Clin Oral Invest (2012) 16:117–127
DOI 10.1007/s00784-010-0475-x

ORIGINAL ARTICLE

Measurement, analysis and interpretation of examiner reliability in caries experience surveys: some methodological thoughts

Jimoh Olubanwo Agbaje · Timothy Mutsvari · Emannuel Lesaffre · Dominique Declerck

Received: 10 December 2009 / Accepted: 28 September 2010 / Published online: 13 October 2010

Abstract Data obtained from calibration exercises are used to assess the level of agreement between examiners (and the benchmark examiner) and/or between repeated examinations by the same examiner in epidemiological surveys or large-scale clinical studies. Agreement can be measured using different techniques: kappa statistic, percentage agreement, dice coefficient, sensitivity and specificity. Each of these methods shows specific characteristics and has its own shortcomings. The aim of this contribution is to critically review techniques for the measurement and analysis of examiner agreement and to illustrate this using data from a recent survey in young children, the Smile for Life project. The above-mentioned agreement measures are influenced (in differing ways and extents) by the unit of analysis (subject, tooth, surface level) and the disease level in the validation sample. These effects are more pronounced for percentage agreement and kappa than for sensitivity and specificity. It is, therefore, important to include information on unit of analysis and disease level (in validation sample) when reporting agreement measures. Also, confidence intervals need to be included since they indicate the reliability of the estimate. When dependency among observations is present [as is the case in caries experience data sets with typical hierarchical structure (surface–tooth–subject)], this will influence the width of the confidence interval and should therefore not be ignored. In this situation, the use of multilevel modelling is necessary. This review clearly shows that there is a need for the development of guidelines for the measurement, interpretation and reporting of examiner reliability in caries experience surveys.

Keywords Caries experience · Agreement measurement · Percentage agreement · Dice coefficient · Kappa · Sensitivity · Specificity

J. O. Agbaje (✉) · D. Declerck
School of Dentistry, Oral Pathology and Maxillofacial Surgery,
Catholic University Leuven,
Kapucijnenvoer 7 blok a bus 7001,
3000 Leuven, Belgium
e-mail: joagbaje@gmail.com

T. Mutsvari · E. Lesaffre
L-Biostat, Catholic University of Leuven,
U.Z. St. Rafael 2nd floor Kapucijnenvoer 35,
3000 Leuven, Belgium

E. Lesaffre
Department of Biostatistics,
Erasmus University Rotterdam Erasmus Medical Centre,
Dr. Molewaterplein 50,
3015 GE Rotterdam, the Netherlands

Introduction

Epidemiological surveys and large scale clinical studies often involve multiple examiners. In order to safeguard reliability, comparability and validity of the obtained results, methodological aspects need to be standardised. This applies also to surveys with repeated measurements in the same individuals, e.g. in follow-up studies.

In the field of caries experience (CE) screening, standardisation guidelines were proposed by different instances. Most widely used protocols are those issued by the World Health Organisation (WHO), the British Association for the Study of Community Dentistry (BASCD) and more recently the International Caries Detection and Assessment System (ICDAS) [1–3]. Each of them includes guidelines on the training of examiners and the organisation of calibration sessions, involving a benchmark examiner or *gold standard*. A benchmark examiner is an experienced

assessor or a validated measuring instrument which is assumed to be error free or nearly so, while a gold standard is an instrument or technique that is regarded as reflecting the absolute truth. In CE surveys, the use of a gold standard is not feasible since this would imply the extraction of teeth. For this reason, invariably, a benchmark scorer is used. Data obtained from calibration exercises are used to assess the level of agreement among examiners (and the benchmark examiner; inter-examiner agreement; reproducibility) and/or between repeated examinations of individuals by the same examiner (intra-examiner agreement; consistency).

The inclusion of information on the outcome of agreement measurement in survey reports allows assessment of the validity of the results obtained; however, different techniques can be used to measure the level of agreement among examiners. Each of these methods shows specific characteristics, has its own shortcomings and needs to be interpreted accordingly. It is important to consider these issues when selecting a specific method and when reporting and interpreting the results.

The WHO advises the calculation of percentage agreement between scores allocated by two examiners or two examining sessions [3]; however, when the prevalence of disease is low (which is often the case in CE surveys), it is known that this measure may not appropriately express reproducibility. In this case, WHO proposes the use of kappa to assess the overall agreement among examiners.

BASCD [4] proposes a stepwise approach. Three essential steps for data comparison are recommended. In the first phase, a comparison of total numbers of affected teeth (or surfaces) by subject and examiner is undertaken making use of simple tables. This allows detection of most problems, e.g. consistent under- or overscoring. A second phase consists of calculation of mean values obtained by the different examiners and the size and direction of the deviation from the mean value obtained by the benchmark examiner. Deviation from the mean value obtained by the benchmark examiner will show if/when an examiner is scoring too high or too low. The calculation of group means and 95% confidence limits forms the third phase. The group mean is the mean dmft/DMFT of all the examiners excluding the benchmark examiner. The 95% confidence limits can be calculated from the group mean dmft/DMFT, the standard deviation and sample size. Examiners with a mean dmft/DMFT value outside the group mean dmft/DMFT range could be scoring at a different level. If, after completing these three steps, there is a need for further analysis, BASCD recommends the use of sensitivity and specificity measures as long as a benchmark examiner was used in all calibration exercises, and the kappa statistic for inter- and intra-examiner comparisons. Alternatively, the Dice concordance index is recommended if only one class is the object of interest, e.g. examiners' agreement on number of decayed teeth [4].

