for their comments which have improved the content and the readability of the paper. We have acknowledged their work in the paper.

Reviewer 1

The manuscript introduces a novel, straightforward, and useful Bayesian Dirichlet-Multinomial model, for the analysis of inter-rater agreement. The new contribution is important. The manuscript appears to be technically correct. The model is convincingly demonstrated on real data sets.

1. The section numbers are not appearing in the text (LaTex issue?).

We have displayed sections and subsections in the LaTex source file, but section numbering is not presented because the format of *Journal Applied Psychological Measurements* does not consider it.

2. An extra line space is appearing before some of the equations.

It has been fixed.

Reviewer 2

This paper submission presents a Bayesian approach based on measures of agreement. This paper is very well written and should supply readers with valuable information that is relatively easy to compute. I noted several strengths such as a great description of the history of agreement studies, general mathematical description for gaining posterior calculations for general agreement measures, a flexible set of priors *including Dirichlet and mixture of Dirichlet, and a nice set of examples. One weakness was the lack of references on Bayesian agreement studies from the statistics literature, for example I know that papers in Statistics in Medicine have addressed this issue of Bayesian kappa agreement and will need to be included in this paper.*

1. P.7, line 42. I suggest making it clear that that Ah is a denominator calculation that is done so that one gets the posterior distribution of the agreement measure, pie(h(rho)—n). Just add a sentence that says this to clarify the goal.(Appendix A is very helpful, but Id like kappa in the main paper).

According to your comment, we have specified that the result of $\hat{A}_h$ is the mean of the sampling posterior distribution of the agreement index. Also, Cohen's Kappa has been included in the main paper as recommended, considering it is one of the most important measures of agreement.

2. P. 11. It would be helpful in the first application to give the reader the summary of the 4x4 table so that readers may apply the approach themselves.

We agree with you, but the space for the article is limited. We have included this information, waiting for the editor's opinion about the length of the article.

Reviewer 3

The authors present an interesting approach for measuring agreement in a qualitative scale. The main contribution is to provide a unified Monte Carlo-based framework to estimate all types of measures of agreement in both informative and non-informative scenarios. The discussion of this methodology and the applications are valuable. However, the following points should be addressed.

1. There are a number of papers that have been published in recent years that may relate to your work. I give a selection below. Please understand, it is not imperative that you use any of these particular references. However, the content they describe is relevant.

- Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. Applied Psychological Measurement,30 (5), 412-431.

- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. Applied Psychological Measurement, 34 (6), 379-392.

- Zopluoglu, C. (2013). CopyDetect An R Package for Computing Statistical Indices to Detect Answer Copying on Multiple-Choice Examinations. Applied Psychological Measurement, 37 (1), 93-95.

- Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. Applied Psychological Measurement, 23 (2), 127-135.

- Blackman, N. J., & Koval, J. J. (1993). Estimating Rater Agreement in 2 x 2 Tables: Correction for Chance and Intraclass Correlation. Applied Psychological Measurement, 17 (3), 211-223.

- Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. Educational and Psychological Measurement, 61(2), 277-289.

These references are very interesting. Most of them have been previously read. Now, we have included some of them in the article.

2. In the Figure 1, all the posterior distributions are obtained by using non informative prior distributions, and, therefore, the influence of the prior information is very low. It is important to explain which would be the effect of informative prior distributions in these cases.

For illustrative purpose, we have solved the problem with non-informative prior distributions. Anyway, an explanation about the effect of informative prior distributions has been included before the figure.

3. The authors use credible intervals instead of confidence intervals, since the approach is based on Bayesian methodology. The authors should give more details on this topic.

We have included an explanation about the main difference between confidence intervals and credible intervals. Now, credible intervals can be better understood and the results in the table should be easier to interpret.

4. According to my understanding measuring agreement is not included in the ISO norms for sensory analysis. Could it be an interesting proposal?

In the ISO norms there is nothing related to measures of agreement up to now. The tests defined for differentiation by ISO are only based on the binomial distribution. Some methods have been proposed in the last years to improve the study of differentiation, and they are used in practice, but ISO has not included these methods in the normalization. Last year, a differentiation method based on measures of agreement was proposed (see Calle-Alonso, F. and Pérez, C. J. (2013). A Statistical Agreement-Based Approach for Difference Testing. Journal of Sensory Studies, 28(5), 358-369). We think your suggestion could be developed to provide an interesting proposal that ISO might address in the future.

5. Although, the paper is well written, there are some mistakes that should be corrected. For example, in the same direction with. I suggest to carefully proofread the manuscript.

The paper has been proofread and some typos and expressions have been corrected.

6. The code presented in the appendix would be more understandable by adding some comments.

Following your advise, we have included more comments to describe each specific step in the code. Now, we feel it is more intelligible.

Reviewer 4

This paper discussed the Bayesian approach to qualitative agreement inference with two priors, respectively, where one is informative and the other is non-informative. Several remarks are listed below.

1. The relation between agreement and intraclass correlation coefficient (author used inter-rater here) should be addressed. Some discussions about when these two are considered the same and when they are not would be of great help.

In the new version of the paper, we have included some information about intraclass correlation and its use. This article refers specially to qualitative scale measures of agreement without the assumption of common marginal distribution among raters, and for this reason we have not used intraclass coefficient.

2. Some modification of the writing style will definitely improve the current version. One example is to be concrete as much as possible. For instance, (1) in the beginning, it was stated that The most popular measure is .., although it has some disadvantages, see. It would be clearer to the readers if you can state what the disadvantages are in one or two sentences. (2) Same paragraph, There are many other available measures, .in different contexts. What are the contexts?

We have modified the writing style in order to follow your suggestion.

3. There are many sentences about applying the current method to multiple raters. Specifically, author considered the two-way (ρ_{ij}) and three-way (ρ_{ijk} in Appendix A) interaction and adopted a transformation (function h) on these interactions. It is not clear to me what to do if there are more than 3 raters. In that case, how to define the multiple-way interaction such as ρ_{ijkl} or ρ_{ijklm} then? Is there a general rule?

The method includes the chosen measure of agreement in the function h . It must be considered that measures of agreement are very different among them and there is not a general rule to estimate all of them with any number of raters. The generalization must be considered for each measure. For example, proportion agreement is easy to denote $\sum_i \rho_{ii\ldots i}$, but others are more difficult to denote. Anyway, computer implementation can be easily addressed for any concrete measure and any number of raters.

4. It was stated first on page 3 that "By treatment the measures of agreement from a Bayesian viewpoint... This issue has not been much explored yet in this context". This is not quite true. The coverage of the reference about existing Bayesian treatment of this topic is too limited. In this manuscript, author focused most of the discussion on previous results, for instance, the case when no covariates are involved. There have been many papers published considering the case when explanatory variables exist. Author should carefully discuss the current state of the research. And clarify how much input the current manuscript can contribute. Here are simply a *few recent references,*

(1) S. Vanbelle et al. (2012) Hierarchical modeling of agreement, Statistics in Medicine, 31, 3667-3680.

(2) M. Tsai (2012) Assessing inter- and intra-agreement for dependent binary data: a Bayesian hierarchical correlation approach, Journal of Applied Statistics, 39, 173-187.

(3) C. Hsiao et al. (2011) Bayesian random effects for interrater and testretest reliability with nested clinical observations, 64, 808-814.

(4) M. Ahmed and M. Shoukri (2010) A Bayesian Estimator of the intracluster Correlation coefficient from Correlated Binary Responses, Journal of Data Science, 8, 127-137.

In the first version of the paper, we did not include all these valuable references because we found this topic out of the scope of the article. In these references, agreement is considered by focusing on both intra-rater and inter-rater. This is very important for problems with test-retest or with nested observations, which is not our case. Also most of them use hierarchical regression methods with explanatory covariables, which are not used in our article. In this article we discuss a method for raters with possibly different expected marginal distributions and without test-retest of the same subjects. Kappa-like measures of agreement should be applied in our case for qualitative scale data. In addition, with our method we do not need to use regression models nor Markov Chain to estimate the measures of agreement. We propose a simpler way to estimate them based on the Monte Carlo method and eliciting prior distributions in two different ways to include information from experts. In the other articles, information from experts is not mentioned to be elicited, so the interest of this proposal is the development of a Bayesian framework where the prior distribution is elicited in two possible ways and a direct computing method for inter-rater measures of agreement is proposed.

However, we agree with you, it is interesting to reference and discuss the articles you mentioned. The references have been incorporated and some explanations have been included to clarify our contribution according to your advice.

5. Some references are missing in the reference list. For example, the four examples on the second paragraph in Introduction showed only the year but not the authors.

On page 21, the reference of Mounchili et al. contains several missing authors. I gave up checking after a while. I will leave the rest to the author.

All the citations in the text have been proofread and all the missing items have been added. Note that APA citation style has been used according to the journal of Applied Psychological Measurement requirements.

For the reference of Mounchili *et al.*, we have followed the 6th edition of the APA manual. In this style manual, when there are entries with more than seven authors the way to reference them is the following: The names of the first 6 authors should be listed, followed by "..., followed by the name of the final author. Nevertheless, we have modified this to allow the inclusion of all the authors of the paper, as you have suggested and it seems logical.

6. On page 3, lines 14-22, the numbers of the section (or maybe the titles of the section) are missing. And, the last sentence of this paragraph is Finally, two appendices contain some interesting information related to the proposal. It would be better to state specifically, what the interesting information is here, rather than simply say it is interesting.

We are sorry for this erratum, the article was written with LaTeX and the format of the journal does not include numbering for the sections. We have rewritten the outline according to your recommendation. Also we have specified the available information in the appendices.

