Research Article

Statistics in Medicine

Received 3 October 2011, Accepted 27 March 2012 Published online 26 June 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5424

Hierarchical modeling of agreement

Sophie Vanbelle,[a*†] Timothy Mutsvari,[b] Dominique Declerck[c] and Emmanuel Lesaffre[b,d]

Kappa-like agreement indexes are often used to assess the agreement among examiners on a categorical scale. They have the particularity of correcting the level of agreement for the effect of chance. In the present paper, we first define two agreement indexes belonging to this family in a hierarchical context. In particular, we consider the cases of a random and fixed set of examiners. Then, we develop a method to evaluate the influence of factors on these indexes. Agreement indexes are directly related to a set of covariates through a hierarchical model. We obtain the posterior distribution of the model parameters in a Bayesian framework. We apply the proposed approach on dental data and compare it with the generalized estimating equations approach. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Cohen's kappa; intraclass; reliability; multilevel; Markov chain Monte Carlo; nested; rater

1. Introduction

Kappa-like agreement indexes are commonly used to quantify agreement between two examiners on a categorical scale. They correct the observed proportion of agreement between the examiners for the effect of chance. The kappa-like family includes Cohen's kappa coefficient [1], the weighted kappa coefficient [2], and the intraclass kappa coefficient [3]. These agreement indexes were extended over the years to the case of several examiners [4–7], of a single examiner against a group of examiners [8, 9], and to the case of two groups of examiners [8, 10]. Authors also proposed extensions to account for a regression structure when examiners classify independent or dependent items (subjects or objects) on a categorical scale. In particular, developments were made on the basis of log-linear models [11, 12], latent class models [13–16], and logistic regression analysis [17–21] when agreement is assessed on a sample of independent items while generalized estimating equations (GEE) [22–27] and weighted least-squares [28] were used when agreement is assessed on a sample of dependent items.

Two major criticisms on kappa coefficients were formulated in the literature. Firstly, several authors [29–32] pointed out that Cohen's kappa coefficient is dependent on the prevalence of the trait under study, which indicates a serious limitation when comparing Cohen's kappa coefficient values among studies with varying prevalence. The dependence studied by Thompson and Walter [29] was relative to the prevalence of the true latent binary variable under study, keeping sensitivity and specificity fixed, whereas Feinstein and Cicchetti [30] studied the dependence of Cohen's kappa coefficient on observed marginal prevalences, keeping the proportion of observed agreement fixed. However, Bloch and Kraemer [33] and Vach [34] criticized the results of Thompson and Walter [29] by noting that the dependence occurs only if one can change the prevalence without changing sensitivity and specificity, which is generally not the case. Moreover, Vach [34] pointed out that the dependence studied by Feinstein and Cicchetti [30] is simply a consequence of the purpose of Cohen's kappa coefficient. This was also noted by Hoehler [35], who remarked that examiner bias, by definition, indicates disagreement. The second major criticism against the kappa-like family of coefficients is that, like correlation coefficients, the

[a]*Department of Methodology and Statistics, University of Maastricht, The Netherlands*
[b]*L-Biostat, Catholic University Leuven, Belgium*
[c]*Department of Oral Health Sciences, Catholic University Leuven, Belgium*
[d]*Department of Biostatistics, Erasmus University Rotterdam, The Netherlands*
**Correspondence to: Sophie Vanbelle, Methodology and Statistics, P. Debyeplein 1, 6229 HA Maastricht, The Netherlands.*
[†]*E-mail: sophie.vanbelle@maastrichtuniversity.nl*

interpretation of kappa statistics is not clear except for 0 and 1 values. Landis and Koch [36] therefore constructed a classification scheme to appreciate the strength of agreement. This classification is widely used but should be avoided because its construction is totally arbitrary. It is preferable to consider a confidence interval to appreciate the value of a kappa estimate, the lower bound being often the only of interest. In that perspective, authors derived several methods to estimate the sampling variability of agreement coefficients belonging in the kappa-like family (e.g., [37–39]). Blackmann and Koval [40] provided a guidance in selecting among some of these methods in small samples, according to the value of the prevalence, the level of agreement, and the sample size. Despite these disadvantages and limitations, kappa-like agreement indexes are popular due to their simplicity and wide applicability to assess agreement when no gold standard is available. It should however be kept in mind that a kappa coefficient mixes two sources of disagreement among examiners, that is, disagreement due to bias among examiners (i.e., different base rate of categories between the examiners) and disagreement that occurs because the examiners evaluate the items differently (i.e., rank order the items differently) [41, 42].

We pay particular attention in the present paper to the general case of agreement between two examiners in the presence of a hierarchical structure in the data. For example, a two-level hierarchical structure arises when graders assess the presence (yes/no) of geographic atrophy in the eyes of patients. The agreement between the two graders can be studied at the eye level (level 1) or at the patient level (level 2). Ignoring the correlation between the two strata constituting the first level of hierarchy (left and right eye, respectively) might lead to an incorrect estimation of agreement and incorrect inference [43]. It is not rare to encounter data with higher hierarchical levels. For example, when dentists assess the presence of caries experience (CE) on a dichotomous basis (yes/no) in the mouth of children, we have a three-level hierarchy. Indeed, agreement can be assessed at the tooth surface level (level 1), the tooth level (level 2), and the mouth level (level 3). More specifically, we may be interested in studying the effect of factors defined at different levels of the hierarchy on the agreement obtained between a pair of graders. For example, we may want to estimate the effect of the type of tooth (deciduous or permanent) on the agreement between the dentists when assessing the presence of CE. When a gold standard (benchmark scorer) is available, agreement is measured by the specificity and sensitivity. It is expected [44, 45] that sensitivity becomes more favorable as we go up in the hierarchy from surface to tooth and subject level. This is because of the logical argument that if an examiner scored at least one surface of a tooth as affected by caries, the attributed score will also be positive at the tooth and subject levels. On the other hand, when an examiner scored negatively at the surface level, it is still possible that the scoring is correct at the tooth and subject levels. For example, when the examiner scored the presence of CE on the wrong surface of a tooth, the tooth is still correctly scored as CE. Mutsvari *et al.* [46], who used a full hierarchical model to estimate sensitivity and specificity of CE assessment at the different levels of the hierarchy, observed this phenomenon. On the other hand, there is no strict order of the results for specificity (see [44] for a detailed explanation). Because kappa-like indexes can be expressed in terms of specificity and sensitivity [3], it is not possible to determine the behavior of kappa-like indexes according to the level of the hierarchy. In particular, we have to distinguish between two situations. On one hand, consider that CE was only detected by one of two examiners in a subject. We expect a decreasing level of agreement between these two examiners when rising in the hierarchy. Indeed, the disagreement between the two examiners will remain at all levels of the hierarchy, whereas the number of units of analysis and therefore the number of agreements will decrease. On the other hand, consider that both examiners detect the presence of CE but on different surfaces of a same tooth. The disagreements between the two examiners will disappear at the tooth level. The level of agreement computed at the tooth level will consequently be higher than at the surface level. Because we expect the occurrence of a combination of the two disagreement situations, it is difficult to predict a general relationship between the level of agreement and the level of the hierarchy.