The ICDAS [2, 5] suggests the use of kappa coefficients to compare agreement between the senior examiner (benchmark examiner) and each examiner participating in a study and to assess intra-examiner reliability. It advises to include row-by-column (contingency) tables of the examiners' scores in all comparisons. In order to ensure the accuracy of kappa coefficients, ICDAS recommends that the marginal homogeneity of the distribution of codes for each examiner be tested. This can be tested by the McNemar's test for 2×2 contingency tables and by the Stuart–Maxwell test for general r×c tables [6–8].

An extension of the kappa statistic to situations where the categories are ordinal is called the weighted kappa. Since different mistakes can have a different impact, weights are assigned to penalise a certain misclassified case more severely than others. For instance, misclassifying category k into category k+2 is worse than misclassifying into category k+1. The weighted kappa may therefore be used to account for the degree of disagreement among observers. The Cicchetti–Allison weights and Fleiss–Cohen weights are two weighting techniques that can be used to compute the weighted kappa coefficients [9, 10].

When the two marginal distributions are not homogeneous (i.e. there is inequality between the row marginal proportions and the corresponding column proportions) then the kappa coefficient may not be a 'good' measure of agreement among the examiners. In such a case, ICDAS recommends statistical modelling for analysis of examiners' reliability, for example by using a log-linear model [11–13]. From the above, it is clear that the use of a kappa statistic to assess examiner reliability is encouraged; however, WHO and BASCD do not advise testing for marginal homogeneity of examiner codes. These two systems also do not consider the type of disagreement (which can have different clinical impact or weight). Further, BASCD is the only system that suggests the use of sensitivity and specificity as a reliability measure (in the presence of a benchmark examiner).

The aim of this contribution is to critically review different techniques for the measurement and analysis of examiner agreement. Different situations are illustrated using data from a recent caries experience survey in young children.

Measures of agreement

An agreement measure can be asymmetrical or symmetrical. Asymmetrical measures assess the scoring behaviour of raters against a gold standard. Examples are sensitivity and specificity measures. As seen above, it is usually not

possible to contrast the scoring behaviour in CE surveys to a gold standard. Therefore, in this context, sensitivity and specificity are most often defined vis-à-vis a benchmark scorer. On the other hand, symmetrical agreement measures compare the scoring behaviour among raters. Examples of symmetrical agreement measures are the percentage agreement, the Dice coefficient and the kappa statistic. A variety of agreement measures are described below. We also elaborate on computational aspects.

Sensitivity (sens) and specificity (spec) are statistical measures of the performance of a binary classification test. The sensitivity is equal to the percentage of actual positives which are correctly identified as such, e.g. the percentage of diseased people who are identified as diseased. The specificity is equal to the percentage of negatives which are correctly identified, e.g. the percentage of healthy people who are identified as healthy. In a dental CE survey, the sensitivity pertaining to a dental examiner is defined as the percentage of patients with true CE that is classified by the dental examiner as having CE. Specificity is the percentage of patients without CE that is classified by the dental examiner as not having CE. The results of the benchmark are used as a reference.

Below the computation of sens and spec is given. To this end, assume that a binary score represents caries experience (= 1) or not (= 0). This variable can be created at surface, tooth and subject level to indicate the presence or absence of caries experience at the respective levels.

Using the following 2×2 table

	Benchmark		
Examiner		1	0
	1	a	b
	0	c	d

and a+b+c+d=n, sens and spec are obtained as follows:

$$sensitivity = a/(a+c) \times 100\%, \quad (1)$$

$$specificity = d/(b+d) \times 100\%. \quad (2)$$

There are no strict cutoff levels as to what constitutes acceptable levels of sensitivity and specificity. Ideally, an examiner should have a high sensitivity and a high specificity. When sensitivity and specificity are less than or equal to 50%, the misclassification process is said to be poorer than by chance [4]. BASCD proposed a sensitivity of 75–80% for dft when compared to the benchmark dft and specificities of at least 90% [4]. On the other hand, Stamm et al. [14] recommended in their study a sensitivity of at least 75% and a specificity of at least 85%.

The percentage agreement, which is also known as the crude or raw agreement, is the simplest method for summarising an agreement for categorical variables. It reflects the percentage of the total number of units inspected where there is agreement between the examiner and the benchmark. Percentage agreement is computed as:

$$\text{percent agreement} = (a+d)/n \times 100\%. \quad (3)$$

From the literature, it remains vague what value should be considered as an acceptable level of agreement. The WHO proposed that acceptable agreement values should be in the range of 85–95% [3].

The Dice coefficient, also called the Jaccard index, is another measure of agreement, which is recommended to be used by Pine et al. [4], when only one class is the object of interest, e.g. agreement between the benchmark and the examiners on the number of teeth with CE. Dice is defined as the weighted average between the benchmark and the examiner agreeing on a tooth that is carious. Using the above 2×2 table, Dice is computed as

$$D = 2a/(2a+b+c) \quad (4)$$

A Dice coefficient of 1.0 indicates identical scoring by both the benchmark and the examiner, whereas a score of 0 means a total disagreement [15]; however, Dice is less used in dental literature since it does not access the overall agreement like kappa does.

The Cohen's kappa statistic (κ) is used to measure agreement of binary values. It is a relative measure that determines the excess of observed agreement to chance agreement. More specifically, kappa for binary outcomes is equal to [16]:

$$\kappa = (p_o - p_e)/(1-p_e), \quad (5)$$

with p_o and p_e the observed and expected agreement by chance, respectively. Using the above 2×2 table $p_o=(a+d)/n$ and $p_e=[(a+b)(a+c)+(c+d)(b+d)]/n^2$.