Abstract

Agreement analysis has been an active research area whose techniques have been widely applied in psychology and other fields. However, statistical agreement among raters has been mainly considered from a classical statistics point of view. Bayesian methodology is a viable alternative that allows the inclusion of subjective initial information coming from expert opinions, personal judgments, or historical data. A Bayesian approach is proposed by providing a unified Monte Carlo-based framework to estimate all types of measures of agreement in a qualitative scale of response. The approach is conceptually simple and it has a low computational cost. Both informative and non-informative scenarios are considered. In case no initial information is available, the results are in line with the classical methodology, but providing more information on the measures of agreement. For the informative case, some guidelines are presented to elicitate the prior distribution. The approach has been applied to two applications related to schizophrenia diagnosis and sensory analysis.

Keywords: Bayesian methodology; Measures of agreement; Monte Carlo methods; Multiple raters; Prior elicitation.

A Monte Carlo-based Bayesian approach for measuring agreement in a qualitative scale

Introduction

Agreement among raters is of great importance for researchers and practitioners who describe and evaluate objects and behaviors in a number of fields, including the social and behavioral sciences. Fleiss *et al.* (1969) and Fleiss (1971) presented two of the most influential articles on measures of agreement. Since then, agreement analysis has been an active research area whose techniques have been widely used in practice. The most popular measure of agreement is Cohen's Kappa, although it has some disadvantages (see Cicchetti and Feinstein (1990a, 1990b)). For instance, the effect of sensitivity and specificity makes Kappa vary decisively, showing very different values with a same proportion of agreement but different marginal distributions. This indicates a serious limitation when comparing Cohen's kappa coefficient values among studies with varying prevalence. There are many other available measures, each one with its own characteristics, that can be used in different contexts. For example, there are specific measures to be used in problems with a gold standard, with ordinal weighted values, or problems with stratified data. Two valuable references on inter-rater measures of agreement are Gwet (2010) and Agresti (1992).

Measures of agreement have been widely used in psychology publications, for example, to measure scales for autism spectrum disorders (Cicchetti (2012)), to validate the results of a mathematical model for brainstorming (Coskun and Yilmaz (2009)), to detect answer copying in tests (Zopluoglu (2013) or Belov and Armstrong (2010)), to measure agreement with interval or nominal multivariate observations (Janson and Olsson (2001)), or to analyze the agreement between tests for developmental coordination disorder (Cairney and Streiner (2011)). In the sensory analysis context, few but interesting results can be found in the literature. Sensory analysis belongs to a psychology area known as psychophysics (see, e.g., Bruce *et al.* (1996)). It can be defined as the knowledge area studying some properties or characteristics of a product that can be perceived by human sensory organs. Sensory analysis techniques provide subjective information from acceptance about different products and they can be used for determining the overall quality. Measures of agreement are appropriate to be applied to data collected from experiments in this area. For example, Wu and Chen (1995) considered the agreement among raters to evaluate the accordance of tea sensory data, whereas Mounchili *et al.* (2005) considered the agreement in a sensory analysis of milk samples. Calle and Pérez (2013) proposed an agreement-based methodology for difference testing.

Bayesian methodology provides a full paradigm for statistical thinking. By design, Bayesian methods natively consider, in opposite to the classical methodology, the uncertainty associated with the parameters of a model. Bayesian methods are recommended as the proper way to make formal use of subjective initial information such as expert opinions and personal judgments or beliefs (see, e.g., Bernardo (2003)). Nevertheless, non-informative prior scenarios can also be considered. By treating the measures of agreement from a Bayesian viewpoint, profitable approaches can be built.

Measuring agreement with Bayesian methodology has been considered in different contexts. When the problem is focused on qualitative data and there are no assumptions about a common marginal distribution for the raters, Kappa-like measures of agreement should be applied to estimate inter-rater agreement (Broemeling (2009)). When intra-rater reliability is also the focus, correlation measures and regression models of agreement are generally used. This is common with test-retest evaluations and also with nested observations. There are some recent publications analyzing correlation between the observations from the same rater and among the different raters at the same time from a Bayesian perspective. For example, Tsai (2012) and Hsiao *et al.* (2011) proposed methods with hierarchical correlation approaches for test-retest observations. It is also possible, but less common, to hierarchically study the agreement with measures from Kappa-like family. For example, Vanbelle *et al.* (2012) proposed two Bayesian hierarchical indices to quantify the agreement between a pair of examiners in the context of multilevel data. From a similar point of view, but just paying attention to intraclass correlation, Ahmed and Shoukri (2010) presented a Bayesian estimator under the Beta-Binomial distribution.

In this paper, a novel Bayesian approach based on measures of inter-rater agreement for qualitative scale response is proposed. The approach considers two models, where two or more raters are involved. Both informative and non-informative scenarios are considered, using the Dirichlet-Multinomial family of distributions. When there is no prior information, the measure of agreement is directly comparable to the classic inter-rater agreement. However, when there is initial information, one or more experts elicitate this information through a prior distribution. Then, the posterior measure of agreement contains the current information from the raters and the expert. The participation of the expert allows including information that the raters could not dispose. A discussion on the main measures of agreement and a Monte Carlo-based framework to calculate them is also presented.

The paper is organized as follows. The first section presents a short non-exhaustive review of the main measures of agreement in a qualitative scale of response. Then, the approach is described using the Dirichlet-Multinomial model and a mixture-based generalization. Following, two applications are presented. The first one relates to schizophrenia diagnosis, whereas the second one considers a sensory analysis for food products. The last section presents the conclusions. Finally, two appendices include the definition of the main measures of agreement and the code to apply the proposed method in R software.

Measures of agreement

The literature on inter-rater agreement has extensively grown and numerous accordance measures have been proposed. This paper is focused on inter-rater agreement when the response variable is nominal or ordinal.

In other contexts, it is also interesting to study the agreement with quantitative variables. For quantitative data, the measure of reliability is often studied by association indices such as the intraclass coefficient (Koch (1982)). On the other hand, accordance among raters in qualitative scale data is usually described with inter-rater Kappa-like agreement measures (Fleiss *et al.* (2013)). In general, when the assumption of a common marginal distribution across raters is not tenable, using measures similar to Cohen's kappa is more appropriate. If each rater uses the same underlying marginal distribution of ratings, then the intraclass correlation is suitable (Bloch and Kraemer (1989)) and the intraclass correlation and Kappa offer the same results. In our case, marginal distributions across raters are not demonstrated to be the same so, from now on, measures of inter-rater agreement for qualitative scale data are considered.

A short non-exhaustive review containing the most known measures of agreement in a qualitative scale is presented. The measures can be classified in two groups: corrected and non-corrected by chance. Some authors have noticed that a certain amount of agreement is to be expected by chance and they try to correct the measures according to this assumption (see, e.g., Gwet (2010)), but others, who believe that there is no need for such adjustment or even that it is wrong (see, e.g., Guggenmos (2006)), use measures that are not chance-corrected. Most of the measures of agreement have been proposed for the two raters case, and very few have been defined for several raters. A summary of measures of agreement is presented in the rest of this section.

Firstly, the accordance without chance correction is considered. Agreement is defined as the association among raters, reflecting if they classify subjects in the same category. For this purpose, and considering two raters, the most elementary index of agreement is the sum of proportions of subjects classified into the same category by both observers (Goodman and Kruskal (1954)). This measure of inter-rater agreement is called agreement proportion. It takes values between 0 (when there is no agreement at all) and 1 (when the agreement is complete). Several authors use this measure as the starting point and then define their indices applying some transformations to it (Armitage *et al.* (1966)).

The conditional agreement for a single response can also be considered. Dice (1945) proposed a measure for two raters and two alternatives that gives the agreement for only one of the alternatives. It is called S_D for the first alternative and S'_D for the second one. It takes values in $[0, 1]$, but later on Goodman and Kruskal (1954) defined a new measure which was simply a rescaled Dice index varying in $[-1, 1]$. Another important conditional measure was proposed by Jaccard (1908). Dice indices are actually the most used measures for conditional agreement.

By other hand, Holley and Guildford (1964) proposed the G coefficient for measuring overall agreement, which was later redefined by Maxwell (1977). This coefficient has some good properties, such as not being affected by prevalence or bias, and it coincides with Bennet's sigma index (Bennet *et al.* (1954)) in the case of two raters. Rogot and Goldberg (1966) defined two measures of agreement A_1 and A_2. The first one is the mean of four conditional probabilities, and it has an interesting property, being 0.5 when the two raters are completely independent. The second measure A_2 is just the mean of S_D and S'_D measures.

All these non-corrected by chance measures have been widely used for two raters through the years, but there are still few generalized measures for more than two raters. Agreement proportion, Dice indices and G coefficient generalizations are the most used measures.

Now, the focus is on measures corrected by chance for two raters. One of the first chance-corrected measure was introduced by Bennet *et al.* (1954), using a fixed chance correction equal to the inverse of the number of alternatives. Later Scott's π (1955) was defined by assuming that the marginal distribution of both raters is uniform. This measure was extended by Cohen (1960) and it came to be known as Cohen's kappa. There is only one applicability condition for Cohen's measure: the raters have to operate independently. There is no restriction on the marginal distribution, what made Kappa a much more used measure than Scott's π. It varies in the interval [-1,1], and the most extended interpretation of this measure was provided by Landis and Koch (1977), i.e., values greater than 0.60 may show a good agreement, values below 0.40 imply a poor agreement, and values between 0.40 and 0.60 show that the agreement is moderate. This rule is clearly subjective, but it has been widely considered as the standard for the interpretation of Cohen's Kappa.

Some measures of agreement defined as non-corrected by chance can be corrected. For example, Rogot and Goldberg's A_1 can be corrected to provide values in [-1,1]. Several authors have noted that some of these coefficients become equivalent after correction and, curiously, many of them coincide with Cohen's Kappa. For example, when correcting by chance Rogot and Goldberg's A_2, Goodman and Kruskal's Lambda or Dice's indices, the Cohen's Kappa is recovered (see Broemeling (2009)).