In the present paper, we propose two agreement indexes belonging to the kappa-like family to quantify the agreement between a pair of examiners in the context of multilevel data. One measure considers examiners randomly chosen from a population and the other considers that they are the only of interest. Then, a method to study the effect of factors on the level of agreement between a pair of examiners is introduced. Rather than simply estimate agreement indexes for various levels of the covariates, emphasis is given in quantifying the impact of the covariates on the level of agreement. This aims helping researchers to identify factors lowering the agreement level between pairs of examiners and find ways to improve it. We give an overview of previous research on the subject in Section 2 and describe the motivating data set in Section 3. We provide a short review of Cohen's and intraclass kappa coefficients and their relationship in Section 4, followed by the presentation of the proposed method in the case of

random and fixed examiners. We introduce the practical implementation of the method in Section 5 and apply it to the motivating data set in Section 6. Finally, we discuss the method in Section 7.

2. Previous research

Relatively few papers paid attention to the problem of agreement in the presence of hierarchical structures and are restricted to the simplest case of a two-level hierarchy. A possible reason for this sparse literature on the subject could be that it is common practice to summarize the data at the higher level of the hierarchy and then compute a classical agreement index. In the geographic atrophy example, we could decide to study agreement at the patient level rather than at the eye level. To summarize the eye's data at the patient level, two different rules can be envisaged: a patient is said to be positive when geographic atrophy is (1) detected in at least one eye or (2) present in both eyes. The two rules might lead to different conclusions [47]. Moreover, rules become more difficult to set up when there are more than two strata (i.e., S). Indeed, every value between 1 and S could be envisaged as cut off when summarizing the data at a higher hierarchical level. Finally, the loss of information (i.e., the number and the position of positive items at the deepest level of the hierarchy) occurring when summarizing data should not be neglected. To circumvent these problems when determining the agreement at the higher level of the hierarchy, Oden [43] proposed to take a weighted average of the agreement coefficients obtained in each strata of the deepest level, whereas Schouten [48] used a weighted agreement index with weights reflecting the strength of agreement between the two examiners. In the geographical atrophy example, a weight of 1.0 was given if the graders agreed on both eyes, a weight of 0.0 if the graders disagreed on both eyes, and a weight of 0.5 if the graders disagree on only one eye. Unfortunately, these agreement indexes do not permit to study the effect of continuous covariates on the level of agreement.

With the advances of the generalized linear mixed models, Klar *et al.* [25], Williamson *et al.* [26], and Gonin *et al.* [27] proposed GEE approaches to model coefficients of agreement of the kappa-like family according to a set of covariates in the presence of repeated measurements. In the mean time, Lipsitz *et al.* [21, 22] developed a method on an heuristic basis and showed that it is in fact equivalent to the GEE approach. Nevertheless, hierarchical and repeated measurements differ in the sense that, for multilevel data, the measurements are not repeated on the same statistical units but clustered. Finally, Ren *et al.* [49] proposed to estimate the intraclass kappa coefficient between several examiners using a multilevel generalized linear model and proposed using a bootstrap method to estimate its standard error.

More recently, agreement in the presence of a hierarchical structure was envisaged in a Bayesian statistical framework. Gajewski *et al.* [50] proposed a Bayesian hierarchical model with latent variables to estimate the inter-examiners reliability of ordinal observations with random examiner responses. Zhang and Cutter [51] used multivariate probit models for unbalanced data sets and then estimated the kappa coefficient on the basis of the posterior samples of the probit regression parameters and the covariance matrix. Finally, Hsiao *et al.* [52] derived an intraclass correlation coefficient among random examiner effects. Although these Bayesian methods take the hierarchical structure of the data into account, none of them permit to directly relate the obtained agreement coefficients to a set of continuous or categorical covariates.

3. Motivating data set

3.1. Epidemiological data set

The Signal Tandmobiel® project is a longitudinal (1996–2001) oral health project in Flanders (North of Belgium). At the first examination, the average age of the children was 7.1 years (SD = 0.4) and varied from 6.1 to 8.1 years. Sixteen trained dentists (examiners) conducted annual examinations on 4468 children (2315 boys and 2153 girls) from 179 primary schools, after they obtained parental consent. Data on oral hygiene and dietary habits were obtained through structured questionnaires, completed by the parents. The children received a clinical examination using the standardized and widely accepted criteria as recommended by the WHO [53] and based on the diagnostic criteria for caries prevalence surveys published by the British Association for the Study of Community Dentistry [54]. The clinical examinations took place in a mobile dental clinic, with a standard dental chair and dental artificial light. Detection was performed by the visual–tactile method, using a disposable mouth mirror and a WHO/CPITN type E probe. No radiographs were taken. For a more detailed description of the Signal Tandmobiel® study, we refer to [55].

3.2. Calibration data set

Training sessions for scoring CE were organized and the scoring behavior of each of 16 dental examiners was compared with that of the benchmark scorer (third author). During the study period (1996–2001), three calibration exercises for scoring CE (1996, 1998, and 2000), involving 92, 35, and 24 children respectively were organized. A large number of children was involved in 1996 compared with the other years. Four sessions were organized in 1996 with each session comprising approximately 25 children. Because of practical reasons, a more efficient organization was needed with only a single session for the years 1998 and 2000. Note that the age of the children for the calibration exercises of 1998 and 2000 was not recorded in the database. However, the ages of children examined in 1996, 1998, and 2000 were age matched with the school children in the first, third, and fifth classes, respectively. During the calibration exercises, children were not sampled at random from the main study. Rather, a school where a relatively high prevalence of CE could be expected was selected. At the end of each of the three calibration exercises, the sensitivity and specificity of each dental examiner *vis-a-vis* the benchmark scorer was determined. In the present work, we combine data of the three calibration exercises. We fit a hierarchical model of agreement, as specified in Section 4.2, to this calibration data set. The purpose of this research is to understand the factors that lower or decrease the agreement between a randomly chosen pair of examiners. In this study, we do not use data from the benchmark scorer to imitate the studies where such a scorer is not available. Because the left quadrant of the children was examined only in 1996, we restricted the analysis to the right quadrant. We excluded a total of 43 children from the analysis, because they were examined by only one dental examiner. The study population finally consisted of 108 (71.5%) children (1261 tooth and 5677 surfaces). There were 60 (55.6%) girls and 48 (44.4%) boys, 49 (45.4%) children were examined in 1996, 35 (32.4%) in 1998, and 24 (22.2%) in 2000.

4. Two-level agreement model

4.1. Cohen's kappa coefficient and intraclass kappa coefficient in the binary case

Cohen's kappa coefficient was initially defined as a descriptive statistic on an ad hoc basis and not in terms of population parameters [1]. However, Bloch and Kraemer [33] derived a population model in the case of a binary scale yielding Cohen's kappa coefficient as maximum likelihood estimator.