Theoretically, kappa ranges from a negative value to +1, but in practice it ranges from 0 to +1. A negative value is assumed if there is complete disagreement, kappa is zero if there is no more agreement that can be expected due to chance and 1 if there is perfect agreement. Landis and Koch [17] gave the following appreciation of observed kappa values: kappa less than 0.40 reflects poor agreement; when between 0.40 and 0.60, a fair to moderate agreement is present; when between 0.60 and 0.80, agreement is classified as good; and kappa values above 0.80 indicate close to perfect or at least excellent [17, 18]. Although this classification is often used in practice, Landis and Koch themselves note that this was but one arbitrary interpretation and hence it should not to be considered as universally appropriate.

The kappa statistic is influenced by disease prevalence. As a result, kappas are seldom comparable across studies, procedures or populations. To overcome this problem, Gwet [19] suggested the use of another agreement measure called the agreement coefficient (AC1). AC1 assumes that the probability of chance agreement is proportional to the portion of rating that may lead to an agreement by chance, thereby reducing the overall agreement by chance to the right magnitude; however, AC1 is rarely used in the dental literature. AC1 is computed as

$$\mathrm{AC1} = (p_o - p_{e*})/(1 - p_{e*}), \quad (6)$$

with p_o and p_{e*} the observed and expected agreement by chance, respectively. Using the above 2×2 table $p_o=(a+d)/n$ and $p_{e*}=2\pi\ (1-\pi)$, where $\pi=(2a+b+c)/2n$.

Another proposed statistic to overcome kappa's deficiencies is the tetrachoric or polychoric correlation [20]. This measure relies on the assumption that there exist continuous latent variables underlying the contingency table. The chance correction in the kappa statistic depends on how a rater makes decisions on scoring a response as positive or negative (the threshold to decide as a positive or negative). Two raters may agree perfectly on the underlying trait, but due to the difference in thresholds, their category scoring may be different. In this case kappa will be low, implying that the raters' responses are similar to chance agreement, when in fact there is perfect agreement on the underlying continuous trait. The kappa statistic therefore mixes two sources of disagreements between two raters [21], (1) the disagreements due to different thresholds in categorising disease status into positive or negative and (2) the disagreement as a result of examiners ranking the categories in a different manner. In the first case, the observed prevalence obtained by the examiners will differ. This can be dealt with by reporting the prevalence index, calculated as $a-d/n$ from the above 2×2 table [22]. Further, Byrt et al. [22] suggested the use of a prevalence-adjusted and bias-adjusted kappa, which is calculated as:

$$\mathrm{PABAK} = [(a + d/n) - 0.5]/(1 - 0.5).$$

Another way is to compute the tetrachoric correlation. The tetrachoric (polychoric) correlation only measures the agreement when the two cutpoints or thresholds had been the same for both examiners. Hence, the tetrachoric correlation is not depending on marginal inhomogeneity and it measures part of the agreement. For example, in Table 1, two raters using the same threshold categorised the outcome into positive and negative. The calculated tetrachoric correlation for both cases was high (1.00) while different kappa values were obtained (0.89, 0.67) for cases A and B, respectively. In this case, the different kappa values were observed due to differences in the marginal totals of the two cases.

Table 1 Comparison of kappa and tetrachoric correlation for a binary outcome from a bivariate normally distributed data at a high correlation (0.9) using the same threshold

		Benchmark			
Examiner		1	0	1	0
	1	6	0	12	0
	0	1	13	3	5
		Case A		Case B	

Validation sample

In order to compare the different systems for assessing examiner reliability, data from the Smile for Life project were used. This oral health promotion intervention study in very young children (and their parents) was launched in 2003 in Flanders (Belgium). Before starting the intervention, baseline data were collected in 3- and 5-year-old children. The results of this survey have been described in detail elsewhere [23]. Examiners participating in the oral health screening were trained according to the criteria published by the BASCD [1]. The recording of CE on individual tooth surfaces was done at the d_1-level (initial lesion), but allowing reporting of results at the d_3-level (cavitation into dentine) [1]. In this contribution, only results at the d_3-level will be presented.

A calibration exercise was organised in age-matched (matching with respect to main study) children (5-year olds, total of 26 children) in order to assess the agreement between the scores obtained by the individual examiners (seven examiners) and the scores obtained by the benchmark examiner (one of the authors, i.e. DD). Written consent was obtained from the parents of all children. All examinations took place on the same day. For the validation exercise, information was available from 26 children; however, for some children not all eight dental examiners (benchmark and seven examiners) scored CE. Most examiners (five) examined 22 children; one examiner, 24 and another examiner, 26 children. All available information was included in the analyses. The prevalence of CE in the validation study was 0.35 according to the scores from the benchmark examiner.

Influence of unit of analysis

In a first step, agreement between examiners from the Smile for Life project was assessed using the above-described

measures. Agreement measures were computed on CE at surface, tooth and subject level. Next, the influence of the unit of analysis on outcome of measurement was explored. Table 2 presents the agreement in scoring CE for the different examiners, calculated at subject, tooth and surface level, using the different methods described above. It is important to note that scoring CE is done at surface level and that the scores on tooth and subject level are derived herefrom.