For some special cases there exist specific measures of agreement. For instance, in the case that the studied population is separated in strata, Barlow *et al.* (1991) defined the stratified Kappa. They used weights based on the size of stratum, on the variance, and on uniform weights. Another common case is the comparison between two raters and a gold standard. It can be very interesting for training raters. Thompson and Walter (1988) solved this problem by dividing the information in two tables and then obtaining the overall agreement for both in one single measure. Another example is the weighted Kappa (Cohen (1968)). This measure is specially defined for those situations where more than two alternatives in an ordinal scale are evaluated by two raters. Fleiss and Cohen (1973) and Cicchetti and Fleiss (1977) proposed a selection of weights for Cohen's weighted Kappa.

Finally, some chance-corrected measures of agreement have been defined for more than two raters by extending already existing measures. For example, Conger (1980) defined a generalized Kappa as an overall agreement measure for three or more raters based on Cohen's Kappa, Fleiss (1971) proposed a generalization of Scott's π, and Mielke *et al.* (2007) proposed a generalized weighted Kappa, giving more weight to the diagonal values for the estimation of the agreement.

For further information on measures of agreement the following reference books are useful. Von Eye and Mun (2005) describe the agreement from different points of view including agreement based on log-linear models, cross-classification indicators, and correlation/covariation structures. Shoukri (2010) focuses on the basics of inter-rater agreement and the practical topics, including many real examples to understand all the concepts without heavy mathematical details. By using Bayesian approaches, Broemeling (2009) provides statistical inferences based on various models of intra and inter-rater agreement using WinBUGS software. Numerous examples, especially from medical research, psychology and sociology are described.

The Bayesian approach

Measures of agreement for m raters and c alternatives in a qualitative scale are analyzed from a Bayesian point of view. In order to simplify and without loss of generality, the notation is considered for two raters ($m = 2$) and two alternatives ($c = 2$). If a qualitative variable X ranges over $1, 2$, then n_{ij} denotes the number of observations for which Rater 1 gives the answer $X = i$ and Rater 2 gives the answer $X = j$, with $i, j = 1, 2$. Table 1 shows the observed and marginal frequencies. The corresponding probabilities, denoted by ρ_{ij}, $i, j = 1, 2$, constitute the parameters of interest in the proposed model.

Table 1

Absolute frequencies in a 2×2 *contingency table.*

	Rater 2		
Rater 1	$X = 1$	$X = 2$	Total
$X = 1$	n_{11}	n_{12}	$n_{1\cdot}$
$X = 2$	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

In classical statistics, data in a contingency table are used to calculate punctual and confidential estimations of the measures of agreement. This is performed by considering $\hat{\rho}_{ij} = n_{ij}/n,\ i, j = 1, 2$. However, in Bayesian methodology, the parameters of interest are random variables instead of fixed quantities. In this case the unknown probability vector is $\boldsymbol{\rho} = (\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22})$ and has a prior density function denoted by $\pi(\boldsymbol{\rho})$. This distribution contains all the initial information obtained independently of the data collection, as, for example, information provided by experts' beliefs and/or by historical data. On the other hand, the experimental results, $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})$, and the chosen model provide the likelihood function $L(\mathbf{n}|\boldsymbol{\rho})$. The prior distribution and the likelihood function are combined by using Bayes' theorem to provide the posterior distribution, i.e.:

$$\pi(\boldsymbol{\rho}|\mathbf{n}) = \frac{\pi(\boldsymbol{\rho})L(\mathbf{n}|\boldsymbol{\rho})}{\int \pi(\boldsymbol{\rho})L(\mathbf{n}|\boldsymbol{\rho})d\boldsymbol{\rho}}. \tag{1}$$

The posterior distribution contains all the current information about the parameter vector $\boldsymbol{\rho}$. This posterior distribution is analytically available for the models presented in this section. Otherwise, it can be estimated by using numerical methods. Anyway, this allows to obtain all the current information on the parameter vector through a probability distribution, instead of providing only punctual and confidential estimations as with the classical methodology. Besides, it is possible to quantify how much our beliefs or historical information have changed after collecting the data. This can be performed by means of the Kullback-Leibler divergence (see Kullback and Leibler (1987))

$$KL(\pi(\boldsymbol{\rho}|\mathbf{n}), \pi(\boldsymbol{\rho})) = \int \pi(\boldsymbol{\rho}|\mathbf{n}) \log \left(\frac{\pi(\boldsymbol{\rho}|\mathbf{n})}{\pi(\boldsymbol{\rho})} \right) d\boldsymbol{\rho}. \tag{2}$$

Now, the interest is focused on estimating posterior measures of agreement. They are defined as

$$A_h = \int h(\boldsymbol{\rho})\pi(\boldsymbol{\rho}|\mathbf{n})d\boldsymbol{\rho}, \tag{3}$$

where $h(\cdot)$ is the agreement function related to a concrete measure (see Appendix A for the main agreement functions). For example, Cohen's κ agreement index function for two raters is defined as,

$$h(\boldsymbol{\rho}) = \kappa(\boldsymbol{\rho}) = \frac{(\rho_{11} + \rho_{22}) - (\rho_{1\cdot}\rho_{\cdot 1} + \rho_{2\cdot}\rho_{\cdot 2})}{1 - (\rho_{1\cdot}\rho_{\cdot 1} + \rho_{2\cdot}\rho_{\cdot 2})}.$$

Note that the previous integral can not be analytically calculated for any agreement function, so numerical integration must be performed. A Monte Carlo-based approach is proposed for this task (see, e.g., Fishman (1996)). Firstly, a random sample $\boldsymbol{\rho}^{(t)}$, $t = 1, 2, \ldots, T$, is generated from the posterior distribution given in (1) based on an independent identically distributed (i.i.d.) sampling. Then, the measure of agreement A_h is estimated by

$$\hat{A}_h = \frac{\sum\limits_{t=1}^{T} h(\boldsymbol{\rho}^{(t)})}{T}, \tag{4}$$

where T is the sample size. This procedure provides consistent and unbiased estimates. The estimation $\hat{A}_h$ is the mean of the sampling posterior distribution of the measure of agreement. The theoretical result supporting the convergence of $\hat{A}_h$ to A_h is the Law of Large Numbers (see Geweke (1989)). The larger the sample size is, the more accurate the estimations are. The Monte Carlo error estimate is

$$\widehat{SE}(\hat{A}_h) = \sqrt{\frac{\sum_{t=1}^{T}(h(\boldsymbol{\rho}^{(t)}) - \hat{A}_h)^2}{T(T-1)}}. \tag{5}$$

This approach provides a unified framework to estimate all types of measures of agreement in a qualitative scale of response, allowing the incorporation of initial information. When initial information is included, the posterior measure of agreement contains the current information from the raters and the expert/s. The participation of the expert/s allows including information that the raters could not dispose. When there is no prior information, the measure of agreement is directly comparable to the traditional inter-rater agreement, but more information can be obtained since the probability distribution for the measure of agreement is available. The approach leads to accurate results with a very low computational cost. Besides, it is conceptually simple. The specific details of the concrete models are presented in the following subsections.

Dirichlet-Multinomial model

The Multinomial distribution is used to describe data where each observation is classified into a number of possible outcomes. In this model, the likelihood $L(\mathbf{n}|\boldsymbol{\rho})$ is considered to be a Multinomial distribution with vector parameter $\boldsymbol{\rho}$ related to the count data $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})$. The Dirichlet distribution is a conjugate prior for the parameters of the Multinomial one, i.e., if $\boldsymbol{\rho} \sim D(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22})$, then the posterior distribution is also a Dirichlet distribution, specifically, $\boldsymbol{\rho}|\mathbf{n} \sim D(\alpha_{11} + n_{11}, \alpha_{12} + n_{12}, \alpha_{21} + n_{21}, \alpha_{22} + n_{22})$. See Lindley (1964) and Good (1965).

A great advantage of this model is that the posterior distribution is analytically known, which allows easy calculations for several quantities of interest (mean, mode, variance...). Besides, random variates from Dirichlet distributions are straightforward to generate, so an i.i.d. random sample $\boldsymbol{\rho}^{(t)}$, $t = 1, 2, \ldots, T$, can be easily obtained (see Warnes (2014)). This allows to estimate the measure of agreement (4) and its Monte Carlo error (5).

The remaining task is the elicitation of the prior distribution parameters. This approach allows to treat both non-informative and informative settings. For the non-informative scenario, the uniform and the least-informative Jeffreys' prior distributions are particular cases of the Dirichlet distribution. They are recovered when $\alpha_{ij} = 1$, $i, j = 1, 2$, and $\alpha_{ij} = 1/2$, $i, j = 1, 2$, respectively. Therefore, the Dirichlet class includes the natural "non-informative" prior distributions where there is no prior information to favor one component over any other. Technically, the improper distribution is not a particular case, since α_{ij} must be greater than 0 for all i, j. However, Lindley (1964) gave special attention to this improper limiting case. From a practical point of view, the improper distribution can be used as a Dirichlet one with parameters $\alpha_{ij} = 0.001$, $i, j = 1, 2$. In the three cases the data will dominate the prior distribution, i.e., the posterior distribution will be more influenced by the data than by the prior distribution. The results obtained by using the three posterior distributions are similar, but even more when the sample size is large.