Consider a population of items (subjects or objects) $\mathcal{I}$. Let $Y_{i,r}$ be the random variable such that $Y_{i,r} = 1$ if examiner r ($r = 1, 2$) classifies a randomly selected item i of population $\mathcal{I}$ in category 1 and $Y_{i,r} = 0$ otherwise. Across the items in the population, $E(Y_{i,r}) = \pi_r$ and $\text{var}(Y_{i,r}) = \sigma_r^2 = \pi_r(1-\pi_r)$. If ρ denotes the correlation between $Y_{i,1}$ and $Y_{i,2}$, Table I corresponds to the population model.

Cohen's kappa coefficient is then defined as

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e} = \frac{2\rho\sigma_1\sigma_2}{1 - \pi_1\pi_2 - (1-\pi_1)(1-\pi_2)}, \tag{1}$$

where the probability of agreement π_o is the sum of the diagonal elements in Table I ($\pi_o = \pi_1\pi_2 + (1-\pi_1)(1-\pi_2) + 2\rho\sigma_1\sigma_2$) and the expected probability of agreement π_e is the product of the marginals ($\pi_e = \pi_1\pi_2 + (1-\pi_1)(1-\pi_2)$).

Suppose that two examiners classify a random sample of N items from population $\mathcal{I}$ on a binary scale. Let n_{ij} be the number of items classified in category i by the first examiner and category j by the second

Table I. Theoretical model in the case of two independent examiners and a binary scale.

	Examiner 2		
Examiner 1	1	2	
1	$E[Y_{i,1}Y_{i,2}]$ $\pi_1\pi_2 + \rho\sigma_1\sigma_2$	$E[Y_{i,1}(1-Y_{i,2})]$ $\pi_1(1-\pi_2) - \rho\sigma_1\sigma_2$	π_1
2	$E[(1-Y_{i,1})Y_{i,2}]$ $(1-\pi_1)\pi_2 - \rho\sigma_1\sigma_2$	$E[(1-Y_{i,1})(1-Y_{i,2})]$ $(1-\pi_1)(1-\pi_2) + \rho\sigma_1\sigma_2$	$1-\pi_1$
	π_2	$1-\pi_2$	1

one ($i, j = 1, 2$). Let $n_{i.}$ (resp. $n_{.j}$) be the total number of items classified in category i (resp. j) by the first (resp. second) examiner. The maximum likelihood estimator of κ is then expressed as

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}, \tag{2}$$

where $p_o = (n_{11} + n_{22})/N$ and $p_e = (n_{1.}n_{.1} + n_{2.}n_{.2})/N^2$.

The intraclass kappa coefficient can be viewed as a special case of Cohen's kappa coefficient where it is assumed that the ratings are interchangeable. In other words, the two examiners are assumed to have the same marginal probability distribution ($\pi_1 = \pi_2$). The resulting index is algebraically equivalent to Scott's index of agreement [56].

Let $\pi_1 = \pi_2 = \pi$, the probability of agreement becomes $\pi_{oI} = \pi^2 + (1-\pi)^2 + 2\rho\sigma^2$ and agreement expected by chance $\pi_e = \pi^2 + (1-\pi)^2$, leading to the *intraclass kappa coefficient*

$$\kappa_I = \frac{\pi_{oI} - \pi_{eI}}{1 - \pi_{eI}} = \rho. \tag{3}$$

The maximum likelihood estimator of κ_I is

$$\hat{\kappa}_I = \frac{p_{oI} - p_{eI}}{1 - p_{eI}}, \tag{4}$$

where $p_{oI} = (n_{11} + n_{22})/N$ and $p_{eI} = [(n_{1.} + n_{.1})/2N]^2 + [(n_{2.} + n_{.2})/2N]^2$ [3].

Cohen's kappa and the intraclass kappa coefficients are equal when there is no examiner bias, that is, when $n_{12} = n_{21}$. Therefore, it is recommended to use Cohen's kappa coefficient when the two examiners are fixed and cannot be considered as interchangeable, whereas the intraclass kappa coefficient is preferred when the absence of examiner bias can be assumed. This could be the case when two examiners come from a common population.

4.2. *Random examiners in the two-level case*

Consider a population $\mathcal{R}$ of examiners. We are interested in the agreement between a randomly chosen pair of examiners from that population on the classification of items on a binary scale. Let there be a random sample of R examiners, then we have $P = R(R-1)/2$ distinct pairs of examiners. Items are also supposed to belong to a population of items $\mathcal{I}$ with a two-level hierarchical structure in the sense that n_j items in the jth cluster ($j = 1, \ldots, N$) were classified in two categories by the R examiners. Each examiner thus classified a total of $\sum_{j=1}^{N} n_j$ items. For example, the presence of CE (yes/no) may be assessed on each of the 28 teeth of 108 children by 16 examiners ($R = 16, n_j = n = 28, j = 1, \ldots, 108$).

Suppose that we randomly choose a pair $p = (r_1, r_2)$ of examiners in the population $\mathcal{R}$. We omit subsequently the dependence notation of p on r_1, r_2 for notation simplicity. Let $Y_{ij,r}$ be the random variable equal to 1 if examiner r from pair p classifies item i from cluster j in category 1 and equal to 0 otherwise. We suppose that $Y_{ij,r} \sim \text{Bern}(\pi_{ij,r})$, where $\pi_{ij,r}$ is the probability of classifying item i from cluster j in category 1 for examiner r of pair p ($i = 1, \ldots, n_j,\ j = 1, \ldots, N,\ r = r_1, r_2,\ p = 1, \ldots, P$). We further assume that the examiners of pair p share common properties and are interchangeable ($\pi_{ij,r1} = \pi_{ij,r2} = \pi_{ij}^p$). This implies the existence of a common underlying probability of classifying items in the two categories of the scale for all examiners of population $\mathcal{R}$ ($E_{\mathcal{R}}(\pi_{ij}^p) = \pi_{ij}$).

Let Y_{ij}^p be the random variable such that $Y_{ij}^p = 1$ if the two examiners of the randomly chosen pair p agree on the classification of item i from cluster j. Suppose that $Y_{ij}^p \sim \text{Bern}(\pi_{o,ij}^p)$ ($i = 1, \ldots, n_j,\ j = 1, \ldots, N,\ p = 1, \ldots, P$). We have

$$Y_{ij}^p = Y_{ij,r_1}Y_{ij,r_2} + (1 - Y_{ij,r_1})(1 - Y_{ij,r_2}). \tag{5}$$

With the use of these notations, the agreement between the two examiners from pair p on item i from cluster j is defined by

$$\kappa_{ij}^p = \frac{\pi_{o,ij}^p - \pi_{e,ij}^p}{1 - \pi_{e,ij}^p}, \tag{6}$$

where $\pi_{e,ij}^p = \left(\pi_{ij}^p\right)^2 + \left(1 - \pi_{ij}^p\right)^2$ is the probability of agreement expected by chance between the two examiners of pair p. The agreement is thus of intraclass form because the examiners are assumed

to be interchangeable $\left(i.e., \pi_{ij}^{p} \text{ is common to examiners } r_1 \text{ and } r_2\right)$ and extends the classical definition given by Equation (3).