From the results, it can be seen that at subject level the percentage agreement was above 81% for all examiners. The kappa value ranged from 0.56 to 0.72 indicating moderate to good agreement in the classification system of Landis and Koch. The sensitivity ranged from 50% to 75% while specificity ranged from 92% to 100%. The Dice coefficient and the AC1 values ranged from 67% to 80%. At tooth level the percentage agreement was higher than the values obtained at subject level and reached more than 91%; however, the kappa values dropped to between 0.36 and 0.44, indicating poor to fair agreement. The sensitivity at tooth level was also less than at subject level and ranged from 29% to 40%. For all examiners a specificity of 97% or higher was obtained. The Dice coefficient was lower at the tooth level compared to subject level, and it ranged from 40% to 48% while the AC1 increased at the tooth level and ranged from 91% to 92%. At surface level, the percentage agreement and AC1 were higher than at mouth and tooth level and reached values above 95% for all examiners. The kappa values dropped again with values between 0.38 and 0.42, indicating poor to fair agreement. The Dice coefficient also dropped to a value between 40% and 44% for all examiners. Sensitivity ranged from 29% to 34%. For most of the examiners, a specificity of 98% or higher was obtained.

At first glance, one would expect that reliability measures become less favourable as we go down in the hierarchy from subject to tooth and surface level. This is because of the logical argument that if a rater scored correct at all surfaces of a tooth, the attributed score will also be correct at tooth and subject level. On the other hand, when an examiner failed at surface level, it is still possible that the scoring is correct at tooth level and subject level, for example when the examiner scored CE on the wrong surface of a tooth, then the tooth is still correctly scored as CE; however, it is clear from the agreement measures obtained in the Smile for Life study that this is not necessarily the case since the percentage agreement was considerably higher at surface level than at subject level. The explanation for this finding can be found in Fig. 2, case 1. There are 14 teeth correctly classified by the examiner, which gives 70% agreement, but 74 out of 88 surfaces are correctly classified resulting in 84% agreement. Clearly, the higher agreement at surface level is obtained because the teeth without CE count as four or five correct agreements. Hence, the percentage agreement at surface level is increased because teeth without CE are given much higher impact on surface level.

From the above, it is clear that the unit of analysis may influence the obtained results considerably [24, 25]. Because of the impact of unit of analysis on the outcome of reliability measures, the level at which reliability assessments were performed should be clearly stated in reports. In addition, it is suggested that authors report and analyse the results of the validation at the same level as chosen for the main study. Consequently, we wish to point out that a comparison between agreement measures from different studies cannot be performed if the level at which they were calculated is not specified.

Impact of disease prevalence in the validation sample on measures of agreement

Often, sensitivity and specificity are considered to represent intrinsic scoring behaviour and thus to be independent of prevalence; however, this may not be the case in practice. For instance, in an ongoing analysis and based on calibration exercises in another study [26], we observed that the specificity depended on the number of surfaces in the mouth that showed CE. This phenomenon is referred to as an informative cluster size [27]. Hence, sensitivity and specificity are directly independent of prevalence (prevalence

Table 2 Ranges of percentage agreement, kappa (95% CI), sensitivity and specificity, Dice coefficient, AC1 and tetrachoric correlation scores obtained by the different examiners (seven in total) involved in the Smile for Life project (data analysed at subject, tooth and surface level)

Level of analysis	Numbers examined	Percentage agreement (%)	Kappa value	Sensitivity (%)	Specificity (%)	Dice coefficient (%)	AC1	Tetrachoric correlation
Subject	26	81–89	0.56–0.72 [0.21; 0.91]–[0.44; 1.00]	50–75	92–100	67–80	0.68–0.80	0.83–1.00
Tooth	476	91–93	0.36–0.44 [0.19; 0.53]–[0.28; 0.60]	29–40	97–99	40–48	0.91–0.92	0.74–0.83
Surface	2104	95–97	0.38–0.42 [0.26; 0.49]–[0.31; 0.53]	29–34	98–99	40–44	0.95–0.96	0.77–0.87

does not occur in the formula) if the population characteristics do not change with prevalence.

On the other hand, percent agreement, the AC1 measure and kappa directly depend on prevalence. That kappa depends on the prevalence is seen above. The dependence of percentage agreement and the AC1 measure follows from percent agreement $= \{\text{spec} \times (1-p) + \text{sens} \times p\} \times 100\%$ and that the AC1 measure is a function of percent agreement.

It is also illustrative to look at the relation between kappa, disease prevalence, sensitivity and specificity, see Fig. 1. The graph shows that kappa increases with increasing disease prevalence to reach a maximum and then decreases. This maximum differs for different sensitivity values. When sensitivity is 50%, a maximum kappa value of 0.43 is obtained at 28% disease prevalence. Increasing the sensitivity to 70% or 90% gives maximum kappa values of 0.63 and 0.80 at 38% and 50% disease prevalence, respectively. The graph also shows that the same kappa (here 0.62) can be obtained from different sensitivity and prevalence combinations, i.e. (1) sensitivity=90%, prevalence=12% and (2) sensitivity=70%, prevalence=36%.

A consequence of the above results is that two studies with a similar sensitivity and specificity for the raters will have different kappa values if the disease prevalence in the two studies is different [22, 28, 29]. Further, if in the validation study the disease prevalence is much different from that in the main study, then the reported kappa values from the validation study are not representative for the main study. Furthermore, a low kappa is obtained from a study with a high or low disease prevalence even when the sens and spec are high. This is due to the high probability of examiners agreeing purely by chance at these two extremes [28, 30]. Kappa used in such a situation could give wrongly the impression of poor examiner agreement. Summarised, it is important to report the prevalence of disease in the validation sample when kappa is reported.

Some practical examples

In order to illustrate above mentioned mechanisms, some examples are presented.

Figure 2 presents four simulated cases and illustrates that similar CE scores can be obtained by benchmark and examiner at subject level despite differences at tooth and/or surface level. In each case, agreement measures are also shown.