When initial information is available, it can be included in the prior distribution through a parameter elicitation. The selection of an appropriate procedure to elicitate subjective probabilities must consider the expert training and/or the historical information available. In this subsection, an expert-based approach is considered to elicitate the prior parameters for the Dirichlet distribution. The expert must not participate in the data collection process and must be an experienced analyst in the field, with knowledge of all the important information on the concrete experiment, the raters involved, and the historical information, if available. The expert can incorporate his/her initial information on the parameters by using the mean and variance of the marginal distributions, i.e.,

$$\mathrm{E}[\rho_{ij}] = \frac{\alpha_{ij}}{\alpha_0}, \quad \text{and} \quad \mathrm{Var}[\rho_{ij}] = \frac{\alpha_{ij}(\alpha_0 - \alpha_{ij})}{\alpha_0^2(\alpha_0+1)} \;\; i,j = 1,2, \tag{6}$$

where $\alpha_0 = \sum_{i,j} \alpha_{ij}$ is known as flattening constant. The expert's best guess about the true value of the parameter vector $\boldsymbol{\rho}$ is denoted by $\boldsymbol{\rho}^* = (\rho^*_{11}, \rho^*_{12}, \rho^*_{21}, \rho^*_{22})$. This information can be included in the prior distribution by using the relationship $\alpha_{ij} = \alpha_0 \rho^*_{ij},\; i,j = 1,2$. The flattening constant controls the marginal variances, so the expert may use it to express the strength of his/her belief on the prior estimate of the parameters. Large (small) values of the flattening constant match to a high (low) degree of belief in the prior estimations. In other words, the larger (smaller) α_0, the less (more) spread out the prior distribution is, and therefore the more (less) confidence the expert has in the prior mean before considering the data. When α_0 is large compared to n, the prior will tend to dominate the data. When n greatly outnumbers α_0, the data will dominate the prior distribution. This procedure has been applied in a maintenance optimization context by van Noortwijk *et al.* (1992).

An Empirical Bayes approach can also be used. The prior parameters can be obtained by directly using historical data or a randomly selected small portion of the current data (see Carlin and Louis (1996)). In an iterative process, the estimated probability parameters of the current posterior distribution can be considered as the parameters for the prior one in the next stage. Anyway, the initial guess for the probability parameters is obtained by using the available information or a non-informative prior. Then, the previous procedure to elicitate the parameters of the prior Dirichlet distribution can be applied.

Finally, this model allows to obtain an analytical expression for the Kullback-Leibler divergence presented in (2) (see, e.g, Penny (2001)), i.e.,

$$\begin{aligned} KL(\pi(\boldsymbol{\rho}|\mathbf{n}), \pi(\boldsymbol{\rho})) &= \log \frac{\Gamma(\alpha_0+n)}{\Gamma(\alpha_0)} + \\ &+ \sum_{i,j} \log \frac{\Gamma(n_{ij})}{\Gamma(\alpha_{ij}+n_{ij})} + \sum_{i,j} n_{ij}[\psi(\alpha_{ij}+n_{ij}) - \psi(\alpha_0+n)], \end{aligned} \tag{7}$$

where Γ and ψ represent the gamma and digamma functions, respectively.

A mixture-based generalized model

Assume that two or more experts are providing initial information on the experiment. The initial information of each rater can be individually elicitated through a Dirichlet distribution, as presented in the previous subsection. Then, the prior distributions are combined into a consensus prior distribution through a mixture. Bayes' theorem is applied to the multinomial likelihood and provides a posterior distribution that is a mixture of Dirichlet ones, so the conjugacy property is kept. Although this model is already known (see, e.g., Holmes *et al.* (2012) for probabilistic modeling of microbial metagenomics data), its use in a systematic procedure in the agreement context is completely new.

In order to simplify notation and without loss of generality, assume that two experts are involved by providing $D_1(\boldsymbol{\alpha})$ and $D_2(\boldsymbol{\beta})$ as the respective Dirichlet prior distributions. They are combined to provide the consensus prior distribution

$$\pi(\boldsymbol{\rho}) = \omega_1 D_1(\boldsymbol{\alpha}) + \omega_2 D_2(\boldsymbol{\beta}),$$

where ω_1 and ω_2 are non-negative weights summing to unity. Then, the posterior distribution is expressed as

$$\pi(\boldsymbol{\rho}|\mathbf{n}) = w_1^* D_1(\boldsymbol{\alpha} + \mathbf{n}) + w_2^* D_2(\boldsymbol{\beta} + \mathbf{n}), \tag{8}$$

where the updated weights are

$$w_1^* = \frac{w_1 C_1}{w_1 C_1 + w_2 C_2}, w_2^* = \frac{w_2 C_2}{w_1 C_1 + w_2 C_2}, \tag{9}$$

with

$$\begin{aligned} C_1 &= \frac{n!\Gamma(\sum_{i,j} \alpha_{ij})\Gamma(\sum_{i,j} \beta_{ij} + n_{ij})}{\prod_{i,j} \Gamma(\alpha_{ij})\Gamma(\beta_{ij} + n_{ij}) n_{ij}!}, \\ C_2 &= \frac{n!\Gamma(\sum_{i,j} \beta_{ij})\Gamma(\sum_{i,j} \alpha_{ij} + n_{ij})}{\prod_{i,j} \Gamma(\beta_{ij})\Gamma(\alpha_{ij} + n_{ij}) n_{ij}!}. \end{aligned}$$

The remaining task is the weight choice. The weights can be chosen to reflect the relative importance of each expert. There are numerous methods that have been proposed in the literature. A natural choice considers fixed weights proportional to the ranking of the experts in terms of expertise. More complex choices can be implemented depending on the objective. Rufo *et al.* (2009; 2010) proposed calculating the weights through Bayesian hierarchical models. Noortwijk *et al.* (1992) discussed on some available methods to determine the weights.

Once the prior parameters for the Dirichlet distributions and the weights are known, generating from the posterior mixture (8) is straightforward. It is based on generating from the individual posterior distributions with probabilities given by the weights. Therefore, the proposed Monte Carlo framework is also applicable to this model. In this case, the two experts can incorporate initial information that the raters could not dispose. The R[1] code for this approach is presented in Appendix B.

Note that the Dirichlet-Multinomial model presented in the previous subsection is recovered when $\omega_1 = 1$ and $\omega_2 = 0$. When no weight is zero, there is no analytical expression for the Kullback-Leibler divergence given in (2). However, it can also be computed by using a Monte Carlo approach. If $\boldsymbol{\rho}^{(t)}$, $t = 1, 2, \ldots, T$, is an i.i.d. random sample generated from the posterior distribution (8), the Monte Carlo estimate for the Kullback-Leibler divergence is

$$\widehat{KL}(\pi(\boldsymbol{\rho}|\mathbf{n}), \pi(\boldsymbol{\rho})) = \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{\pi(\boldsymbol{\rho}^{(t)}|\mathbf{n})}{\pi(\boldsymbol{\rho}^{(t)})}\right). \tag{10}$$

The next section shows two applications of the proposed approach, where non-informative and informative scenarios are considered.

Applications

Schizophrenia study

Young *et al.* (1982) analyzed four different methods for schizophrenia diagnosis in 196 patients from the Illinois State Psychiatric Institute and classified them by using data from the Present State Examination (1974). The four methods were: Taylor and Abrams (1978), Research Diagnostic Criteria (RDC) (Spitzer *et al.* (1978)), Flexible 6 (Carpenter *et al.* (1973)), and Schneider (1959). Table 2 shows the diagnostics and frequencies, meaning S schizophrenia and NS non-schizophrenia. They studied a pattern of relationship among the diagnoses with latent class analysis, and indicated that the four methods estimated a single underlying diagnosis, but with different degrees of accuracy. The agreement among methods was low. The classification of the disease was better with the Taylor and Abrams' method.

The proposed methodology is applied to these data. In this case, there is no initial information available, so a non-informative approach will be used. The three non-informative prior distributions shown in the previous sections are considered. Firstly, Kullback-Leibler divergence is estimated between posterior and prior distributions. As expected, the distance tends to infinity, because the posterior and the prior distributions are extremely different, since the posterior distribution is highly influenced by the data.

Posterior distributions and Kappa measures were calculated by using the proposed Monte Carlo method with the function κ (see Appendix A). Figure 1(a) shows the estimated posterior distributions of overall Kappa with the different prior distributions. This figure also presents the Kappa measures obtained for the pairwise comparisons between all the methods, according to the following prior distributions: (b) Improper,

[1] www.r-project.org

Table 2

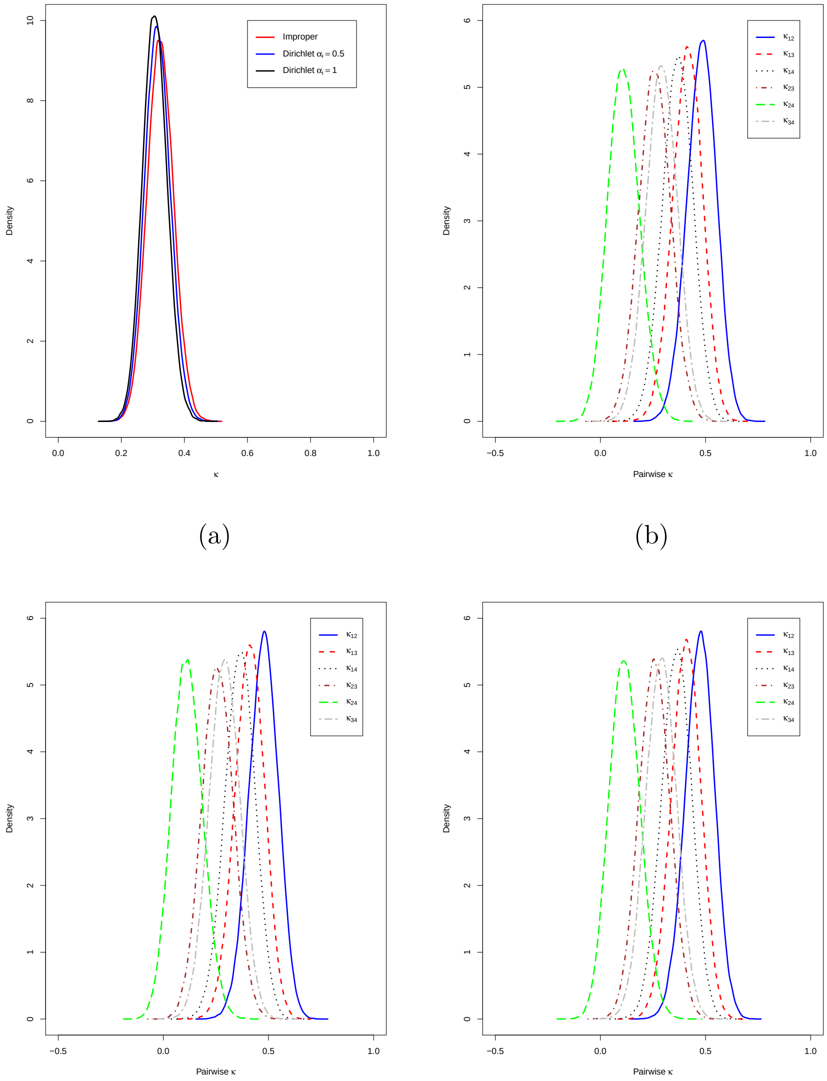
Four methods for schizophrenia diagnosis.

