Effect of Different Diagnostic Thresholds on Dental Caries Calibration

Andréa Videira Assaf, DDS, MSc, PhD; Marcelo de Castro Meneghim, DDS, MSc, PhD; Luciane Zanin, DDS, MSc; Karine Laura Cortelazzi, DDS, MSc, DrPH; Antonio Carlos Pereira, DDS, MSc, PhD; Gláucia Maria Bovi Ambrosano, DDS, MSc, PhD

Abstract

Objectives: *a)to analyze the intra- and inter-examiner reproducibility (reliability) of a calibration trial, at different diagnostic thresholds of dental caries; b) to verify the accuracy (benchmark validity) though sensitivity (S), specificity (SP), positive (PPV) and negative predictive (NPV) values.* ***Participants:*** *A group of dental examiners (n=11), who had previous experience in epidemiological surveys and six to seven-year-old children. Children were selected according to the dmft and dental caries activity.* ***Methods:*** *Theoretical and clinical training and calibration exercises were arranged for a total of 28 hours. WHO criteria including the active initial lesions (IL) were used. Main outcome measures: WHO and WHO+IL diagnostic thresholds according to tooth and dental surface.* ***Results:*** *Excellent mean results of intra and inter-examiner Kappa values were found for both diagnostic thresholds, according to tooth and surface, during the calibration phase. The most relevant errors were related to IL diagnosis and to the first permanent molars. When assessed against a benchmark examiner, moderate to high validity values were observed (0.71-1.00), with some loss mainly for sensitivity and positive predictive value, when including IL.* ***Conclusion:*** *It was possible and feasible to use the proposed methodology of this study in epidemiological surveys, even with the inclusion of IL. However, further examiner calibration studies are still needed in order to improve and establish a methodology of calibration with this new diagnostic threshold.*

Key Words: epidemiology, dental caries, diagnosis

Introduction

Dentistry continually demands strict control of research development principles in order to obtain more qualified and consistent results. Calibration, by determining the examiners' reproducibility before and during a dental caries epidemiological survey, is an important factor for better understanding of the examiners in relation to the research criteria used. It also sets the standard to which examiners are expected to work and it provides information to establish whether the survey results are reliable. Lack of examiner agreement could indicate inaccuracy and lead to problems of data interpretation and lack of comparability with other datasets (1, 2).

The process of examiner calibration for epidemiological diagnosis of dental caries can be divided into theoretical and clinical training and calibration exercises. Calibration exercises and reproducibility and validity analyses enable formal results of the examiners' understanding to be obtained, thus allowing them to participate in the survey (1, 2).

The worldwide decrease in the prevalence of dental caries in children and adolescents, reported over the past 20 years, has generated clinical consequences such as the presence of a larger number of non-cavitated carious lesions or initial lesions (IL), with a reduction in cavitated carious lesions and the predominance of activity on occlusal surfaces (3,4,5,6).

Most epidemiological caries studies employ the WHO (World Heath Organization) diagnostic criteria in which a tooth or a tooth surface is recorded as being decayed when cavitation is obvious, but excludes all initial lesions. This threshold has been termed the D_3 diagnostic threshold (1, 7). Therefore, initial lesions (IL) in enamel, or even non-cavitated lesions in dentin (hidden lesion) are ignored. This may no longer be sufficient to reflect changes in the incidence of caries in populations showing a slow overall rate of caries progression. It also means that they fail to benefit from some form of preventive management strategy (3,4,8,9).

Among the problems of recording IL appropriately are the difficulty of diagnosing IL under epidemiological conditions and the great difficulty inherent to calibrating examiners at more sensitive thresholds (10,11). However, scientific literature has shown the real possibility of doing so. Nyvad *et al.* (1999), Fyfee *et al.* (2000) and Warren *et al.* (2002) have shown acceptable reliability (reproducibility) and validity (accuracy) results at more sensitive diagnostic thresholds (6,8,9).