In the first case, the benchmark and the examiner scored (several) different teeth and surfaces for CE, but they both obtained a similar summary score at subject level (dmft as well as dmfs). The kappa values are low (0.20, 0.13 respectively), but the percentage agreement remains high (70%, 84%). In the second case, the benchmark and the examiner obtained a similar score at tooth level with the examiner overscoring at surface level (three surfaces). There is perfect agreement at tooth level. At surface level, both the percentage agreement and kappa values are high (94%, 0.85). In the third case, the benchmark and the examiner have similar scores at tooth level with the examiner missing caries experience on one single surface and this in the case of low disease severity. Here, the percentage agreement remains high while kappa (at surface level) and also sensitivity drops considerably. The same mistake in a patient with higher disease severity (case four) would yield a kappa value of 0.93 (at surface level) which is considerably higher. This illustrates the impact of disease prevalence on agreement measures.

From the above, it is clear that a high value for overall percentage agreement can be obtained even though a

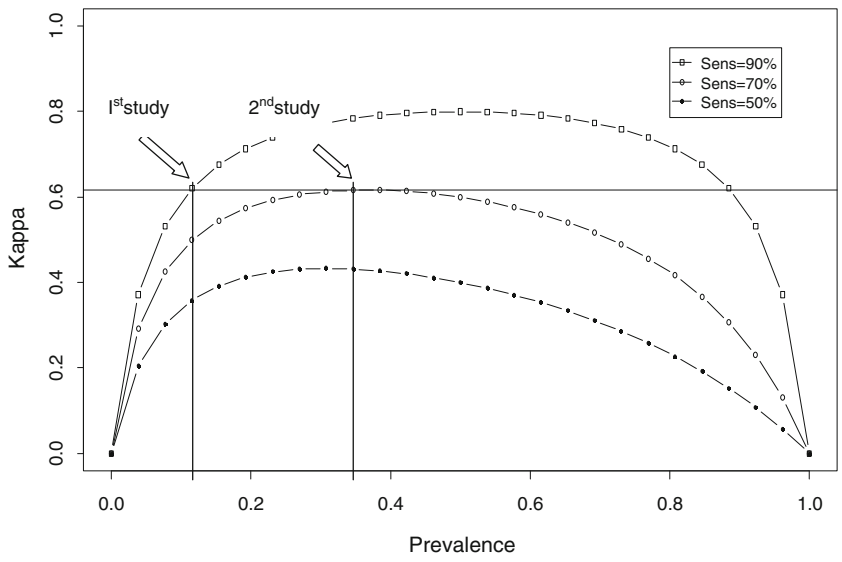


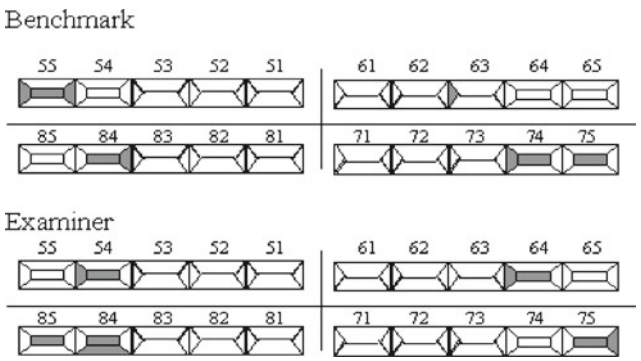
Fig. 1 Kappa scores obtained at different levels of disease prevalence with fixed specificity (90%) and different sensitivity (50%, 70%, 90%) levels

considerable amount of disease cases were misclassified (30%, e.g. in case 1 in Fig. 2). This high value of percentage agreement results from the large number of non-diseased cases on which the examiners agreed. Thus, at low disease prevalence a high value of percentage agreement may conceal significant disagreement between the examiners. It is, therefore, important to combine the percentage agreement with other agreement measures (e.g. sensitivity and specificity) when the proportion of non-diseased cases in the validation sample is high (which is often the case in contemporary caries experience surveys).

Statistical uncertainty

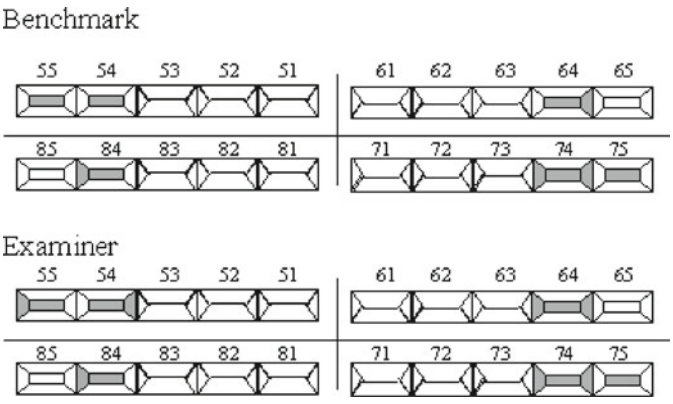
Another issue to consider is statistical uncertainty. In reports on caries epidemiological surveys agreement measures are rarely reported with an indication of their

Case 1: Similar scores at subject level for the benchmark and the examiner but with several different teeth and surfaces scored.