TA	RDC	Flexible	Schneider	Frequency
S	S	S	S	10
S	S	S	NS	8
S	S	NS	S	4
S	S	NS	NS	7
S	NS	S	S	9
S	NS	S	NS	6
S	NS	NS	S	7
S	NS	NS	NS	7
NS	S	S	S	0
NS	S	S	NS	2
NS	S	NS	S	0
NS	S	NS	NS	7
NS	NS	S	S	7
NS	NS	S	NS	13
NS	NS	NS	S	15
NS	NS	NS	NS	94

(c) Jeffrey, and (d) Uniform. The diagnostic methods are numerically denoted by (1) Taylor and Abrams, (2) RDC, (3) Flexible 6, and (4) Schneider. Since prior distributions are non-informative, the differences among the distributions are small. In the case that informative prior distributions were used, the differences among posterior distributions would be greater, and the Kullback-Leibler divergence between the prior and posterior distributions would be reduced.

Figure 1 (a) shows density estimations of the overall Kappa statistic provided by the three non-informative prior distributions. The lowest agreement is achieved with the uniform prior distribution, whereas the highest one is achieved with the Jeffrey's prior distribution. Nevertheless, the differences among the three posterior distributions are very small. Data dominate the prior distributions.

The classic estimation for Kappa is $\hat{\kappa}_c = 0.325$ with a confidence interval $(0.267, 0.384)$. Following Landis and Koc's (1977) scale, it shows a fair agreement. By using classic estimation, only a punctual and confidential estimation can be provided. However, by using the Bayesian approach a probability distribution is obtained



(a) (b)

(c) (d)

Figure 1. (a) Overall κ, (b) Pairwise κ with improper prior distribution, (c) Pairwise κ with prior distribution $D(\alpha_{ij} = 0.5)$, (d) Pairwise κ with prior distribution $D(\alpha_{ij} = 1)$.

for the Kappa measure. This provides more information about the Kappa measure, and the statistics of interest can be easily calculated from this distribution. Table 3 presents the statistics summary for the Kappa measures. Attending to the mean (or median because of the symmetry), the agreement among all the diagnostic methods together is also fair. The highest value is achieved when using the improper prior distribution, i.e., 0.324. Note that 95% credibility intervals are presented, which are conceptually different from confidence intervals. Credible intervals capture the current uncertainty in the location of the parameter values and thus can be interpreted as probabilistic statement about the parameter. In contrast, confidence intervals capture the uncertainty about the obtained interval and they can not be interpreted as a probabilistic statement about the true parameter values.

Table 3

Descriptive summary of the posterior distribution for Kappa by using three non-informative prior distributions.

Prior	$\hat{A}_\kappa$	$SE(\hat{A}_\kappa)$	Median	P_5	P_{95}	Cred. Int.
Improper	0.324	0.00042	0.323	0.256	0.394	(0.242,0.406)
$D(\alpha_{ij} = 0.5)$	0.315	0.00041	0.315	0.250	0.383	(0.236,0.395)
$D(\alpha_{ij} = 1)$	0.307	0.00039	0.307	0.244	0.373	(0.231,0.384)

When measuring agreement for more than two diagnostic methods the pairwise comparisons may provide complementary information to that of the global measure. By this way, the influence of every pair on the general agreement may be uncovered. The partial agreement measures are lower than expected because of the dimensional effect. Figure 1 (b-d) presents the posterior estimations of the partial Kappa distributions for all the diagnostic methods by using the three prior distributions. Table 4 presents the statistics summary. The highest agreement values between methods correspond to the comparisons between Taylor and Abrams' method and the other three. All the estimated pairs including Taylor and Abrams' method achieve agreement average values between 0.36 and 0.48 (fair/moderate), whereas the rest remain between 0.11 and 0.29. This shows that Taylor and Abrams' method truly discovers the core of the diagnosis and the others only some groups of features. By other hand, the agreement found between RDC, Flexible and Schneider is fair/slight. They greatly influence the low value obtained for overall Kappa estimation, specially RDC and Schneider's methods with an average agreement κ close to 0.11. For these two methods, the credibility intervals include null values of κ, meaning that there is the same agreement as if the diagnosis was just performed by chance. But also RDC/Flexible and Flexible/Schneider's achieve low agreement values (under 0.3 in all the cases), concluding that these methods are not suitable to precisely diagnose the disease.

Health diagnoses are expected to have a very high accuracy. Only Taylor and Abrams' method seems to be appropriated and reliable enough. The low overall agreement together with the pairwise agreement results show that the other three methods can be applied to diagnose some kinds of schizophrenia, but not individually or as a gold standard.

Table 4

Descriptive statistics for pairwise Kappa.

Pair	Prior	$\hat{A}_\kappa$	$SE(\hat{A}_\kappa)$	Median	P_5	P_{95}	Cred. Int.
κ_{12}	Improper	0.481	0.00070	0.482	0.364	0.593	(0.345,0.617)
	$D(\alpha_{ij} = 0.5)$	0.476	0.00069	0.478	0.360	0.587	(0.340,0.610)
	$D(\alpha_{ij} = 1)$	0.472	0.00069	0.473	0.356	0.583	(0.337,0.607)
κ_{13}	Improper	0.414	0.00071	0.415	0.295	0.528	(0.276,0.551)
	$D(\alpha_{ij} = 0.5)$	0.410	0.00070	0.411	0.291	0.523	(0.272,0.547)
	$D(\alpha_{ij} = 1)$	0.406	0.00070	0.407	0.290	0.519	(0.270,0.543)
κ_{14}	Improper	0.367	0.00072	0.368	0.246	0.484	(0.225,0.508)
	$D(\alpha_{ij} = 0.5)$	0.365	0.00072	0.366	0.245	0.481	(0.224,0.504)
	$D(\alpha_{ij} = 1)$	0.361	0.00071	0.362	0.242	0.477	(0.221,0.500)
κ_{23}	Improper	0.259	0.00075	0.259	0.138	0.383	(0.113,0.406)
	$D(\alpha_{ij} = 0.5)$	0.258	0.00074	0.258	0.136	0.380	(0.113,0.403)
	$D(\alpha_{ij} = 1)$	0.257	0.00074	0.257	0.137	0.379	(0.113,0.402)
κ_{24}	Improper	0.111	0.00073	0.110	-0.007	0.235	(-0.032,0.255)
	$D(\alpha_{ij} = 0.5)$	0.113	0.00073	0.111	-0.004	0.236	(-0.029,0.256)
	$D(\alpha_{ij} = 1)$	0.115	0.00073	0.114	-0.001	0.237	(-0.026,0.257)
κ_{34}	Improper	0.292	0.00074	0.292	0.169	0.414	(0.147,0.437)
	$D(\alpha_{ij} = 0.5)$	0.289	0.00074	0.290	0.168	0.410	(0.146,0.434)
	$D(\alpha_{ij} = 1)$	0.212	0.00088	0.214	0.065	0.354	(0.041,0.385)

Sensory analysis

Sensory analysis belongs to a psychology area known as psychophysics (see, e.g., Bruce *et al.* (1996)). It can be defined as the branch of psychology concerned with the relationship between physical stimuli perceived by human sensory organs and the effects they produce in the mind. In sensory analysis, some products are evaluated with the sensory organs and described using the perception. Sensory analysis techniques provide subjective information about these products and it can be used for determining the overall quality or differences. It has been widely applied to analyze food and drink appearance, texture, touch, odor, or taste (see e.g., Lawless and Heymann (2010)). Bayesian methodology has almost not been used in this context. Bi (2011) provided an interesting Bayesian approach to non-replicated sensory preference, difference and equivalence tests.

There are many methods to evaluate if there are any perceptible differences between two products, but only three of them have been standardized: the triangle test (ISO 4120:2004 (2004)), the paired comparison test (ISO 5495:2005, (2005)) and the duo-trio test (ISO 10399:2004, (2004)). All three are supported by the International Standardization Organization (ISO), which has developed an international standard for sensory analysis to ensure that products and services are safe, reliable and of good quality. The triangle test is statistically the most efficient one. In this kind of sensory analysis each panelist (rater) receives three product samples, two of the same one and a third different. The panelists are asked to choose the odd sample from the three. Then, the differences are inferred by studying the proportion of right answers above the expected by chance with the binomial distribution.