Therefore, in the face of the changes in the dental caries pattern, the difficulty of diagnosing the disease appropriately, particularly under epidemiological conditions, and the problems with calibrating examiners, especially when the initial stages of disease are included in the examinations, the aims of this study were: a) to analyze the intra and inter-examiner reproducibility (reliability) of calibration trial, at different diagnostic thresholds of dental caries; b) to verify the accuracy (benchmark validity)

Send correspondence and reprint requests to: Dr. Marcelo de Castro Meneghim or Andréa Videira Assaf, Departamento de Odontologia Social, Avenida Limeira 901, Piracicaba, SP, Brazil CEP: 13414-903. E-mail: meneghim@fop.unicamp.br. Drs. Assaf, de Castro Meneghim, Zanin, Cortelazzi, Pereira and Ambrosano are affiliated with the Department of Community Dentistry - FOP - UNICAMP- Piracicaba, SP, Brazil. Manuscript received 7-08-05; returned to authors for revisions 8/30/05; final version accepted for publication 9/24/05.

though sensitivity (S), specificity (SP), positive predictive (PPV) and negative predictive (NPV) values, according to the diagnostic thresholds, respectively.

Material and Methods

Ethical approval was obtained form the Ethical Committee in Research at the Piracicaba Dentistry School/UNICAMP (State University of Campinas), Protocol No. 068/2002, in agreement with Resolution 196/96 of the National Committee of Health/Health Department (BZ). The schools granted permission for the study and informed consent was obtained from the parents.

Study design: *Sample and examiner selection*- A heterogeneous group of 11 dentists (4 salaried public health dentists and 7 post-graduate students), all experienced examiners in dental caries epidemiological surveys according to the WHO codes and criteria, participated in the study (1).

Six- to seven-year-old children from public schools in Piracicaba-SP-Brazil, which has had 0.7 ppmF in the water supply since 1997, were previously selected in an outdoor setting (schoolyard), by a professional who did not participate in the experimental phase. The examiner used a dental mirror, a ball-ended CPI (Community Periodontal Index) probe with a diameter of 0.5mm, and previous dental brushing and drying for the exams. Therefore, 10 to 12 different children with mixed dentition were selected for each training session and for the calibration period. All children were born in the town or had lived there since the age of two and had consent for participation included in this study. Children were selected beforehand, in order to represent a variety of findings. Priority was given to children with caries experience, i. e. with cavitated and initial lesions, although some caries-free children were also included, so that the examiners did not go into the exercise with the preconception that all children had caries (2,12). Children that had local or general problems, such as the use of fixed orthodontic devices, severe fluorosis and hypoplasia, or serious systemic diseases were excluded from the sample.

Diagnostic criteria and codes - The criteria and codes were those based on the WHO recommendations (1). For the IL, active caries with intact surfaces were recorded; an adaptation of the criteria according to Nyvad et al. (1999) and Fyffe et al. (2000) (8,9). Thus, an IL (initial lesion) was defined as active caries which, through visual assessment by a calibrated examiner, indicates intact surface, no clinically detectable loss of dental tissue, a rough, whitish/yellowish colored area of increased opacity, with loss of luster and presumed to be carious (when the probe is used its tip should be moved gently across the surface). Smooth surface: caries lesion typically located close to gingival margin. Fissure/pit: intact fissure morphology: lesion extending along the walls of the fissure. In this study, localized surface defects (active microcavities) restricted to enamel only were included (use of the same code – W) in the IL group. Active white spot lesions and microcavities contiguous to sealants, restorations and cavitations were also recorded (Table 1).

Examiner training and calibration - A benchmark dental examiner ("Gold Standard") conducted the complete process of training and calibrating examiners. The benchmark examiner (dentist who routinely uses the WHO

TABLE 1
Summary of the criteria and codes, according to WHO and WHO+IL diagnostic threshold for caries, restorations, sealants and other dental conditions

WHO Codes		Criteria	WHO+IL Codes		Criteria
Prim	Perm		Prim	Perm	
A	0	Sound	A	0	Sound, excluding W (white spot)
			W	WP	W (active white spot/ surface discontinuity in enamel only)
B	1	Decayed	B	1	Decayed without W (chronic lesion)
			BW	1W	Decayed with W (active lesion)
C	2	Filled, with decay	C	2	Filled, with decay (chronic lesion)
			CW	2W	Filled, with W + decay (active lesion)
D	3	Filled, no decay	D	3	Filled, no decay
			DW	3W	Filled with W
E	4	Missing, as a result of caries	4	4	Missing, as a result of caries
-	5	Missing, any other reason	5	5	Missing, any other reason
F	6	Fissure sealant	F	6	Fissure sealant
			FW	6W	Fissure sealant with W
G	7	Bridge abutment, special crown crown or venner/implant	7	7	Bridge abutment, special crown or venner/implant
-	8	Unerupted tooth		8	Unerupted tooth
T	T	Trauma (fracture)	T	T	Trauma (fracture)
-	9	Not recorded	-	9	Not recorded