We next introduce covariates and show how they are related to the hierarchical agreement coefficient. We will propose a relationship between the hierarchical agreement index κ_{ij}^{p} and the covariates. However, because κ_{ij}^{p} is a ratio and the probability distribution of kappa-like agreement indexes is difficult to establish, we will rephrase the relationship to an equivalent set of two separate models. The first one directly models $\pi_{o,ij}^{p}$, whereas the second indirectly models $\pi_{e,ij}^{p}$, both according to a set of covariates.

Suppose that each item has an item/cluster specific $(f-1)\times 1$ covariate vector $\boldsymbol{x}_{ij}^{*}$ and let $\boldsymbol{x}_{ij}$ denote the $f \times 1$ vector $\left(1, \boldsymbol{x}_{ij}^{*'}\right)'$. We propose to link the agreement index to the covariates with

$$g\left(\kappa_{ij}^{p} | \boldsymbol{x}_{ij}, \alpha_j\right) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \alpha_j, \tag{7}$$

where $g(.)$ is a link function, $\boldsymbol{\beta}$ a $f \times 1$ vector of parameters, and $\alpha_j \sim N\left(0, \sigma_C^2\right)$ a random intercept varying for each cluster.

Because agreement coefficients belonging to the kappa-like family vary between -1 and 1 [1], we have that $\left(1+\kappa_{ij}^{p}\right)/2$ vary between 0 and 1. A natural choice for the link function to avoid constraints on the parameters $\boldsymbol{\beta}$ is then the complementary log–log function, that is,

$$g(.) = \ln[-\ln(1-.)]. \tag{8}$$

Because

$$\ln\left[-\ln\left(1-\frac{1+\kappa_{ij}^{p}}{2}\right)\right] = \ln\left[-\ln\left(\frac{1-\kappa_{ij}^{p}}{2}\right)\right] = \ln\left[-\ln\left(\frac{1-\pi_{o,ij}^{p}}{2\left(1-\pi_{e,ij}^{p}\right)}\right)\right], \tag{9}$$

Equation (7) is equivalent to

$$-\ln\left(1-\pi_{o,ij}^{p} | \boldsymbol{x}_{ij}, \alpha_j\right) = -\ln\left[2\left(1-\pi_{e,ij}^{p}\right)\right] + \exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \alpha_j\right). \tag{10}$$

For known $\pi_{e,ij}^{p}$, this corresponds to a hierarchical generalized linear model with a known offset but the probability $\pi_{e,ij}^{p}$ is rarely known in practice. One possibility is to estimate $\pi_{e,ij}^{p}$ and use this estimate in the model of $\pi_{o,ij}^{p}$, similar to that in [21, 22]. This implies

$$-\ln\left(1-\pi_{o,ij}^{p} | \boldsymbol{x}_{ij}, \alpha_j\right) = -\ln\left[2\left(1-\hat{\pi}_{e,ij}^{p}\right)\right] + \exp\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \alpha_j\right). \tag{11}$$

To obtain $\hat{\pi}_{e,ij}^{p}$, we first estimate the marginal probabilities $\pi_{ij,r}$ with the two-level hierarchical model

$$\text{logit}(\pi_{ij,r} | \boldsymbol{x}_{ij}, \delta_j, \eta_r) = \boldsymbol{x}_{ij}'\boldsymbol{\lambda} + \delta_j + \eta_r, \tag{12}$$

where $\boldsymbol{\lambda}$ is a vector of parameters, $\eta_r \sim N\left(0, \sigma_R^2\right)$ the random effect relative to the examiners, and $\delta_j \sim N\left(0, \sigma_d^2\right)$ the random intercept pertaining to the clusters ($r = r_1, r_2$). Then, the marginal probability π_{ij}^{p} is estimated by

$$\hat{\pi}_{ij}^{p} = \frac{\hat{\pi}_{ij,r_1} + \hat{\pi}_{ij,r_2}}{2}. \tag{13}$$

Finally, we obtain

$$\hat{\pi}_{e,ij}^{p} = \left(\hat{\pi}_{ij}^{p}\right)^2 + \left(1-\hat{\pi}_{ij}^{p}\right)^2. \tag{14}$$

The agreement index κ_{ij}^{p} is thus directly related to the covariates through the model parameters $\boldsymbol{\beta}$ $\left(\text{common to the model of } \kappa_{ij}^{p} \text{ and } \pi_{o,ij}^{p}\right)$ and indirectly through the model parameters $\boldsymbol{\lambda}$. If the sign of a parameter estimate $\hat{\beta}_m$ is positive (negative), the level of agreement increases (decreases) as the value of the mth covariate increases.

To estimate the model parameters given in Equations (11) and (12), the maximum likelihood estimates will be replaced by sampled values of a posterior probability distribution obtained in a Bayesian framework, as described in Section 5.

4.3. Fixed examiners in the two-level case

In the previous section, we developed a method to study the effect of covariates on the hierarchical agreement index on a pairwise basis when interest was to generalize the finding to any pair of examiners in the population. If examiners do not share common properties and are the only examiners of interest, they can be considered as fixed. Therefore, suppose now that instead of a sample of $R > 2$ examiners, we only dispose of the data of two particular examiners, namely examiners 1 and 2, and consider them as fixed.

Let $Y_{ij,r}$ be defined as in the previous paragraph ($i = 1, \ldots, n_j,\ j = 1, \ldots, N,\ r = 1, 2$). The agreement between the two examiners on item i from cluster j is defined by

$$\kappa_{ij} = \frac{\pi_{o,ij} - \pi_{e,ij}}{1 - \pi_{e,ij}}, \tag{15}$$

where $\pi_{o,ij} = Y_{ij,1}Y_{ij,2} + (1 - Y_{ij,1})(1 - Y_{ij,2})$ and $\pi_{e,ij} = \pi_{ij,1}\pi_{ij,2} + (1 - \pi_{ij,1})(1 - \pi_{ij,2})$ are the probability of agreement and the probability of agreement expected by chance between the two examiners, respectively. This extends the classical definition of Cohen's kappa given in Equation (1) to the two-level case.