An experiment has been specifically designed and performed to discriminate two trademarks of Spanish sausage from the highest quality (Iberian extra) through the proposed approach. There were two panelists who participated in six different tasting sessions. Each panelist tasted six samples in each session. The number of sessions was large in order to avoid sensory fatigue. The layout of the products was the same for both panelists. The samples were presented on a plate forming a triangle with one different and two alike pieces of sausage slices of the same shape and thickness (2 mm). Six possible order combinations were randomized across panelists: AAB, ABA, BAA, BBA, BAB, and ABB, being A and B the respective trademarks. The panelists used a document to record their answers. The results of the experiment are shown in Table 5.

Table 5

Sausage tasting results.

	Panelist 1		
Panelist 2	Right	Wrong	Total
Right	26	5	31
Wrong	5	0	5
Total	31	5	36

The most common event is that the two panelists success in the differentiation, happening 26 times out of 36, and it never happened that both of them simultaneously fail at the differentiation. If the interest is focused on measuring the agreement between panelists, Cohen's Kappa is not an appropriate measure because the frequencies are very asymmetrically distributed (across the second diagonal) and this extremely affects to the Kappa index value. Cicchetti and Feinstein (1990b) defined the paradoxes where Kappa should not be used, and partial agreement measures should be considered. Calle-Alonso and Pérez (2013) proposed the use of Dice indices as proper measures of agreement in this context. This allows to separately evaluate the positive and negative agreement (agreement on right and wrong responses), giving information on the discrimination problem. High S_D and low S'_D indicate that the raters are differentiating the products.

There were two experts controlling the experiment. These experts have managed many related experiments and they had knowledge of all the aspects involved. In fact, they knew the panelist staff and specifically the two panelists involved in the experiment. Also, the historical information on the laboratory was known to them. The experts did not participate in the data collection process. The first expert's best guess about the true value of the parameter vector was $(0.66, 0.18, 0.15, 0.01)$, whereas for the second expert it was $(0.55, 0.2, 0.2, 0.05)$. The first expert provided a flattering constant equal to $\alpha_0 = 60$, which is larger than $n = 36$. This means that he had a high degree of belief in his prior estimation. The second one had a less degree of belief in his prior estimation, providing a flattering constant equal to $\alpha_0 = 40$. This allowed to build two Dirichlet prior distributions with parameters (39.6,10.8,9,0.6) and (22,8,8,2), respectively. The experts decided that the initial weights were $w_1 = 0.75$ and $w_2 = 0.25$, giving more importance to the first prior distribution. By this way, they built in a mixture of Dirichlet distributions as a consensus prior distribution. This prior distribution contained all the initial information available to the experts.

The distributions of the three agreement indices (Cohen's Kappa and Dice indices) have been estimated by using the proposed Monte Carlo-based approach with simulated samples of size 10,000 by following the previous specifications. The approach has also been applied in a non-informative scenario (uniform prior distribution) for comparative purposes. Kullback Leibler (KL) divergence has been calculated as a way to evaluate the divergence between the prior and posterior distributions in each scenario. The lowest KL divergence is achieved when the informative prior distribution is used, providing an estimated value of 0.3854 with a Monte Carlo error estimation equal to 0.0063. When the uniform prior distribution is used, the KL divergence is approximately ten times more, i.e., 3.890 with a Monte Carlo error estimate equal to 0.0135. Then, the prior and the posterior distributions are closer when the initial information provided by the experts is considered.

The distributions of the indices have been summarized in Table 6 by means of $\hat{A}_h$, $SE(\hat{A}_h)$, median, 5-th and 95-th percentiles, and the 95% credible interval.

With the mixture prior distribution, the estimated value for Kappa is -0.1393, indicating a very poor agreement. As it has been previously mentioned, this is influenced by the asymmetric data distribution. A high positive agreement $S_D = 0.8120$ and a low negative agreement $S'_D = 0.0086$ are obtained. According to Calle-Alonso and Pérez (2013), this indicates a high degree of differentiation between the two products. Analogously, in the non-informative scenario, Dice indices indicate a high positive ($S_D = 0.8147$) and a low negative agreement ($S'_D = 0.1353$), but credibility intervals are wider than in the informative scenario. Besides, the value of the negative agreement 0.1353 is greater than the one for the informative scenario, i.e.,

Table 6

Summarized distributions of the agreement measures with mixture and uniform prior distributions.

	$\hat{A}_h$	$SE(\hat{A}_h)$	Median	P_5	P_{95}	Cred. Int.
Mixture						
κ	-0.1393	0.0006	-0.1491	-0.2223	-0.0189	(-0.2479,-0.0038)
S_D	0.8120	0.0004	0.8141	0.7516	0.8665	(0.7422,0.8790)
S'_D	0.0086	0.0001	0.0042	0.0001	0.0322	(0.0000,0.03218)
Uniform						
κ	-0.0366	0.0013	-0.0618	-0.2085	0.2156	(-0.2518,0.2456)
S_D	0.8147	0.0005	0.8199	0.7199	0.8922	(0.7102,0.9124)
S'_D	0.1353	0.0011	0.1062	0.0089	0.3639	(0.0000,0.3640)

0.0086. Anyway, there is also evidence to confirm the differentiation between products.

Comparing the results from the two scenarios set out, it is apparent that prior distributions should not be thought of as an innocuous tool. On the contrary, consensually informed prior distributions permit cumulative scientific knowledge to rationally affect conclusions drawn from new observations (see Kruschke (2010)). In this example the data are fairly clear to observe the differentiation, and the prior distribution has a moderate effect. However, there are other situations where the prior distribution can be determinant, leading to very distinct results.

Finally, the distribution of positive and negative Dice indices are shown in Figure 2, S_D in the left histogram and S'_D in the right histogram. White bars represent the simulated values of Dice with the mixture prior distribution, and dark bars represent dice indices simulated with the uniform prior distribution. Intermediate shaded bars are the overlapped values for both distributions. It can be appreciated that the values obtained with non-informative prior distributions are more dispersed, whereas the use of an informative prior distribution produces more concentrated values.

Conclusion

The proposed approach is conceptually simple and computationally efficient to estimate all types of measures of agreement on a qualitative scale of response. The two presented models allow the inclusion of subjective initial information coming from expert opinions, personal judgments or beliefs, or from historical data. They provide probability distributions for the measures of interest instead of only punctual and confidential estimations, as happens with the classical statistical methodology.

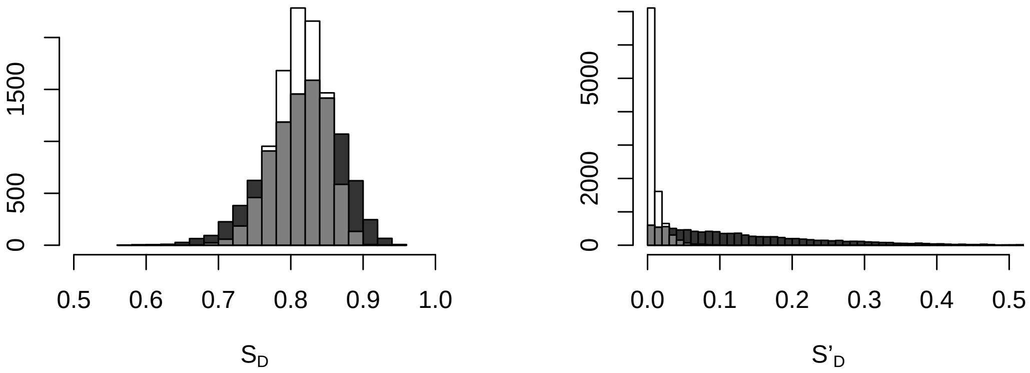


Figure 2. Histograms for Dice indices with non-informative and informative prior distributions

The use of this approach in non-informative settings can be useful in psychology and other related fields. For example, it has been used to analyze the agreement among four different methods to diagnose schizophrenia. However, even more interesting is the approach for informative settings. For example, the proposed approach has been proved to be useful to discriminate among food products in a sensorial analysis context by using Dice indices. The power of this approach can be obtained through real experiments in applied research fields. The proposed approach can be considered as a small step in this research area. Extending this type of Bayesian approach to other experimental structures for quantitative scale of response should be addressed in the immediate future. Broemeling (2009) focused on non-informative settings. Therefore, the challenging task is to develop elicitation techniques to describe the initial information obtained from expert opinions or other ways. Applying these techniques to real experiments will make them more popular and will allow more practitioners to benefit from their advantages.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, *1*(2), 201-218.

Ahmed, M., & Shoukri, M. (2010). A Bayesian estimator of the intracluster correlation coefficient from correlated binary responses. *Journal of Data Science*, *8*, 127–137.

Armitage, P., Blendis, L., & Smyllie, H. (1966). The measurement of observer disagreement in the recording of signs. *Journal of Royal Statistical Society*, *129*, 98-109.

Barlow, W., Lai, M. Y., & Azen, S. P. (1991). A comparison of methods for calculating a stratified Kappa. *Statistics in Medicine*, *10*, 1465-1472.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, *34*(6), 379–392.

Bennet, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, *18*, 303-308.

Bernardo, J. M. (2003). *Encyclopedia of life support systems: Bayesian statistics.* UNESCO.

Bi, J. (2011). Bayesian approach to sensory preference, difference and equivalence tests. *Journal of Sensory Studies*, *26*(5), 383–399.

Bloch, J. M., & Kraemer, H. C. (1989). 2x2 Kappa coefficients: Measures of agreement or association. *Biometrics*, *45*, 269-287.

Broemeling, L. (2009). *Bayesian methods for measures of agreement.* Chapman and Hall.

Bruce, V., Green, P., & Georgeson, M. (1996). *Visual perception (3rd ed.).* Psychology Press.

Cairney, J., & Streiner, D. (2011). Using relative improvement over chance (RIOC) to examine agreement between tests: Three case examples using studies of developmental coordination disorder (DCD) in children. *Research in Developmental Disabilities*, *32*(1), 87 - 92.