Note: code W- presence of white spot or surface discontinuity in enamel in dental surfaces (W, WP), as well as in sealants (FW, 6W), filled (DW, 3W) and other conditions.

criteria for exams) had been previously trained and calibrated in the diagnosis of IL and had routinely examined using these criteria in another study (11).

The training and calibration sequences were as follows: training sessions performed in a total of 5 periods (1 theoretical, 4 clinical training) and 2 calibration sessions conducted after the clinical training.

Theoretical discussions were first held and clinical photographic slides shown to provide visual examples of each criterion, in order to verify examiners' knowledge about epidemiological diagnosis, according to WHO (1); to instruct the examiner in the use of the criteria and the examination method; and finally, to achieve an initial standardization of the 11 examiners as regards the criteria used in the study (e.g.: to measure the ability to diagnose IL, mainly focusing on the clinical characteristics, according to the location). Two tests, one before and one after the benchmark examiner's explanation were applied to the examiners, using 25 clinical photographic slides in each one. In the first test, the examiners described what they thought about the slides representing the various clinical conditions. Afterwards, the benchmark examiner gave them all theoretical explanations about the criteria and codes besides other general information, which could be important for the development of the study. Later, in the second test, other 25 different slides were projected and answered by the examiners in another paper sheet. For this second one, examiner had to mark both, the code and condition regarded to each slide.

The clinical training sessions were held, followed by the calibration exercises. Both clinical training and calibration were done in an outdoor setting under standardized conditions such as adequate natural light (e.g. on sunny days), with dental mirror and ball-ended CPI probes with a diameter of 0.5mm (to clear up doubts about visual diagnosis, remove debris and assess presence of fissure sealants), dental drying and previous tooth brushing. Tooth brushing was done before the individuals were examined, according to the modified Bass technique with fluoridated dentifrice for a standardized time of 2 minutes. Prior dental drying was carried out during the examinations for about 5 seconds per tooth with the use of compressed air through a dental compressor (Proquest Delivery System, model 4010, Compressor Technologies LTD, Englewood, Colorado, USA). During the examinations, all the examiners were helped by note takers.

For each period of clinical training, each dentist examined 10 to 12 children, with different dental caries prevalence. Discussions were held among the examiners and the benchmark examiner during the training, regarding clinical diagnosis, codes and criteria used, recording and other errors.

After the clinical training exercises, the examiners undertook two calibration exercises with an interval of one week between them (2 periods of 4 hours). As mentioned above, the children presented different clinical situations, especially cavitations in dentine and IL. In the calibration phase a 12th benchmark examiner joined the examiners. They examined the same group of children (n=12) in both periods. No discussion was permitted among the examiners and the benchmark examiner with regard to interpretation of the criteria during these calibration phases.

It is important to mention that to avoid problems with getting children to cooperate and growing tired, in each training and calibration period, the same children were always first examined by 6 dentists. After a 1 1/2 hour interval they returned to be examined by other 5 dentists (including the benchmark examiner in the calibration section), in the same place and with all examination conditions standardized. The mean time for each examination period was 55 minutes, and another 20 minutes spent to take children from the classes and to perform previous dental brushing.

Main outcome measures - Two diagnostic thresholds were used to calculate the reliability/reproducibility and the validity/accuracy of examiners: WHO (1) diagnostic threshold, in which caries is considered a cavitated lesion, and WHO+IL diagnostic threshold, including those active IL. For the WHO+IL, an adaptation of the WHO codes with the inclusion of IL was developed (Table 1). For the analysis of the results, both units of measurement, tooth and dental surface, were used for primary and permanent teeth.