Summary scores	Benchmark	Examiner
	dmft = 5	dmft = 5
	dmfs = 9	dmfs = 9
Agreement measures	**Tooth level**	**Surface Level**
% agreement	70	84
Kappa	0.20	0.13
Sensitivity	40	22
Specificity	80	91

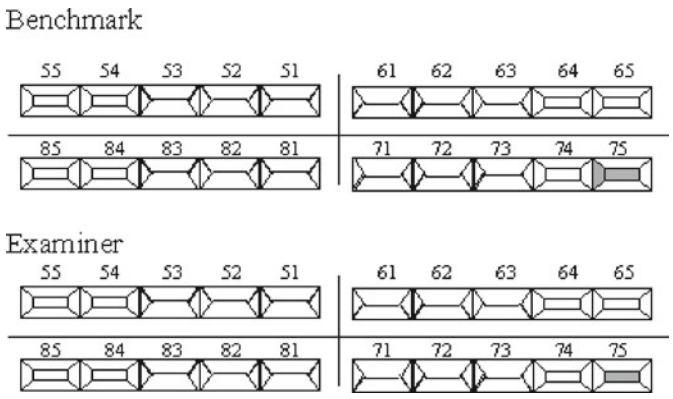
Case 2: Similar scores obtained at subject and at tooth level by the benchmark and the examiner, but the examiner is over-scoring at surface level.



Summary scores	Benchmark	Examiner
	dmft = 6	dmft = 6
	dmfs = 10	dmfs = 13
Agreement measures	**Tooth level**	**Surface Level**
% agreement	100	94
Kappa	1.00	0.85
Sensitivity	100	100
Specificity	100	96

Fig. 2 Simulated situations showing similar dmft and/or dmfs scores for benchmark and examiner at subject level in spite of differences at tooth and/or surface level and the impact on agreement measures. *Case 1:* Similar scores at subject level for the benchmark and the examiner but with several different teeth and surfaces scored. *Case 2:* Similar scores obtained at subject and at tooth level by the benchmark and the examiner, but the examiner is overscoring at surface level. *Case 3:* Similar scores obtained by benchmark and examiner at subject and at tooth level with the examiner missing caries experience on one single surface (low disease severity). *Case 4:* Similar scores obtained by benchmark and examiner at subject and at tooth level with the examiner missing caries experience on one single surface (high disease severity)

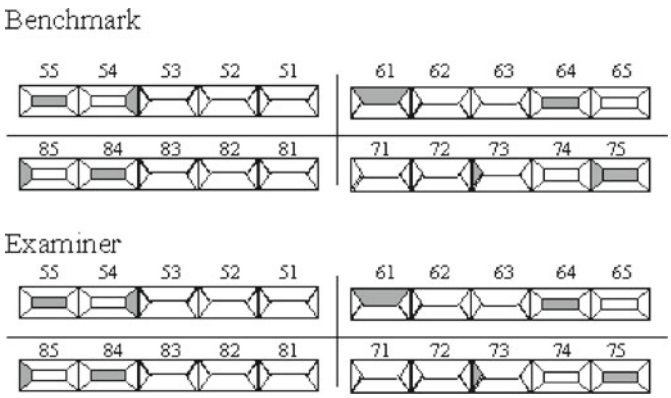
Case 3: Similar scores obtained by benchmark and examiner at subject and at tooth level with the examiner missing caries experience on one single surface (low disease severity).



Summary scores	Benchmark	Examiner
	dmft = 1	dmft = 1
	dmfs = 2	dmfs = 1

Agreement measures	Tooth level	Surface Level
% agreement	100	98
Kappa	1.00	0.66
Sensitivity	100	50
Specificity	100	100

Case 4: Similar scores obtained by benchmark and examiner at subject and at tooth level with the examiner missing caries experience on one single surface (high disease severity).



Summary scores	Benchmark	Examiner
	dmft = 8	dmft = 8
	dmfs = 9	dmfs = 8

Agreement measures	Tooth level	Surface Level
% agreement	100	98
Kappa	1.00	0.93
Sensitivity	100	89
Specificity	100	100

Fig. 2 (continued)

uncertainty, i.e. the confidence interval (CI) [31, 32]. The uncertainty comes from the fact that most often the reported agreement measure is based on a random sample from a population and hence is only an estimate of the true agreement measure. Clearly, there is more uncertainty if the agreement measure is based on a small sample and this is reflected by the 95% CI. Although technically different, the interpretation of the 95% CI for a true value in practice is that we are certain with 0.95 probability that the true value is included in that interval. Thus, the adjective 95% indicates a degree of certainty that the CI includes the true value (of agreement here). Generally, the confidence limits are reported as two-sided (lower and upper limit) but in practice the lower limit is of interest.

The effect of clustering on agreement measures

In oral health research, data often have a multilevel or hierarchical structure [33, 34]. This results in a grouping/dependency effect at subject and tooth level. Teeth from the same subject and surfaces from the same tooth are assumed to share similar characteristics which may have an important effect on the CE measurement [26]. As such, an

examiner is more likely to have a consistent scoring behaviour in a particular subject than across subjects.

Traditional measures of agreement, e.g. kappa, specificity, sensitivity and the percentage agreement are calculated under the assumption of independence, i.e. observations are assumed to be independent. Although the estimates of these measures are relatively insensitive to the assumption of independence, the clustered structure of data may have a strong impact on the calculation of the 95% CI [35].

Multilevel modelling is a statistical approach that takes into account the hierarchy in data. An example of a multilevel approach to investigate factors that influence the development of dental caries was presented by Burnside et al. [36]. This approach can be invoked to analyse sensitivity and specificity in the presence of clustering. Covariates can be included to explain sensitivity and specificity on characteristics of the mouth, tooth, surface, examiner, etc. An example of such an approach is given in [37]. A kappa statistic for dependent data has been suggested by Williamson et al. [38], who suggested a generalised estimating equations (GEE) [39] approach for assessing the dependent agreement measures when the response is categorical. In this approach, two sets of equations are incorporated in the GEE; the first set is to model the marginal distributions of the categorical responses and the second set is to model all pairwise correlations of the ratings using the kappa coefficient. In both sets of estimating equations, covariates can be added.