Calle-Alonso, F., & Pérez, C. J. (2013). A statistical agreement-based approach for difference testing. *Journal of Sensory Studies*, *28*(5), 358-369.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis.* Chapman & Hall.

Carpenter, W., Strauss, J., & Bartko, J. (1973). Flexible systems for the diagnosis of schizophrenia: Report from the WHO pilot study of schizophrenia. *Science*, *182*, 1275-1278.

Cicchetti, D. (2012). On scales of measurement in autism spectrum disorders (ASD) and beyond: Where smitty went wrong. *Journal of Autism and Developmental Disorders.*

Cicchetti, D. V., & Feinstein, A. R. (1990a). High agreement but low Kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-559.

Cicchetti, D. V., & Feinstein, A. R. (1990b). High agreement but low Kppa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551-558.

Cicchetti, D. V., & Fleis, J. L. (1977). Comparison of the null distribution of weighted Kappa and the cordinal statistic. *Applied Psychological Measurementes*, *1*, 195-201.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-220.

Conger, A. (1980). Integration and generalization of Kappas for multiple raters. *Psychological Bulletin*, *88*, 322-328.

Coskun, H., & Yilmaz, O. (2009). A new dynamical model of brainstorming: Linear, nonlinear, continuous (simultaneous) and impulsive (sequential) cases. *J. of Mathematical Psychology*, *53*(4), 253 - 264.

Dice, L. (1945). Measurements of the amount of ecologic association between species. *Ecology*(26), 297-302.

Fishman, G. S. (1996). *Monte Carlo. concepts, algorithms,and applications.* Springer.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of the weighted Kappa and the interclass Kappa. *Psychological Bulletin*, *33*(3), 613-619.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of Kappa and weighted Kappa. *Psychological Bulletin*, *72*(5), 323-327.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley & Sons.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317-1339.

Goddman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*(49), 732-764.

Good, I. (1965). *The estimation of probabilities: An essay on modern Bayesian methods.* Cambridge, MA: MIT Press.

Guggenmoos-Holzmann, I. (2006). How reliable are chance-corrected measures of agreement? *Statistics in Medicine*, *12*(23).

Gwet, K. (2010). *Handbook of inter-rater reliability.* Advanced Analytics, LLC.

Holley, J. W., & Guildford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measures*, *32*, 281-288.

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, *7*(2), e30126.

Hsiao, C. K., Chen, P. C., & Kao, W.-H. (2011). Bayesian random effects for interrater and test–retest reliability with nested clinical observations. *Journal of clinical epidemiology*, *64*(7), 808–814.

ISO 10399:2004. Sensory analysis. Methodology. Duo-Trio test. (2004). International Organization for Standarization.

ISO 4120:2004. Sensory analysis. Methodology. Triangular test. (2004). International Organization for Standarization.

ISO 5495:2005. Sensory analysis. Methodology. Paired comparison test. (2005). International Organization for Standarization.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societé Vaudoise des Sciences Naturelles*, *44*, 223-270.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*(2), 277–289.

Koch, G. G. (1982). Intraclass correlation coefficient. *Encyclopedia of statistical sciences.*

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences*, *14*(7), 293–300.

Kullback, S. (1987). The Kullback-Leibler distance. *The American Statistician*, *41*(4), 340-341.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lawless, H., & Heymann, H. (2010). *Sensory evaluation of food, principles and practices* (2nd ed.). Springer.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, *35*, 1622-1643.

Maxwell, A. E. (1977). Coefficient of agreement between observers and their interpretation. *British Journal of Psychiatry*, *130*, 79-83.

Mielke, P., Berry, K., & Johnston, J. (2007). The exact variance of weighted kappa with multiple raters. *Psuchological Reports*, *101*, 655-660.

Mounchili, A., Wichtel, J., Bosset, J., Dohoo, I., Imhof, M., Altieri, D., Mallia, S., & Stryhn, H. (2005). HS-SPME gas chromatographic characterization of volatile compounds in milk tainted with off-flavour. *International Dairy Journal*, *15*(12), 1203 - 1215.

Penny, W. D. (2001). *Kullback-Leibler divergences of Normal, Gamma, Dirichlet and Wishart densities* (Technical Report). Wellcome Department of Cognitive Neurology.

Rogot, E., & Goldberg, I. (1966). A proposed index for measuring agreement in test-retest studies. *Journal* *of chronic diseases*(19), 991-1006.

Rufo, M. J., Martín, J., & Pérez, C. J. (2009). Inference on exponential families with mixture of prior distributions. *Computational Statistics and Data Analysis*, *53*, 3271-3280.

Rufo, M. J., Pérez, C. J., & Martín, J. (2010). Merging experts opinions: A Bayesian hierarchical model with mixture of prior distributions. *European Journal of Operational Research*, *207*, 284-289.

Schneider, K. (1959). *Clinical psychopathology.* Grune and Stratton.

Scott, W. (1955). Reliability of content analysis: the cases of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321-325.

Shoukri, M. (2010). *Measures of interobserver agreement and reliability.* Chapman and Hall.

Spitzer, R., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry*, *35*(6), 773-782.

Taylor, M., & Abrams, R. (1978). The prevalence of schizofrenia: A reassessment using modern diagnostic criteria. *American Journal of Psychiatry*, *135*, 945-948.

Thompson, W. D., & Walter, S. D. (1988). A reppraisal of the Kappa coefficient. *Journal of Clinical Epidemiology*, *41*(10), 949-958.

Tsai, M.-Y. (2012). Assessing inter-and intra-agreement for dependent binary data: a Bayesian hierarchical correlation approach. *Journal of Applied Statistics*, *39*(1), 173–187.

Vanbelle, S., Mutsvari, T., Declerck, D., & Lesaffre, E. (2012). Hierarchical modeling of agreement. *Statistics in Medicine*, *31*(28), 3667–3680.

van Noortwijk, J. M., Dekker, R., Cooke, R. M., & Mazzuchi, T. A. (1992). Expert judgement in maintenance optimization. *IEEE Transactions on Reliability*, *41*(3), 427-432.

Von Eye, A., & Mun, E. (2005). *Analyzing rater agreement, manifest variable methods.* Lawrence Erlbaum Associates.

Warnes, G. R. (2014). gtools: Various R programming tools [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=gtools` (R package version 3.3.1)

Wing, J., Cooper, J., & Sartorius, N. (1974). *The measurement and classification of psychiatric symptoms.* Cambridge University Press.

Wu, H., & Chen, L. (1995). Sensory analysis in quality control - the agreement among raters. *Botanical Bulletin of Academia Sinica*, *36*, 121-133.

Young, M. A., Tanner, M. A., & Meltzer, H. Y. (1982). Operational definitions of schizophrenia what do they identify? *Journal of Nervous and Mental Disease*, *170*(8), 443-447.

Zopluoglu, C. (2013). Copydetect an R package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied Psychological Measurement*, *37*(1), 93–95.

Appendix A. Main agreement functions

APPENDICES FOR PUBLICATION AS ONLINE SUPPLEMENTS

Measures for two raters

- Proportion of agreement [14]:

$$P(\boldsymbol{\rho}) = \rho_{11} + \rho_{22}.$$

- Dice indices [7]:

$$S_D(\boldsymbol{\rho}) = \frac{\rho_{11}}{(\rho_{\cdot 1} + \rho_{1\cdot})/2}, \quad S'_D(\boldsymbol{\rho}) = \frac{\rho_{22}}{(\rho_{2\cdot} + \rho_{\cdot 2})/2}.$$

- Goodman and Kruskal λ_r [9]:

$$\lambda_r(\boldsymbol{\rho}) = \frac{2\rho_{11} - (\rho_{12} + \rho_{21})}{2\rho_{11} + (\rho_{12} + \rho_{21})}, \quad \lambda'_r(\boldsymbol{\rho}) = \frac{2\rho_{22} - (\rho_{12} + \rho_{21})}{2\rho_{22} + (\rho_{12} + \rho_{21})}.$$

- Jaccard indices [11]:

$$J(\boldsymbol{\rho}) = \frac{\rho_{11}}{\rho_{11} + \rho_{12} + \rho_{21}}, \quad J'(\boldsymbol{\rho}) = \frac{\rho_{22}}{\rho_{22} + \rho_{12} + \rho_{21}}.$$

- Concordance indices [16]:

$$C(\boldsymbol{\rho}) = \frac{\rho_{11}/(\rho_{11} + \rho_{21}) + \rho_{11}/(\rho_{12} + \rho_{11})}{2}, \quad C'(\boldsymbol{\rho}) = \frac{\rho_{22}/(\rho_{22} + \rho_{21}) + \rho_{22}/(\rho_{12} + \rho_{22})}{2}.$$

- G coefficient [12]:

$$G(\boldsymbol{\rho}) = (\rho_{11} + \rho_{22}) - (\rho_{21} + \rho_{12}).$$

- Rogot and Goldberg A_1 [14]:

$$A_1(\boldsymbol{\rho}) = \frac{1}{4}\left(\frac{\rho_{11}}{\rho_{1\cdot}} + \frac{\rho_{11}}{\rho_{\cdot 1}} + \frac{\rho_{22}}{\rho_{2\cdot}} + \frac{\rho_{2}}{\rho_{\cdot 2}}\right).$$

- Rogot and Goldberg A_2 [14]:

$$A_2(\boldsymbol{\rho}) = \frac{\rho_{11}}{\rho_{1\cdot} + \rho_{\cdot 1}} + \frac{\rho_{22}}{\rho_{2\cdot} + \rho_{\cdot 2}}.$$