Statistical analysis- The results of the first calibration exercise were used to calculate the inter-examiner reproducibility and the validity (S, SP, PPV, NPV) while the first and the second calibration exercises (after an interval of one week) were used to calculate the intra-examiner reproducibility, according to different diagnostic thresholds (WHO and WHO+IL), for both units of measurement. High reproducibility values were considered for Kappa>0.85, for inter and intra-examiner reproducibility (1). Data from the first calibration exercise were also used to calculate inter-examiner Kappa values according to each tooth.

It is important to point out that mean Kappa values (inter-examiner reproducibility) are the results of the final mean which were obtained through crossovers among the 11 examiners (examiner versus examiner), e.g.: examiner 1 *versus* examiner 2, examiner 1 *versus* examiner 3 ... 1x11; 2x3; 2x4....10x11. The benchmark examiner's results were not compared with those of examiners to calculate the inter-examiner reproducibility. Inter, intra-examiner Kappa values and Kappa for each tooth were unweighted Kappa, as all disagreements were considered equally serious.

Validity values were calculated comparing the examiners with the benchmark examiner, and obtaining a final mean value for each measurement (S, SP, PPV, NPV). It is important to observe that it is not possible to assess diagnosis validity against a true (histological) 'gold standard' in

TABLE 2
Mean inter and intra-examiner Kappa values for teeth and surfaces at WHO and WHO+IL diagnostic thresholds during the calibration phase, Piracicaba, SP, Brazil, 2004

	Tooth		Surface	
	WHO+IL	WHO	WHO+IL	WHO
INTRA	0.97 (0.99)	0.99 (1.00)	0.99 (1.00)	1.00 (1.00)
	[0.93—1.00]	[0.96—1.00]	[0.98—1.00]	[0.98—1.00]
INTER	0.90	0.95	0.96	0.98
	[0.85—0.96]	[0.93 – 0.99]	[0.95 – 0.98]	[0.97—0.98]

( ) Kappa Value from the benchmark examiner.
[ ] Examiners' Kappa Value Intervals

caries prevalence surveys. Therefore, it is usual to substitute the diagnoses made by a benchmark examiner as a gold standard, in order to facilitate the assessment of benchmark validity (9).

To complement mean DMFT and dmft, either including the IL or not, for 6 to 7-year-old children by examiner (n=11), the size and direction of deviation (d) from the benchmark examiner (BE) were calculated for the first period of calibration.

Results

The mean Kappa for all the examiners in the final theoretical exercise was 0.86.

Excellent mean intra and inter-examiner Kappa value results (K≥0.90) according to tooth and surface were found for both diagnostic thresholds during the calibration phase. However, the intra and inter-examiner Kappa values according to the WHO diagnostic threshold were slightly higher than those according to the WHO+IL diagnostic threshold. For instance, for the inter-examiner reproducibility according to the WHO+IL threshold, the mean Kappa value was 0.90 (examiners' range: 0.85-0.93) for tooth, and 0.96 (examiners' range: 0.95-0.98) for surface. Kappa value, according to the WHO threshold, was 0.95 (examiners' range: 0.93-0.99) for tooth, and 0.98 (examiners' range: 0.97-0.98) for surface (Table 2).

Analysis according to each tooth showed lower mean inter-examiner Kappa values for the posterior teeth, mainly according to the WHO+IL threshold, compared with the front teeth, showing that it was more difficult to calibrate examiners in those regions. The lowest values were for the permanent molars with a variation from 0.51 to 0.69 and for the second primary molar (number 85) with a Kappa value of 0.63 (table 3).

Results of high accuracy (S, SP, PPV and NPV) were found for the WHO diagnostic threshold, while specificity and negative predictive mean values were high (K≥0.96) and S and positive predictive values were considered moderate to high according to WHO+IL diagnostic threshold. For instance, S mean value was 0.71 (examiners' range: 0.46-0.84) for tooth, and 0.82 (examiners' range: 0.58-0.99) for surface, according to the WHO+IL threshold. Positive predictive mean value was 0.75 (examiners' range: 0.65-0.90) for tooth, and 0.85 (examiners' range: 0.79-0.92) for surface, according to the WHO+IL threshold (Table 4).

In general, there were small variations for the dmft and dmft+IL in either direction from the benchmark. However, deviation values were considered proportionally higher for the DMFT whether or not the IL was included (2) (Table 5).