The relationship given by Equation (7) thus becomes

$$\ln\left[-\ln\left(\frac{1-\kappa_{ij}}{2}|\boldsymbol{x}_{ij}, \alpha_j\right)\right] = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \alpha_j, \tag{16}$$

where $\boldsymbol{\beta}$ is the vector of parameters and $\alpha_j \sim N\left(0, \sigma_C^2\right)$ a random intercept varying for each cluster. This model simplifies to

$$-\ln(1 - \pi_{o,ij}|\boldsymbol{x}_{ij}, \alpha_j) = -\ln[2(1 - \hat{\pi}_{e,ij})] + \exp\left(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \alpha_j\right), \tag{17}$$

where $\hat{\pi}_{e,ij}$ is obtained by

$$\hat{\pi}_{e,ij} = \hat{\pi}_{ij,1}\hat{\pi}_{ij,2} + (1 - \hat{\pi}_{ij,1})(1 - \hat{\pi}_{ij,2}). \tag{18}$$

The estimates of the marginal probabilities $\hat{\pi}_{ij,1}$ and $\hat{\pi}_{ij,2}$ are obtained by a two-level hierarchical logistic model

$$\text{logit}(\pi_{ij,r}|\boldsymbol{x}_{ij}, \delta_j) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\lambda}_r + \delta_j, \tag{19}$$

where $\boldsymbol{\lambda}_r$ is the vector of parameters and $\delta_j \sim N\left(0, \sigma_d^2\right)$ the random effect relative to the clusters ($r = 1, 2$). We assume different vectors of parameters $\boldsymbol{\lambda}_r$ in contrast to the approach in Section 4.2.

5. MCMC approach

To estimate the parameters in the models mentioned previously, we used a Bayesian approach. In a Bayesian approach, the prior knowledge about the parameters is combined with the observed data (likelihood) to yield the posterior distribution. We obtained the posterior summary measures of the parameters by using the MCMC sampling approach (e.g., [57]). We performed the MCMC calculations in OpenBUGS [58]. We used non-informative priors expressing that we do not have prior information on the parameters. For the regression coefficients, we assumed vague independent priors to follow a normal distribution with mean 0 and large variance, that is, $\beta \sim \text{N}(0, 10^6)$. The prior distribution for all the standard deviations of the random effects, that is, mouth, tooth, and examiner was taken as uniform, for example, $\sigma \sim \text{Uniform}[0, 100]$. We ran three parallel MCMC chains, each for 30,000 iterations for all the models with a burn-in period of 2500 iterations. We checked the convergence of these MCMC by using the CODA package in R [59]. In particular, we used the Gelman and Rubin's diagnostics measure R [59], and this value was close to 1 for all the parameters, which means there was no evidence against convergence.

6. Application

The following data are part of the Signal Tandmobiel® study described in Section 3. The presence of CE (yes/no) was assessed in the right quadrant of 108 children by 16 dental examiners on each surface

of the tooth. The resulting data thus have three levels of hierarchy; namely, the surface, the tooth, and the mouth. Characteristics related to the children [gender (boy, girl), year of examination (1996, 1998, 2000)], to the tooth [dentition type (permanent, deciduous), jaw (upper, lower), type (canine, molar, premolar, incisor)], and to the surface [location (distal, mesial, lingual, occlusal, buccal)] were recorded. The aim is to study the influence of these factors on the agreement obtained between any pair of examiners and more particularly to detect factors lowering the level of agreement. Because the 16 examiners did not assess all the children, we dispose of data from 74 of the 120 possible pairs of examiners. Because deciduous molars are replaced by permanent premolars, molars and premolars were grouped in a single category, namely '(pre)molar'.

The intraclass kappa coefficient, as described in Section 4.1, was computed for each pair of examiners assessing at least two children at the mouth, tooth, and surface levels (Table II). The prevalence of CE decreased when going deeper in the level of the hierarchy. It was approximately 52.9%, 11.3%, and 5.1% at the mouth, tooth, and surface levels, respectively. However, the dependence of the data needs to be taken into account when computing intraclass kappa coefficients at the surface and tooth levels because the definition of chance agreement no longer applies. Similar agreement levels were found on the surface and tooth levels. On the other hand, the results slightly differed between the surface and mouth levels. This could be explained by the fact that a tooth is composed of four or five surfaces, whereas a half mouth included between 40 and 54 surfaces in our data. When summarizing the information on CE at the mouth level, the number of agreements on caries-free surfaces is thus reduced to a greater extent than at the tooth level.

6.1. Modeling agreement

We then applied the method described in Section 4.2 to take the hierarchical structure into account. If the subscript s denotes the surface, t the tooth, and m the mouth level, the hierarchical agreement index can be related to the covariates by

$$\ln\left[-\ln\left(\frac{1-\kappa^{p}_{stm}}{2}|\boldsymbol{x}_{stm},\alpha_t,\gamma_m\right)\right]=\boldsymbol{x}'_{stm}\boldsymbol{\beta}+\alpha_t+\gamma_m, \tag{20}$$

where $\alpha_t \sim N(0,\sigma^2_T)$ is the random effect relative to the tooth level and $\gamma_m \sim N(0,\sigma^2_M)$ to the mouth level. Because the dentition type (deciduous, permanent), subsequently denoted by D_{stm}, is a covariate varying within clusters (i.e., mouths), the covariate was decomposed in between-cluster ($\overline{D}_m$) and within-cluster component ($D_{stm}-\overline{D}_m$) to obtain a proper interpretation of the regression coefficients [60,61]. The regression coefficient relative to the between-cluster covariate refers to the effect of increasing the proportion of permanent teeth in a mouth by one unit, whereas the within-cluster regression coefficient refers to the effect of the actual teeth type within a given mouth. Note that the between-cluster component is highly related to the age of the child. The proportion of permanent teeth in the mouth of one child was 40.1 ± 12.6 %, 64.0 ± 16.3%, and 90.9 ± 13.9% for children in the first, third, and fifth classes, respectively. Therefore, only the type of tooth was used as covariate in the models. We provide the posterior mean of the model parameters in Table III with 95% credibility interval and the

Table II. Signal Tandmobiel® study: classical intraclass kappa distribution obtained for the 74 pairs of dental examiners assessing caries experience in the mouths of 108 children.

	Level	Mean	SD	Median	Range
Intraclass kappa coefficient ($\hat{\kappa}_I$)	Surface	0.78	0.098	0.78	0.55–0.93
	Tooth	0.79	0.093	0.79	0.45–0.94
	Mouth	0.73	0.20	0.75	−0.33–1.00
Observed proportion of agreement (p_{oI})	Surface	0.98	0.0081	0.98	0.94–0.99
	Tooth	0.96	0.018	0.96	0.92–0.99
	Mouth	0.87	0.086	0.89	0.50–1.00
Expected proportion of agreement (p_{eI})	Surface	0.92	0.022	0.93	0.83–0.95
	Tooth	0.82	0.036	0.83	0.68–0.88
	Mouth	0.52	0.037	0.52	0.50–0.72

Summary is presented at the surface, mouth, and tooth levels.

