Title: Reliability analysis using the in-lux examination method for dental indices in adolescents for use in epidemiological studies

Rafaela da Silveira Pinto[1]
Mario Vianna Vettore[2]
Mauro Henrique Guimarães Nogueira de Abreu[1]
Andrea Clemente Palmier[1]
Rosa Núbia Vieira de Moura[1*]
Angelo Giuseppe Roncalli[3]

1 Department of Community and Preventive Dentistry, School of Dentistry, Federal University of Minas Gerais.
2 Department of Health and Nursing Sciences, University of Agder
3 Department of Dentistry, Federal University of Rio Grande do Norte

* Correspondence author: Rosa Núbia Vieira de Moura
Address: Avenida Antônio Carlos, 6627 – Belo Horizonte, Minas Gerais, Brazil, ZIP Code 31270.901; Phone number: +55 31 34092448; e-mail: bdmoura96@gmail.com

- **Acknowledgements:** This study was funded by the Postgraduate program in dentistry at the School of Dentistry, Federal University of Minas Gerais.

- **Ethics and inform consent:** This study was approved by the Research Ethics Council of the Federal University of Rio Grande do Norte (Registration 32077014.8.0000). The consent form was signed by the parents of adolescents under 18 years of age, by the adolescents, and by the examiners.

- **Data availability:** The data that support the findings of this study can be requested to the first author in the above list (rafaelasilveirapinto@gmail.com).

- **Conflict of interest:** The authors declare no conflict of interest.

- **Authors' contribution: RSP** Conception and design of the study, acquisition, analysis, and interpretation of data, drafting of the article. **MVV, MHGNA, AGR:** Contributions to conception and design, analysis, and interpretation of data, drafting of the article. **ACP, RNVM:** Contributions to conception and design, acquisition of data, drafting of the article. All authors approved the final version.

Abstract

Objective: To evaluate the *in-lux* examination method to assess the reliability of examiners for oral health surveys.

Methods: A calibration study involving 10 examiners and 27 patients was conducted, and high-quality photos of dental epidemiological indices were projected. The latter refers to the *in-lux* examination method. Two groups, comprised of five examiners each, were trained to assess dental caries (DMFT index) and malocclusion (DAI). The first group carried out *in vivo* (clinical) and *in-lux* examinations in the same patients. The second group performed *in-lux* examinations only. The measurements were repeated to obtain intraexaminer weighted kappa coefficients.

Results: Inter-examiner weighted kappa coefficients of the *in vivo* examination method for DMFT and DAI ranged from 0.597 to 0.851 and from 0.574 to 0.844, respectively. The values for *in-lux* examination were between 0.440 and 0.856 (DMFT), and between 0.524 and 0.783 (DAI). The intraexaminer kappa coefficients of the *in vivo* examination method ranged between 0.569 and 0.851 (DMFT), and between 0.644 and 0.834 (DAI). In the *in-lux* method, these values were between 0.426 and 0.831 (DMFT), and between 0.341 and 0.838 (DAI). Three examiners did not reach the minimum acceptable kappa value (k=0.610) for DMFT, and one for DAI. Of these, one examiner managed to reach the minimum kappa coefficient for DMFT after additional training. The three others maintained the kappa coefficients at lower than acceptable limits.

Conclusions: Calibration using both *in vivo* and *in-lux* examination methods were able to discriminate the examiners regarding their reliability to reproduce dental indices. The *in-lux* examination method was considered reliable and may replace the *in vivo* examination method. The *in-lux* method might be more feasible to evaluate the reliability of examiners when clinical calibration is unrealistic due to logistic issues and when a large number of examiners are involved in the survey.

Keywords: Oral Health; Survey Method; Epidemiological Studies.

Introduction

Population-based oral health surveys are conducted to estimate the prevalence of the main oral health conditions, resulting in relevant information to support public policies and for the adequate planning and organization of dental services. Planning and carrying out these surveys are complex and expensive actions, and require adequate methodological quality assurance to generate unbiased oral health measurements. Clinical or *in vivo* calibration of examiners prior to fieldwork data collection has been recommended by the World Health Organization (WHO) to obtain consistent and reliable measurements when registering dental indices during fieldwork data collection. Clinical calibration in oral health epidemiological studies refers to the assessment of the reliability of examiners by repeating dental examinations with the same participants, performed by different examiners (interexaminer reliability), or by the same examiner at different time points (intraexaminer reliability).[1,2] Most dental indices used in oral health surveys result in ordinal measures. Therefore, the weighted kappa coefficient is considered the most appropriate measure to assess reliability in the use of dental indices.[3,4]

The most commonly used international manuals for assessing the oral health status and the oral health needs of a population in oral health surveys were proposed by the WHO[5] and by the British Association for the Study of Community Dentistry.[6] These manuals recommend conducting a calibration study prior to fieldwork data collection, which consists of training the examiners and then evaluating their consistency to register dental indices, using an experienced examiner as the gold-standard or reference. An interval period of at least two days between the theoretical and fieldwork training is recommended so that the examiners can incorporate the knowledge.[5] Photographs of different clinical conditions representing the different codes of dental indices are frequently used in theoretical discussions as part of the training of examiners before calibration[7] and can be employed in the calibration itself when the dental condition is of low prevalence.[4]

Oral health surveys conducted in England,[8] Belgium,[9] the USA,[10,11] and Brazil[12,13] have calibrated the examiners through *in vivo* (clinical) examinations before the survey's data collection. The use of the *in vivo* method has a number of possible obstacles, mainly in large population-based oral health surveys involving a considerable number of examiners who are distributed throughout the country. Consequently, several participants should be clinically examined at different time points to assess the inter and intraexaminer agreement. The *in vivo* examination method also has logistic issues, since they must be conducted in specific settings (e.g. schools) where the participants can be easily recruited.[4] In addition, the use of the *in vivo* examination method has important ethical implications due to the long time taken by participants who are examined multiple times. The repetition of dental exams may also cause pain, discomfort, and possible damage to the soft oral tissues during periodontal assessment.[14]