Advanced statistical modelling offers additional possibilities. Based on the misclassification information obtained in the validation data set, correction of data of the main study can be implemented. In this way, differences in outcome measures that could be related to differences in scoring behaviour of the examiners involved can be accounted for. For an example of this application, we refer to Lesaffre et al. [40].

Finally, we wish to point out that often (perhaps too often) the reliability studies are too small yielding too wide confidence intervals for the agreement measures especially because of the often clustered nature of the data. It is, therefore, advisable to perform a priori a sample size calculation to achieve a particular degree of certainty for the agreement measures, see [41, 42]. This is also important if correction for misclassification is envisaged, see next section.

Gold standard and benchmark examiners

A ‘gold standard’ is regarded as reflecting the absolute truth, i.e. an instrument or technique yielding information about the condition of interest that is regarded as infallible. Therefore, in CE screening a histological section of the tooth (surface) considered might be regarded as providing a gold standard assessment. Unfortunately, in practice such a high degree of certainty can never be achieved. In screening, assessment is often based on visual–tactile evaluations and the scoring of a reference examiner, called benchmark examiner, is used as the standard. This benchmark examiner is an experienced assessor with consistent scoring behaviour in the past. In CE surveys, at best a benchmark scorer is included allowing to compute sensitivity and specificity but often such a benchmark is not available and then only symmetric agreement measures such as kappa can be determined. It should be clear that sensitivity and specificity obtained from a benchmark depends on his/her correct scoring and intra-observer variability. Indeed, when benchmark scores with error a distorted picture of the reliability will be obtained and the estimates used for correction in regression models will be biased [43]. Brenner [44] explored the effects of using a reference which is less than perfect, called the alloyed gold standard and shows that using estimates from such a reference results in an overcorrection in the main model. It is, therefore, important that the performance of benchmark examiners is subjected to regular quality control.

Conclusions

The validity of results presented in a CE survey depends, amongst other items, on the reliability of the measurements. It is, therefore, important that information on reliability measurement is included in survey reports. In spite of efforts regarding the standardisation of screening methodology, many uncertainties remain with respect to the measurement, analysis and interpretation of examiner reliability.

In this report, we have shown that agreement measures frequently used in CE surveys (percentage agreement, kappa, sensitivity and specificity) are influenced (in differing ways and extents) by the unit of analysis (mouth, tooth or surface level) and the disease level in the validation sample. It is, therefore, advised to include information on the unit of analysis applied (preferably identical in validation and main study) and disease level of the validation sample in the survey reports.

In this study, we looked at caries experience as a binary outcome. Caries experience can also be scored using dmft/DMFT or dmfs/DMFS. In that case, other methods such as the intraclass correlation, the concordance correlation coefficient or the Bland and Altman method can be used. Note, however, that such study operates on a subject level (since dmft is a summation over teeth and surfaces) and much valuable information is lost on the scoring behaviour.

When reporting reliability measures, the confidence interval should be presented, as an indication of reliability

of the estimate. In addition, the dependency of observations should not be disregarded as this has an impact on the width of the CI. Advanced statistical modelling offers new perspectives by exploring the information contained in a validation dataset to a larger extent. It should be underlined that the quality of the scoring by the benchmark examiner is crucial when reliability measures are calculated and should therefore be subjected to regular evaluations.Finally, from the above it is clear that there is a need for the development of guidelines for the measurement, analysis, interpretation and reporting of examiner reliability in CE surveys.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Calculation of kappa from sensitivity, specificity and disease prevalence.

Let Y^*=examiner score, Y=benchmark score

$$P_{jk} = P(Y^* = j,\ Y = k) = P(Y^* = j/Y = k) \times P(Y = k)$$

P_{jk} = probability of examiner score j and benchmark score k

j and k take values of 0 and 1 e.g. if j=1 and k=1we obtain

$$P_{11} = P(Y^* = 1,\ Y = 1) = P(Y^* = 1/Y = 1) \times P(Y = 1)$$

$$P(Y^* = 1/Y = 1) = \text{sensitivity} = \text{sens}$$
$$P(Y = 1) = \text{prevalence} = p$$

if j=0 and k=0 we obtain $P_{00} = P(Y^* = 0,\ Y = 0) = P(Y^* = 0/Y = 0) \times P(Y = 0)$

$$P(Y^* = 0/Y = 0) = \text{specificity} = \text{spec}$$
$$P_{11} = \text{sens} \times p$$
$$P_{00} = \text{spec} \times (1 - p)$$

Similarly P_{01} and P_{10} are obtained

$$p_o = P_{00} + P_{11}$$
$$p_e = (P_{00} + P_{01}) \times (P_{00} + P_{10}) + (P_{10} + P_{11}) \times (P_{01} + P_{11})$$
$$p_o = \{\text{spec} \times (1 - p) + \text{sens} \times p\}$$
$$p_e = \{[\text{spec} \times (1 - p) + (1 - \text{sens}) \times p] \\ [\text{spec} \times (1 - p) + (1 - \text{spec}) \times (1 - p)] \\ + [(1 - \text{spec}) \times (1 - p) + \text{sens} \times p] \\ [(1 - \text{sens}) \times p + \text{sens} \times p]\}$$
$$\kappa = (p_o - p_e)/(1 - p_e)$$