Table III. Signal Tandmobiel® study: parameter estimates of the hierarchical complementary log-log model for kappa and through the method of Lipsitz *et al.*

		Proposed approach			Lipsitz *et al.* method		
Parameter		Posterior mean (SD)	2.5%	97.5%	GEE Estimate (SD)	2.5%	97.5%
Fixed effects	Intercept	0.77 (0.61)	−0.46	2.02	0.70 (0.20)	0.30	1.10
Gender	Girl	−0.0040 (0.20)	−0.38	0.37	0.12 (0.12)	−0.11	0.35
	Boy						
Dentition type (between)	Permanent	−0.28 (0.51)	−1.29	0.72	−0.085 (0.24)	−0.55	0.38
	Deciduous						
Dentition type (within)	Permanent	−0.28 (0.21)	−0.71	0.14	−0.25 (0.17)	−0.58	0.077
	Deciduous						
Jaw	Upper	−0.14 (0.17)	−0.48	0.19	−0.11 (0.091)	−0.29	0.073
	Lower						
Type	Canine	−0.38 (0.65)	−1.72	0.88	–[a]		
	(Pre)molar	0.15 (0.41)	−0.94	0.65	–[a]		
	Incisor						
Random effects	σ^2_{mouth}	0.27 (0.14)	0.081	0.61			
	σ^2_{tooth}	0.72 (0.20)	0.42	1.18			

[a]The algorithm did not converge with type of tooth in the model.

results provided by the GEE approach. For the sake of comparability, we adapted the GEE method of Lipsitz *et al.* [22] to make use of the complementary log–log as link function.

As seen in Table III, the results of the Bayesian approach and the GEE approach lead to the same conclusion. The level of agreement between a pair of dental examiners assessing the presence of CE was not related to any of the available covariates. Note that in the multilevel approach, the magnitude of the random effects is not negligible compared with that of the fixed effect. This reflects the existence of an heterogeneity in the agreement levels for the children and the teeth, remaining unexplained after adjustment for the available covariates. We computed the agreement level between pairs of examiners for the (pre)molars in the lower jaw of an median boy (i.e., all random effects equal to 0), aged 8 years (i.e., with 64% of permanent teeth in our data set) with respect to the type of tooth (deciduous or permanent) (Figure 1). The posterior median is 0.70 for deciduous teeth and 0.49 for permanent teeth.

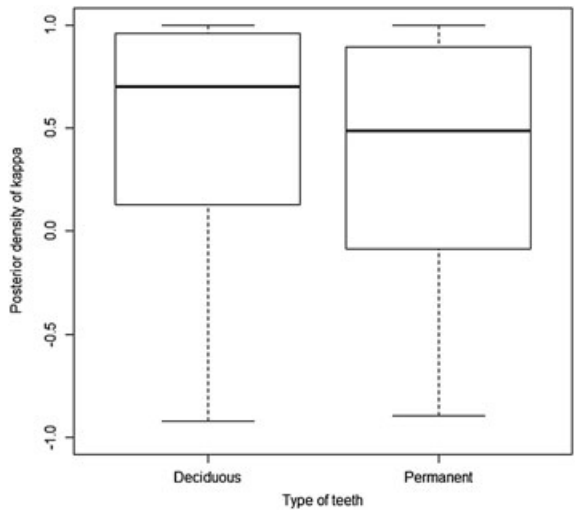

Figure 1. Signal Tandmobiel® study: posterior density of the agreement coefficient for a (pre)molar in the lower jaw of a median 8-year-old boy (i.e., with 64% of permanent teeth) according to the type of teeth (permanent or deciduous).

Table IV. Signal Tandmobiel® study: parameter estimates obtained using the hierarchical logistic regression and Lipsitz *et al.* method for the marginal probabilities of CE assessment.

		Proposed approach			Lipsitz *et al.* method					
Parameter		Posterior mean (SD)	2.5%	97.5%	GEE estimate (examiner 1)	2.5%	97.5%	GEE estimate (examiner 2)	2.5%	97.5%
Fixed effects	Intercept	−13.5 (1.44)	−16.4	−10.7	−5.40 (0.95)	−7.26	−3.54	−5.08 (0.94)	−6.92	−3.24
Gender	Girl Boy	−1.16 (0.71)	−2.51	0.24	0.085 (0.36)	−0.63	0.79	0.012 (0.39)	−0.76	0.78
Dentition type (between)	Permanent Deciduous	0.060 (1.60)	−2.96	3.33	1.01 (0.83)	−0.61	2.64	0.97 (0.89)	−0.77	2.70
Dentition type (within)	Permanent Deciduous	−3.27 (0.57)	−4.41	−2.17	−1.78 (0.55)	−2.85	−0.70	−1.91 (0.54)	−2.96	−0.85
Jaw	Upper Lower	−0.96 (0.41)	−1.74	−0.16	−0.38 (0.16)	−0.70	−0.067	−0.39 (0.16)	−0.71	−0.076
Type	Canine (Pre)molar Incisor	0.19 (1.11) 6.89 (0.88)	−2.00 5.27	2.39 8.72	−0.36 (0.65) 2.01 (0.24)	−1.64 1.53	0.92 2.49	−0.44 (0.59) 1.74 (0.24)	−1.59 1.27	0.71 2.21
Random effects	σ^2_{mouth}	10.8 (2.77)	6.3	17.1						
	σ^2_{tooth}	17.7 (2.33)	13.5	22.6						
	$\sigma^2_{\text{examiner}}$	0.19 (0.098)	0.071	0.43						

6.2. *Modeling the probability of positive caries experience assessment*

Table IV shows the posterior distribution of the probability of detecting CE, namely $\pi_{stm,r}$, as given by the hierarchical logistic model

$$\text{logit}(\pi_{stm,r}|\boldsymbol{x}_{stm}, \delta_t, \zeta_m, \eta_r) = \boldsymbol{x}'_{stm}\boldsymbol{\lambda} + \delta_t + \zeta_m + \eta_r. \tag{21}$$

In the GEE approach, the probability of assessing the presence of CE is different for the two examiners of a pair, whereas it is assumed to be common to all examiners in the proposed approach.

According to the multilevel logistic regression, the probability of detecting CE was higher for a deciduous than a permanent tooth, on the lower jaw than the upper jaw, and on premolars than on incisors. When calculating the proportion of variance attributable to each level of the hierarchy in the multilevel logistic model following the latent variable approach [62], we found that 55% of the variance was attributable to the variation between teeth within individuals, 34% was attributable to variation at mouth level, and only 0.58% was attributable to the examiners. The degree of similarity in the assessment of CE scoring, as assessed by the intraclass correlation coefficient, was equal to 0.89 for surfaces of the same tooth as compared with 0.34 for surfaces of different teeth within a given mouth assessed by a particular examiner. The results of the multilevel and the GEE approach are similar, although they have a different interpretation.

Figure 2 shows the posterior distribution of the examiner random effects corresponding to model (21) for each examiner. Almost all posterior intervals included the value 0, indicating that these random effects do not deviate from 0. That overall exchangeability is however perhaps not satisfied because examiner 9 clearly deviates from the other examiners.