TABLE 3
Mean inter-examiner Kappa values for each tooth, according to WHO and WHO +IL diagnostic thresholds. Piracicaba, SP, Brazil, 2004

Tooth	WHO	WHO +IL	Tooth	WHO	WHO +IL
16	0.72	0.55	36	0.60	0.51
55	0.90	0.78	75	0.89	0.72
54	0.90	0.78	74	0.89	0.79
53	0.95	0.90	73	1.00	1.00
52/12	1.00	1.00	72/32	1.00	1.00
51/11	0.97	0.98	71/31	1.00	1.00
61/21	0.96	0.95	81/41	1.00	1.00
62/22	1.00	0.98	82/42	1.00	1.00
63/23	1.00	1.00	83	1.00	1.00
64	0.97	0.85	84	0.98	0.78
65	0.83	0.78	85	0.85	0.63
26	0.80	0.69	46	0.77	0.52

Discussion

This aim of this study was to verify the reproducibility (reliability) and validity (accuracy) of the calibration at two different thresholds: WHO (1) diagnostic threshold, which is usually used in surveys and (2) the WHO+IL diagnostic threshold, which could generate more diagnostic errors among the examiners due to the inclusion of IL. The choice of the "heterogeneous" group (although all examiners had already had some experience in WHO epidemiological surveys) was made to evaluate whether standardization would be achieved after the training phase, even with these examiners, whose clinical experience and professional backgrounds were different. Kwan *et al.* (1996) also used a heterogeneous group (dentists and dental auxiliaries) to verify the reproducibility of caries diagnoses according to WHO criteria (1,13). Fyfee *et al.* (2000) used examiners with clinical experience, public or non-public workers, to verify the effect of distinct diagnostic thresholds on the reliability and validity of epidemiological diagnosis of dental caries (9).

In general, intra and inter-examiner reproducibility values, as well as accuracy values (S, SP, PPV e PNV) were high for the WHO threshold, therefore, indicating that the expecta-

TABLE 4

Diagnostic performance measurements for the examiners. Data from calibration at different diagnostic thresholds (WHO and WHO+IL), using the benchmark examiner as a 'gold standard' – unit of evaluation, tooth and surface, Piracicaba, SP, Brazil, 2004

Threshold	unit	S	SP	PPV	NPV
WHO+IL	tooth	0.71 (0.46—0.84)	0.97 (0.95—0.99)	0.75 (0.65—0.90)	0.96 (0.93—0.98)
WHO	tooth	0.90 (0.84—0.96)	0.98 (0.97—0.99)	0.89 (0.82—0.96)	0.98 (0.98—0.99)
WHO+IL	surface	0.82 (0.58—0.99)	0.99 (0.98—0.99)	0.85 (0.79—0.92)	0.98 (0.97—1.00)
WHO	surface	0.96 (0.92—1.00)	1.00 (0.99—1.00)	0.96 (0.91—1.00)	1.00 (0.99—1.00)

Mean: S=sensitivity, SP=specificity, PPV-positive predictive value, NPV= negative predictive value

( ) Range of the measurements.

tions for calibrating this 6 to 7-year age group were perfectly achieved (Tables 2 and 4). The visual tactile method associated with diagnostic adjuncts such as prior dental brushing and drying under natural light were used in the examinations.

Some criticism could be leveled at the use of such diagnostic adjuncts, mainly in the case of tooth drying, which is not used in WHO surveys and it might facilitate the diagnosis of not only the IL but also of cavitated lesions. However, Assaf et al. (2004), when comparing epidemiological examinations by the visual tactile method (WHO standard method), with or without the association of tooth drying and brushing, and the examinations in a dental setting for groups of low and moderate caries, showed that dental drying does not improve the diagnosis of lesions at the WHO diagnostic threshold but brushing does for the low caries prevalence groups. On the other hand, for examinations according to WHO +IL threshold, both adjuncts should be used to facilitate the diagnosis of IL, mainly dental drying (11). Therefore, dental drying is not important for facilitating the detection of cavitated lesion but is relevant for the diagnosis of IL.