- Rescaled standard deviation index [1]:

$$RSD(\boldsymbol{\rho}) = \sqrt{\frac{\rho_{11} + \rho_{22} - (\rho_{11} - \rho_{22})^2}{1 - 1/4(\rho_{1\cdot} + \rho_{\cdot 1} + \rho_{2\cdot} + \rho_{\cdot 2})^2}},$$

- Bennet's σ [3]:

$$\sigma(\boldsymbol{\rho}) = 2(\rho_{11} + \rho_{22}) - 1.$$

- Scott's π [15]:

$$\pi(\boldsymbol{\rho}) = \frac{4(\rho_{11}\rho_{22} - \rho_{12}\rho_{21}) - (\rho_{12} - \rho_{21})^2}{(\rho_{1\cdot} + \rho_{\cdot 1})(\rho_{2\cdot} + \rho_{\cdot 2})}.$$

- Cohen's κ [4]:

$$\kappa(\boldsymbol{\rho}) = \frac{(\rho_{11} + \rho_{22}) - (\rho_{1\cdot}\rho_{\cdot 1} + \rho_{2\cdot}\rho_{\cdot 2})}{1 - (\rho_{1\cdot}\rho_{\cdot 1} + \rho_{2\cdot}\rho_{\cdot 2})}.$$

- r_{11} index [13]:

$$r_{11}(\boldsymbol{\rho}) = \frac{2\rho_{11}\rho_{22} - \rho_{12}\rho_{21}}{\rho_{1\cdot}\rho_{2\cdot} + \rho_{\cdot 1}\rho_{\cdot 2}}.$$

- Chance corrected Rogot and Goldberg A_1 [14]:

$$CA_1(\boldsymbol{\rho}) = \frac{(\rho_{11}\rho_{22} - \rho_{12}\rho_{21})(\rho_{1\cdot}\rho_{2\cdot} + \rho_{\cdot 1}\rho_{\cdot 2})}{2\rho_{1\cdot}\rho_{2\cdot}\rho_{\cdot 1}\rho_{\cdot 2}}.$$

- Conditional Kappa [5]:

$$\kappa_i(\boldsymbol{\rho}) = \frac{\rho_{ii} - \rho_{i\cdot}\rho_{\cdot i}}{\rho_{i\cdot} - \rho_{i\cdot}\rho_{\cdot i}}.$$

- Stratified Kappa [2]:

1. Uniform weights.

$$\kappa_{ave}(\boldsymbol{\rho}) = \frac{1}{S}\sum_{s=1}^{S}\kappa_s.$$

2. Stratum size.

$$\kappa_{siz}(\boldsymbol{\rho}) = \frac{\displaystyle\sum_{s=1}^{S} n_s\kappa_s}{\displaystyle\sum_{s=1}^{S} n_s}.$$

3. Deviation inverse.

$$W_s = \frac{Var^{-1}(\kappa_s)}{\sum_{s=1}^{S} Var^{-1}(\kappa_s)},$$

$$\kappa_{var}(\boldsymbol{\rho}) = \sum_{s=1}^{S} W_s\kappa_s.$$

Measures for more than two raters

For the general case with m raters, there exists few measures of agreement available. These are natural extensions of the ones proposed for two raters. The agreement for more than two raters decrease by the effect of the dimension, and this kind of measures are not used very often in the statistical applications. Without loss of generality, the main measures for three raters are presented.

- Agreement proportion:

$$\sum_i \rho_{iii}.$$

- Dice indices [17]:

$$S_{D_i} = \frac{\rho_{iii}}{(\rho_{i\cdot\cdot} + \rho_{\cdot i\cdot} + \rho_{\cdot\cdot i})/3}.$$

- G coefficient [10]:

$$G = 2\sum_{i=1}^{c} \rho_{iii} - 1.$$

- Bennet's σ [6]:

$$\sigma = \frac{\displaystyle\sum_{i=1}^{c} c\rho_{iii} - 1}{c-1},$$

- Fleiss' κ [8]:

$$\kappa = \frac{\displaystyle\sum_{i=1}^{c} \rho_{iii} - \sum_{i=1}^{c} \rho_{i\cdot\cdot}\rho_{\cdot i\cdot}\rho_{\cdot\cdot i}}{1 - \displaystyle\sum_{i=1}^{c} \rho_{i\cdot\cdot}\rho_{\cdot i\cdot}\rho_{\cdot\cdot i}}.$$

Appendix B. R code

```
#Libraries
library(TeachingDemos)
library(gtools)
#Data
n<-c(26,5,5,0)
#Guessing probabilities
alphastar<-c(0.66,0.18,0.15,0.01)
betastar<-c(0.55,0.2,0.2,0.05)
#Flattering constants
alpha0<-60
beta0<-40
#Dirichlet parameters
alpha<-alpha0*alphastar
beta<-beta0*betastar
#Initial weights
w<-c(0.75,0.25)
#Auxiliary constants
c1<-(gamma(sum(alpha))*factorial(sum(n))*prod(gamma(alpha+n)))/
/(prod(gamma(alpha))*prod(factorial(n))*gamma(sum(alpha+n)))
```

```
c2<-(gamma(sum(beta))*factorial(sum(n))*prod(gamma(beta+n)))/
/(prod(gamma(beta))*prod(factorial(n))*gamma(sum(beta+n)))
#Final weights
wstar<-c((w[1]*c1)/(w[1]*c1+w[2]*c2),(w[2]*c2)/(w[1]*c1+w[2]*c2))
#Simulation size
T<-10000
#Auxiliary matrix
samp<-matrix(NA,ncol=4,nrow=T)
#Generation proccess
for (i in 1:T){
 u<-runif(1,0,1)
 ifelse(u<wstar[1],samp[i,]<-rdirichlet(1,alpha+n),
 samp[i,]<-rdirichlet(1,beta+n))
}
#Agreement generated values
agree<-(samp[,1]+samp[,4])
cagree<-(samp[,1]+samp[,2])*(samp[,1]+samp[,3])+
+(samp[,3]+ samp[,4])*(samp[,2]+samp[,4])
kappai<-(agree-cagree)/(1-cagree)
Dice1i<-2*samp[,1]/(2*samp[,1]+samp[,2]+samp[,3])
Dice2i<-2*samp[,4]/(2*samp[,4]+samp[,2]+samp[,3])
KLesti<-log(wstar[1]*ddirichlet(samp,alpha+n)+
+wstar[2]*ddirichlet(samp,beta+n))-
-log((w[1]*ddirichlet(samp,alpha)+w[2]*ddirichlet(samp,beta)))
#Kappa and Monte Carlo error estimations
Kappa<-mean(kappai,na.rm=TRUE)
MCEKappa<-sqrt(sum((Kappa-kappai)^2)/(T*(T-1)))
#Dice indices and Monte Carlo error estimations
Dice1<-mean(Dice1i,na.rm=TRUE)
MCEDice1<-sqrt(sum((Dice1-Dice1i)^2)/(T*(T-1)))
Dice2<-mean(Dice2i,na.rm=TRUE)
MCEDice2<-sqrt(sum((Dice2-Dice2i)^2)/(T*(T-1)))
KLD<-mean(KLesti,na.rm=TRUE)
MCEKLD<-sqrt(sum((na.omit(KLesti)-KLD)^2)/
/(length(KLesti)*(length(KLesti)-1)))
#Summary
result<-matrix(c(Kappa,MCEKappa,median(kappai),
quantile(kappai,c(.05,.95)),emp.hpd(kappai,conf=0.95)[1],
emp.hpd(kappai,conf=0.95)[2],Dice1,MCEDice1,median(Dice1i),
quantile(Dice1i, c(.05, .95)),emp.hpd(Dice1i,conf=0.95)[1],
emp.hpd(Dice1i,conf=0.95)[2],Dice2,MCEDice2,median(Dice2i),
quantile(Dice2i, c(.05,.95)),emp.hpd(Dice2i,conf=0.95)[1],
emp.hpd(Dice2i,conf=0.95)[2]),nrow=3,ncol=7,
byrow=TRUE, dimnames=list(c("kappa","Dice1","Dice2"),
c("A","MCE","Median","P5","P95","L-BCI","U-BCI")))
#Print summary
result
```

[1] Armitage, P., Blendis, L., and Smyllie, H. (1966). The measurement of observer disagreement in the recording of signs. *Journal of Royal Statistical Society*, 129:98–109.

[2] Barlow, W., Lai, M. Y., and Azen, S. P. (1991). A comparison of methods for calculating a stratified Kappa. *Statistics in Medicine*, 10:1465–1472.

[3] Bennet, E. M., Alpert, R., and Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, 18:303–308.

[4] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

[5] Coleman, J. S. (1966). Measuring concordance in attitudes. Unpublished manuscript. Department of Social Relations, Johns Hopkins University.

[6] Conger, A. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88:322–328.

[7] Dice, L. (1945). Measurements of the amount of ecologic association between species. *Ecology*, (26):297–302.

[8] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[9] Goddman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, (49):732–764.

[10] Gwet, K. (2010). *Handbook of Inter-Rater Reliability*. Advanced Analytics, LLC.

[11] Jaccard, P. (1912). The distribution of the flora in the Alpine Zone.1. *New Phytologist*, 11(2):37–50.

[12] Maxwell, A. E. (1977). Coefficient of agreement between observers and their interpretation. *British Journal of Psychiatry*, 130:79–83.

[13] Maxwell, A. E. and Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology*, 21:105–116.

[14] Rogot, E. and Goldberg, I. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of chronic diseases*, (19):991–1006.

[15] Scott, W. (1955). Reliability of content analysis: the cases of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

[16] Shoukri, M. (2010). *Measures of interobserver agreement and reliability.* Chapman and Hall.

[17] Warrens, M. J. (2008). *Similarity Coefficients for Binary Data.* PhD thesis, University of Leiden.