We computed the posterior probability of being classified as experiencing caries for (pre)molars in the lower jaw of a median boy, aged 8 years (i.e., with 64% of permanent teeth in our data set) with respect to the type of tooth (deciduous or permanent) (Figure 3). The posterior median was 0.98% for deciduous teeth and 0.037% for permanent teeth.

7. Discussion

Hierarchical data are frequently encountered in research and need specific analysis methods. In this paper, we defined two hierarchical agreement indexes when interest lies in the pairwise agreement

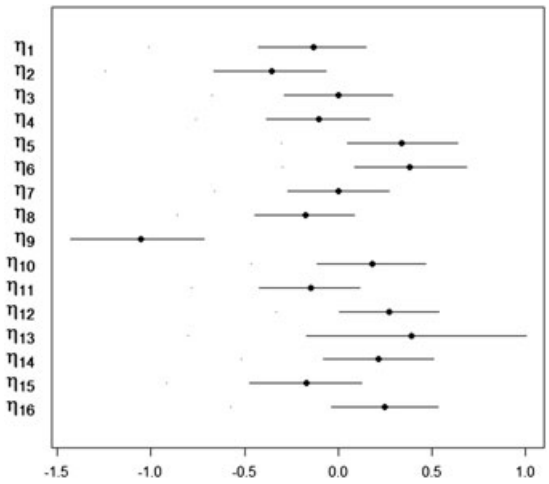

Figure 2. Signal Tandmobiel® study: posterior distribution of the examiner random effects obtained in the hierarchical logistic model of the marginal for each of the 16 examiners (posterior mean (•) and 95% credibility interval).

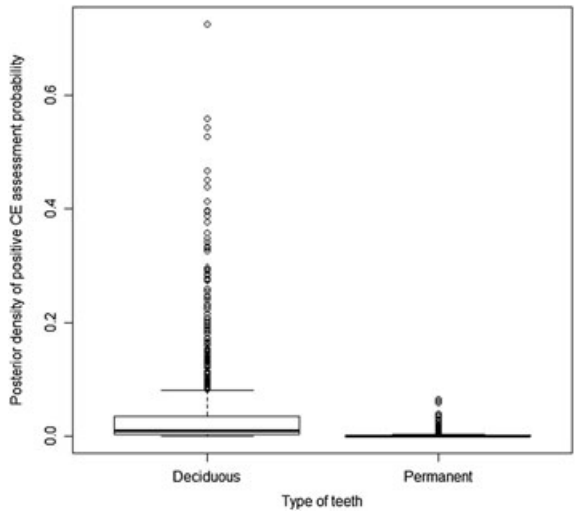

Figure 3. Signal Tandmobiel® study: posterior probability distribution for being classified as experiencing caries on (pre)molars in the lower jaw of a median 8-year-old boy (i.e., with 64% of permanent teeth) according to the type of tooth.

between examiners on a binary scale. The first index quantifies the underlying common agreement existing between any pair of examiners belonging to a common population of examiners. The examiners are assumed to be interchangeable, providing an agreement index of the intraclass form. The second one was derived in the context of fixed examiners and is of Cohen's form. It is designed to quantify pairwise agreement when the examiners in the study are the only of interest. This method easily extends to the case of $R > 2$ fixed examiners, as using the 2-wise definition of agreement introduced by Conger and David and Fleiss [4, 6]. We then proposed a method to directly evaluate the impact of categorical and continuous covariates, defined at different levels of the hierarchy, on the obtained agreement indexes. In practice, the probability of agreement expected by chance $\pi_{e,ij}$ is estimated using two hierarchical logistic models and used as a known offset in a hierarchical complementary log–log model relating the observed probabilities of agreement $\pi_{o,ij}$ to a set of covariates. The regression part of these three models are obtained simultaneously using a Bayesian approach. The posterior mean of the parameters obtained when modeling the probabilities $\pi_{o,ij}$ corresponds to the posterior mean of parameters obtained by modeling directly the hierarchical agreement index to the same set of covariates. This method permits to identify factors lowering the level of agreement between examiners on a pairwise basis. In our example,

there is no evidence of an effect of the recorded covariates on the agreement between a pair of examiners assessing the presence of CE. The method has the advantage, by making a direct relationship between a set of categorical and continuous covariates and the agreement index, to permit an immediate evaluation of the influence of these covariates on the level of agreement. Nevertheless, the use of the complementary log–log link function presents one drawback. It does not allow for an obvious interpretation of the model parameters. Because agreement coefficients corrected for chance vary between -1 and $+1$, Fisher Z-transform was also envisaged. However, besides no simple interpretation of the parameters, it was not possible to isolate the probability of agreement $\pi_{o,ij}$ in a simple way.

According to the number of iterations, the convergence was relatively fast (30,000 iterations) but the computation time per iteration was rather slow. This could be explained by the complexity of the model and the size of the data set. Indeed, we estimated two hierarchical logistic models and a complementary log–log hierarchical model simultaneously using information of 5677 surfaces assessed by several examiners, leading to 45,372 observations. The estimation process will be faster with smaller data sets and less hierarchical levels. The implementation of the method of Lipsitz *et al.*, on the other hand, presented another drawback. The standard errors of the model parameters had to be computed using the sandwich estimator. Because the probability of detecting CE was very low at the surface level, the jackknife procedure necessary to compute the sandwich estimators led to quasi-complete separation problems and resulted in the non-convergence of the procedure when modeling the agreement level. One other major limitation of the GEE approach is that the examiners are always considered as fixed [27], preventing generalization to other raters. The Bayesian and the GEE approach led to similar conclusions in our example, but this could not always be the case. Moreover, results should be interpreted conditionally on the random effects in the multilevel approach and have a marginal interpretation in the GEE model.

In conclusion, we proposed a method to directly evaluate the effect of covariates on the level of agreement in the hierarchical framework. This could help researchers to identify factors influencing negatively the agreement between pairs of examiners. Further research is needed to evaluate the minimum number of examiners needed to provide a representative sample of the population of examiners.

Acknowledgements

The Catholic University Leuven (research grant OT/05/60) partially supported this investigation; Unilever, Belgium, supported the data collection. The Signal Tandmobiel® project has the following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands and L-Biostat, Catholic University of Leuven, Leuven, Belgium), and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

1. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
2. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; **70**:213–220.
3. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 1979; **44**:461–472.
4. Conger AJ. Integration and generalization of kappas for multiple raters. *Psychological Bulletin* 1980; **88**:322–328.
5. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd ed. John Wiley: New York, 1981.
6. Davies M, Fleiss JL. Measuring agreement for multinomial data. *Biometrics* 1982; **38**:1047–1051.
7. Schuster CS, Smith DA. Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. *Psychometrika* 2005; **70**:135–146.
8. Schouten HJA. Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica* 1982; **36**:45–61.
9. Vanbelle S, Albert A. Agreement between an isolated rater and a group of raters. *Statistica Neerlandica* 2009; **63**:82–100.
10. Vanbelle S, Albert A. Agreement between two independent groups of raters. *Psychometrika* 2009; **74**:477–491.
11. Graham P. Modelling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine* 1995; **14**:299–310.
12. Perkins SM, Becker MP. Assessing rater agreement using marginal association models. *Statistics in Medicine* 2002; **21**:1743–1760.
13. Uebersax JS. Validity inferences from interobserver agreement. *Psychological Bulletin* 1988; **104**:405–416.
14. Agresti A. Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research* 1992; **1**:201–218.