As the intention was to standardize all examination conditions according to both diagnostic thresholds, brushing and drying were used, but not artificial light. Although artificial light has been used in examinations with criteria that include IL lesions (7,8,9), this diagnostic adjunct is not currently used in WHO surveys. Thus, agreement among examiners in calibration according to the WHO threshold could have been be overestimated if artificial light had been used in the present study. Moreover, the intention was to verify whether good agreement results could be achieved at the WHO+IL diagnostic threshold even under natural light in an outdoor setting, as the natural light is usually recommended in tropical countries because of the good light conditions during the year. The same cannot be said of cold countries, where examinations are only possible under artificial light in indoor spaces.

The literature has shown a decrease in the prevalence of dental caries in many areas all over the world (5,14). It is interesting to note that even now, most of the surveys still consider caries as a cavitated lesion, ignoring all the IL. However, even in studies that include the IL diagnosis, the lack of standardized criteria and terminology among them is evident (6,8,9). Therefore, the decision to use less sensitive criteria, such as the WHO diagnostic threshold, was still justified for a number of reasons: restorative interventions are usually carried out at the dentin cavitation stage; adequate levels of reliability in the dental caries diagnosis are more difficult to achieve when the IL are included in evaluations, mainly when large numbers of examiners have to be used; epidemiological conditions do not allow a detailed examination when compared with standard conditions, such as in a dental setting (1,4,9,11).

Therefore, doubts have been raised with regard to reliability at more sensitive diagnostic thresholds. However, reliable and successful recording of IL (with the use of portable equipment or with the association of diagnostic adjuncts during epidemiological examinations) has been observed (6,8,9). In this study, except for the S (0.71-tooth and 0.82-surface) and positive predictive value PV (0.75-tooth) results, for the WHO+IL diagnostic threshold, excellent reproducibility (intra and inter agreement) and accuracy (S, SP, PPV, NPV) results were found for both evaluation units (tooth and surface) (Tables 2 and 4).

TABLE 5

Mean DMFT and dmft either including the IL or not, for 6- to 7-year-old children by examiner (n=11) and the size and direction of deviation (d) from the benchmark examiner (BE) for the first period of calibration, Piracicaba, SP, Brazil, 2004

Examiner	dmft (d)	DMFT (d)	dmft+IL (d)	DMFT+IL (d)
BE	3.25	0.00	3.92	0.08
1	3.08 (-0.17)	0.00 (0.00)	3.75 (-0.17)	0.00 (-0.08)
2	3.08 (-0.17)	0.00 (0.00)	3.50 (-0.42)	0.00 (-0.08)
3	3.00 (-0.25)	0.00 (0.00)	3.67 (-0.25)	0.08 (0.00)
4	3.00 (-0.25)	0.00 (0.00)	3.42 (-0.50)	0.08 (0.00)
5	3.58 (+0.33)	0.08 (+0.08)	4.08 (+0.16)	0.25 (+0.17)
6	3.58 (+0.33)	0.08 (+0.08)	4.17 (+0.25)	0.17 (+0.09)
7	3.42 (+0.17)	0.08 (+0.08)	4.00 (+0.08)	0.25 (+0.17)
8	3.42 (+0.17)	0.17 (+0.17)	4.17 (+0.25)	0.58 (+0.50)
9	3.42 (+0.17)	0.17 (+0.17)	4.17 (+0.25)	0.25 (+0.17)
10	3.58 (+0.33)	0.25 (+0.25)	4.42 (+0.50)	0.75 (+0.67)
11	3.33 (+0.08)	0.08 (+0.08)	4.00 (+0.08)	0.17 (+0.09)