References

1. Pitts NB, Evans DJ, Pine CM (1997) British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97. Community Dent Health 14(Suppl 1):6–9
2. Ismail AI, Sohn W, Tellez M, Amaya A, Sen A, Hasson H, Pitts NB (2007) The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries. Community Dent Oral Epidemiol 35:170–178
3. World Health Organization (1997) Oral health surveys. Basic methods. World Health Organization, Geneva
4. Pine CM, Pitts NB, Nugent ZJ (1997) 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 2. Community Dent Health 14(Suppl 1):18–29
5. International Caries Detection and Assessment System Coordinating Committee (2005) Criteria manual - International Caries Detection and Assessment System (ICDAS II)
6. Maxwell AE (1970) Comparing the classification of subjects by two independent judges. Br J Psychiatry 116:651–655
7. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153–157
8. Stuart AA (1955) A test for homogeneity of the marginal distributions in a two-way classification. Biometrika 42:412–416
9. Cicchetti DV, Allison T (1971) A new procedure for assessing reliability of scoring EEG sleep recordings. Am J EEG Technol 11:101–109
10. Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
11. Kingman A (1986) A procedure for evaluating the reliability of a gingivitis index. J Clin Periodontol 13:385–391
12. Uebersax JS (1993) Statistical modeling of expert ratings on medical treatment appropriateness. J Am Stat Assoc 88:421–427
13. Uebersax JS (1987) Diversity of decision-making models and the measurement of interrater agreement. Psychol Bull 101:140–146
14. Stamm JW, Stewart PW, Bohannan HM, Disney JA, Graves RC, Abernathy JR (1991) Risk assessment for oral diseases. Adv Dent Res 5:4–17
15. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302
16. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
17. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
18. Tooth LR, Ottenbacher KJ (2004) The kappa statistic in rehabilitation research: an examination. Arch Phys Med Rehabil 85:1371–1376
19. Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 61:29–48
20. Uebersax JS (2005) Statistical methods for rater agreement: the tetrachoric and polychoric correlation coefficient. http://www.john-uebersax.com/stat/tetra.htm. Accessed 10 May 2010
21. Hutchinson TP (1993) Focus on psychometrics. Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. Res Nurs Health 16:313–316
22. Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. J Clin Epidemiol 46:423–429
23. Declerck D, Leroy R, Martens L, Lesaffre E, Garcia-Zattera MJ, Vanden BS, Debyser M, Hoppenbrouwers K (2008) Factors associated with prevalence and severity of caries experience in preschool children. Community Dent Oral Epidemiol 36:168–178
24. Assaf AV, de Castro MM, Zanin L, Tengan C, Pereira AC (2006) Effect of different diagnostic thresholds on dental caries calibration—a 12-month evaluation. Community Dent Oral Epidemiol 34:213–219
25. Assaf AV, Tagliaferro EP, Meneghim MC, Tengan C, Pereira AC, Ambrosano GM, Mialhe FL (2007) A new approach for

interexaminer reliability data analysis on dental caries calibration. J Appl Oral Sci 15:480–485
26. Vanobbergen J, Lesaffre E, Garcia-Zattera MJ, Jara A, Martens L, Declerck D (2007) Caries patterns in primary dentition in 3-, 5- and 7-year-old children: spatial correlation and preventive consequences. Caries Res 41:16–25
27. Williamson JM, Datta S, Satten GA (2003) Marginal analyses of clustered data when cluster size is informative. Biometrics 59:36–42
28. Hoehler FK (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol 53:499–503
29. Spitznagel EL, Helzer JE (1985) A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 42:725–728
30. Maclure M, Willett WC (1987) Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126:161–169
31. Braga MM, Oliveira LB, Bonini GA, Bonecker M, Mendes FM (2009) Feasibility of the International Caries Detection and Assessment System (ICDAS-II) in epidemiological surveys and comparability with standard World Health Organization criteria. Caries Res 43:245–249
32. Fyffe HE, Deery C, Nugent ZJ, Nuttall NM, Pitts NB (2000) Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee Selectable Threshold Method for caries diagnosis (DSTM). Community Dent Oral Epidemiol 28:42–51
33. Gilthorpe MS, Griffiths GS, Maddick IH, Zamzuri AT (2000) The application of multilevel modelling to periodontal research data. Community Dent Health 17:227–235
34. Tu YK, Gilthorpe MS, Griffiths GS, Maddick IH, Eaton KA, Johnson NW (2004) The application of multilevel modeling in the analysis of longitudinal periodontal data-part I: absolute levels of disease. J Periodontol 75:127–136
35. Williams FM, Nan G (2006) Estimation of sensitivity and specificity of clustered binary data. Statistics and data analysis, SUGI 31 Proceedings, SAS Proceedings
36. Burnside G, Pine CM, Williamson PR (2007) The application of multilevel modelling to dental caries data. Stat Med 26:4139–4149
37. Mutsvari T, Lesaffre E, Garcia-Zattera MJ, Diya L, Declerck D (2010) Factors that influence data quality in caries experience detection: a multilevel modeling approach. Caries Res 44:438–444
38. Williamson JM, Lipsitz SR, Manatunga AK (2000) Modeling kappa for measuring dependent categorical agreement data. Biostatistics 1:191–202
39. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
40. Lesaffre E, Mwalili SM, Declerck D (2004) Analysis of caries experience taking inter-observer bias and variability into account. J Dent Res 83:951–955
41. Barnhart HX, Williamson JM (2002) Weighted least-squares approach for comparing correlated kappa. Biometrics 58:1012–1019
42. Lin HM, Williamson JM, Lipsitz SR (2003) Calculating power for the comparison of dependent κ-coefficients. J R Stat Soc C Appl Stat 52:391–404
43. Wacholder S, Armstrong B, Hartge P (1993) Validation studies using an alloyed gold standard. Am J Epidemiol 137:1251–1258
44. Brenner H (1996) Correcting for exposure misclassification using an alloyed gold standard. Epidemiology 7:406–410