15. Schuster C, Smith DA. Indexing systematic rater agreement with a latent-class model. *Psychological Methods* 2002; **7**:384–395.
16. Schuster C, Smith DA. Estimating with a latent class model the reliability of nominal judgments upon which two raters agree. *Educational and Psychological Measurement* 2006; **66**:739–747.
17. Coughlin SS, Pickle LW, Goodman MT, Wilkens LR. The logistic modeling of interobserver agreement. *Journal of Clinical Epidemiology* 1992; **45**:1237–1241.
18. Shoukri MM, Martin SW, Mian IUH. Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statistics in Medicine* 1995; **14**:83–99.
19. Shoukri MM, Mian IUH. Maximum likelihood estimation of the kappa coefficient from bivariate logistic regression. *Statistics in Medicine* 1996; **15**:1409–1419.
20. Barlow W. Measurement of interrater agreement with adjustment for covariates. *Biometrics* 1996; **52**:695–702.
21. Lipsitz SR, Parzen M, Fitzmaurice GM, Klar N. A two-stage logistic regression model for analyzing inter-rater agreement. *Psychometrika* 2003; **68**:289–298.
22. Lipsitz SR, Stuart R, Williamson J, Klar N, Ibrahim J, Parzen M. A simple method for estimating a regression model for κ between a pair of raters. *Journal of the Royal Statistical Society, Series A* 2001; **164**:449–465.
23. Thomson JR. Estimating equations for kappa statistics. *Statistics in Medicine* 2001; **20**:2895–2906.
24. Williamson JM, Manatunga AK. Assessing interrater agreement from dependent data. *Biometrics* 1997; **53**:707–714.
25. Klar N, Lipsitz SR, Ibrahim JG. An estimating equations approach for modelling kappa. *Biometrical Journal* 2000; **42**:45–58.
26. Williamson JM, Lipsitz SR, Manatunga AK. Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* 2000; **1**:191–202.
27. Gonin R, Lipsitz SR, Fitzmaurice GM, Molenberghs G. Regression modelling of weighted κ by using generalized estimating equations. *Journal of the Royal Statistical Society, Series C* 2000; **49**:1–18.
28. Barnhart HX, Williamson JM. Weighted least-squares approach for comparing correlated kappa. *Biometrics* 2002; **58**:1012–1019.
29. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* 1988; **41**:949–958.
30. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology* 1990; **43**:543–549.
31. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 1990; **43**:551–558.
32. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 1993; **46**:423–429.
33. Bloch DA, Kraemer HC. 2×2 kappa coefficients: measures of agreement or association. *Biometrics* 1989; **45**:269–287.
34. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology* 2005; **58**:655–661.
35. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology* 2000; **53**:499–503.
36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–174.
37. Bishop YYM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis. Theory and Practice*. MIT Press: Cambridge, 1975.
38. Fleiss JL, Davies M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology* 1982; **115**:841–845.
39. Garner JB. The standard error of Cohen's kappa. *Statistics in Medicine* 1991; **10**:767–775.
40. Blackman NJ, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* 2000; **19**:723–741.
41. Lesaffre E, Mwalili SM, Declerck D. Analysis of caries experience taking inter-observer bias and variability into account. *Journal of Dental Research* 2004; **83**:951–955.
42. Mielke PW, Berry KJ, Johnston JE. Resampling probability values for weighted kappa with multiple raters. *Psychological Reports* 2008; **102**:606–613.
43. Oden NL. Estimating kappa from binocular data. *Statistics in Medicine* 1991; **10**:1303–1311.
44. Lesaffre E, Küchenhoff H, Mwalili SM, Declerck D. On the estimation of the misclassification table for finite count data with an application in caries research. *Statistical Modelling* 2009; **9**:99–118.
45. Agbaje JO, Mutsvari T, Lesaffre E, Declerck D. Measurement, analysis and interpretation of examiner reliability in caries experience surveys: some methodological thoughts. *Clinical Oral Investigations* 2012; **16**:117-127.
46. Mutsvari T, Lesaffre E, García-Zattera MJ, Diya L, Declerck D. Factors that influence data quality in caries experience detection: a multilevel modeling approach. *Caries research* 2010; **44**:438–444.
47. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977; **33**:363–374.
48. Schouten HJA. Estimating kappa from binocular data and comparing marginal probabilities. *Statistics in Medicine* 1993; **12**:2207–2217.
49. Ren S, Yang S, Lai S. Intraclass correlation coefficients and bootstrap methods of hierarchical binary outcomes. *Statistics in Medicine* 2006; **25**:3576–3588.
50. Gajewski BJ, Hart S, Bergquist-Beringer S, Dunton N. Inter-rater reliability of pressure ulcer staging: ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine* 2007; **26**:4602–4618.
51. Zhang X, Cutter G. A Bayesian method of estimating kappa coefficient with application to a rheumatoid arthritis study. *Communications in Statistics - Theory and Methods* 2009; **38**:3432–3444.
52. Hsiao CK, Chen PC, Kao WH. Bayesian random effects for interrater and test-retest reliability with nested clinical observations. *Journal of Clinical Epidemiology* 2011; **64**:808–814.
53. WHO. *Oral Health Surveys: Basic Methods*, 4th ed. World Health Organization: Geneva, 1997.

54. Pine CM, Pitts NB, Nugent ZJ. British Association for the Study of Community Dentistry (BASCD) guidance on the statistical aspects of training and calibration of examiners for surveys of child dental health. A BASCD coordinated dental epidemiology programme quality standard. *Community Dental Health* 1997; **14 (Suppl 1)**:18–29.
55. Vanobbergen J, Martens L, Lesaffre E, Declerck D. The Signal Tandmobiel® project, a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry* 2000; **2**:87–96.
56. Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 1955; **19**:321–325.
57. Spiegelhalter D, Thomas A, Best N, Gilks W. *BUGS 0.5: Bayesian Inference Using Gibbs Sampling - Manual (version ii)*. Medical Research Council Biostatistics Unit: Cambridge, 1996.
58. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
59. Plummer M, Best N, Cowles K, Vines K. coda: Output Analysis and Diagnostics for MCMC. R package version 0.13-3, 2007. (Available from: http://CRAN.Rproject.org/)
60. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; **54**:638–645.
61. Mancl LA, Leroux BG, DeRouen TA. Between-subject and within-subject statistical information in dental research. *Journal of Dental Research* 2000; **79**:1778–1781.
62. Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. *Understanding Statistics* 2002; **1**:223–231.