However, the most relevant errors were related to IL diagnosis (isolated lesions and lesions associated with decayed teeth - active cavitated lesions). In addition, most of the difficulties with calibrating examiners were concentrated on the posterior molars, mainly the permanent ones, showing moderate to good Kappa results (variation from 0.51 to 0.69) (Table 3). The difficulties with correctly diagnosing posterior teeth are justifiable because they are difficult to see without the use of artificial light, and there are also inherent difficulties with diagnosing IL, mainly under epidemiological conditions. Therefore, differently from examinations according to WHO threshold, the results of the present study showed the importance of using all diagnostic adjuncts, including artificial light, when examining IL in surveys. New studies with similar methodologies and the use of the latter adjunct might confirm this observation. Some discussion could also be raised because of the diagnoses made by a benchmark examiner, as a gold standard does not represent the 'true diagnosis', such as the histological evaluation or even the determination of the depth of lesion penetration after minimal operative intervention currently used in clinical research. Therefore, the validity values reported in this study should be interpreted with some caution. The aim of the analysis of the size and direction of deviation (d) of the examiners' DMFT/dmft and DMFT+IL/dmft+IL means for 6 to 7-year-old children, in relation to the benchmark examiner (BE), for the first period of calibration at both diagnostic thresholds, was to measure the distance between each examiner and the benchmark examiner, who was assumed to have found the 'real diagnostic value'. In general, this study presented small variations in deviations for dmft at both diagnostic thresholds (= 0.5), while deviation values were considered proportionally higher for the DMFT whether or not the IL was included (Table 5) (2). This analysis, although recommended in epidemiological surveys (2), does not correctly reflect the coincidence of diagnosis in relation to the benchmark examiner. Therefore, differently from the Kappa statistic, which aims to make a comparison between two examiners according to each tooth/surface examined, equal DMF and dmf mean values between two examiners do not always represent agreement in the diagnosis, mainly when the diagnosis of cavitated and IL are included. In this case, Kappa would be the most appropriate statistic for the comparison among examiners and the benchmark examiner.

It is important to point out that this study presents some limitations and different reproducibility and validity results could be achieved if, for instance, the following factors had been present: a smaller group of examiners and examiners with different experiences (e.g. inexperienced versus experienced in surveys) had participated in the study; artificial light had been used during the examinations; children from different regions had participated without being pre-selected, and others. For these reasons, further studies about examiner calibration according to this new diagnostic threshold are still needed in order to improve and establish a calibration methodology that can be used by dental caries epidemiologists and the scientific community.

Acknowledgements

The authors wish to thank the director, the dentists, the students and those responsible for the children of the public school located in the city of Piracicaba where the research was conducted.

References

1. Oral health surveys: basic methods. 4th ed, p 66. Geneva: WHO, 1997.
2. Pine C M, Pitts N B, Nugent Z J. 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 epidemiology program quality standard. Community Dent Oral Epidemiol 1997; 14:18-29 (supplement).
3. Ismail A I. Clinical diagnosis of precavitated carious lesions. Community Dent Oral Epidemiol 1997; 25:13-23.
4. Pitts N B. Diagnostic tools and measurements-impact on appropriate care. Community Dent Oral Epidemiol 1997; 25:24-35.
5. Barmes D E. A global view of oral diseases: today and tomorrow. Community Dent Oral Epidemiol 1999; 27:2-7.
6. Warren J J, Steven M L, Kanellis M J. Dental caries in the primary dentition: assessing prevalence of cavitated and non-cavitated lesions. J. Public Health Dent 2002; 62:109-114.
7. Pitts N B, Fyfee H E. The effect of varying diagnostic thresholds upon clinical caries data for a low prevalence group. J. Dent Res 1988; 67:592-596.
8. Nyvad B, Machiulskiene V, Baelum V. Reliability of a new caries diagnostic system differentiating between active and inactive caries lesions. Caries Res 1999; 33:252-260.
9. Fyffe H E, Deery C, Nugent Z J *et al.* 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 2000; 28:42-51.
10. Rimmer P A, Pitts N B. Effects of diagnostic threshold and overlapped approximal surfaces on reported caries status. Community Dent Oral Epidemiol 1991; 19:205-212.
11. Assaf AV, Meneghim MC, Zanin L, Mialhe FL, Pereira AC, Ambrosano GMB. Assessment of different methods for diagnosing dental caries in epidemiological surveys. Community Dent Oral Epidemiol 2004; 32:418-25.
12. Brazil. Health Ministry. Secretary's Office of Health Politics. Department of Basic Attention. Technique Área of Oral Health. Project SB 2000: oral health conditions of the Brazilian populations in the year 2000: manual of examiners' calibration. Brasília, 2001. Series C. Projects, Programs and Reports, n.54 (in Portuguese).
13. Kwan S Y L, Prendergast M J, Williams S A. The diagnostic reliability of clinical dental auxiliaries in caries prevalence surveys-a pilot study. Community Dental Health 1996; 13:145-149.
14. Marthaler TM. Changes in dental caries 1953-2003. Caries Res 2004; 38:173-181.