The *in-lux* examination method follows the same principles as the *in vivo* method. However, the participants' clinical examinations are replaced by the projection of high-quality photos of the epidemiological dental indices, which supports the use of the term "*in-lux*". Despite not precisely reproducing the fieldwork conditions, the *in-lux* examination method foresees several advantages over the *in vivo* method. First, it is possible to evaluate the reliability of all conditions of the same dental index, including those that are less prevalent and therefore unlikely to occur during the *in vivo* calibration study. Second, the *in-lux* examination method enables the assessment of the reliability of a large number of examiners in a short period of time. Thus, the interval between calibration and fieldwork data collection is significantly lower when compared to the *in vivo* method. Third, the *in-lux* method enables the evaluation of reliability when the examiners are from different locations, avoiding large displacements and saving resources. Finally, the *in-lux* examination method does not require participants and a large infrastructure for dental exams. However, it should be acknowledged that the *in-lux* examination method is not suitable to assess the reliability of some oral health conditions, particularly those where the exam is not merely visual, such as periodontal conditions when periodontal probing is necessary to register CPI and PIP.[15]

Calibration of examiners using the *in-lux* method has already been used for epidemiological oral health surveys, including in Brazil, at the municipal,[16] state,[17] and national levels[13], as well as in Italy.[18] However, there is a dearth of studies comparing the reliability of calibration using the *in-lux* and *in vivo* methods. The aim of this study was to evaluate the *in-lux* examination method to assess the reliability of dental examiners for epidemiological oral health surveys.

Methods

The present study was approved by the Research Ethics Committee of the Federal University of Rio Grande do Norte (UFRN) (protocol No. 32077014.8.0000). A signed consent form was obtained from the parents of participants aged 18 years or younger, adolescents, and examiners.

A calibration study involving 10 examiners and 27 adolescent patients was conducted. All examiners were experienced dentists (mean age = 35.5 years), who on average completed the undergraduate dental course 10.6 years ago and have completed at least one postgraduate course in dentistry.

The 5th version of the WHO Manual for Epidemiological Surveys suggests that calibration be performed on adolescents aged 12 to 15 years.[5] The patients were young apprentices from two companies that provide services to Federal University of Minas Gerais (UFMG) UFRN. The Young Apprentice Program is available for teenagers aged 12 to 19. They were registered at the dental clinic of the School of Dentistry of the UFMG in Brazil. Inclusion criteria were the patient's age, between 12 and 19 years, and availability and agreement to be dentally examined twice and to have photographs of their mouth taken. Initially, 35 adolescents were invited. Of them, five adolescents were excluded according to the eligibility criteria.

The evaluation of consistency using the *in-lux* examination method involved ten examiners (dentists), divided into two groups: Group 1: examiners 1 to 5, dental examinations using *in vivo* and *in-lux* methods, and Group 2: examiners 6 to 10, dental examinations using the *in-lux* method. Both groups of examiners participated in a theoretical training, consisting of a two-hour video recorded lecture with explanations and comments of possible doubts on the use of the dental indexes. Group 1 then performed a simulation of the *in vivo* dental exams involving three adolescents. The remaining 32 adolescents were scheduled to attend dental examinations on two occasions and to attend a session for the photographs to be taken. Of them, five adolescents did not attend all sessions and were thus excluded, resulting in a final sample of 27 participants. Clinical examinations were performed privately in dental office rooms in a public university in Brazil under natural light. The dental examinations were carried out using a ball-ended WHO probe and plain dental mirror to record the number of decayed, missing, and filled teeth (DMFT index) and the Dental Aesthetics Index (DAI). The DAI evaluates malocclusion according to three dimensions of dentition, spacing, and occlusion based on 10 occlusal characteristics. The scores of DAI are calculated by adding the item scores, which are multiplied by their coefficients (weights). A constant is then added to the summated score. Higher values indicate worse malocclusion and greater orthodontic treatment need. DAI scores are grouped into four categories: no abnormality or minor malocclusion (DAI ≤25); definite malocclusion (DAI = 26–30); severe malocclusion (DAI = 31–35); and very severe or handicapping malocclusion (DAI ≥36).[19]

After the dental examinations, a consensus was reached among the examiners on the oral conditions of the participants for DMFT and DAI indexes. Any disagreements were resolved by a further examination of the adolescent until a consensus was reached. Dental examinations were carried out twice by group 1 within a 48-hour period.[20] Photographs of the clinical conditions of the adolescents were taken to evaluate the *in-lux* examination method. Twelve photographs of each volunteer were taken, in different positions, to test the *in-lux* method. A Nikon camera model D750, Macro Lens 100 mm, and Nikon circular and twin flashes were used. The professional photographer had taken dental photography courses totaling 200 hours of training.

Four months after the *in vivo* examinations, groups 1 and 2 examiners attended an additional theoretical training on dental indices to assess their reliability by the *in-lux* calibration method. The same recorded video lecture employed in the initial training was used, followed by a discussion concerning the doubts about the dental indices. The two groups of examiners simultaneously registered the dental conditions of the adolescents using the photographs taken previously at the time of the *in vivo* examinations on a 32-inch monitor in a low-light room. Each group of examiners performed the consensus of the dental conditions on separate days. Any disagreements were resolved by discussion among the examiners of each group with the discretion to review the photographs of the tooth images until a consensus was reached. The dental indices were recorded again, using the photographs one week after the first exam.

The examiners who did not reach a substantial agreement[21] (weighted kappa coefficient ≥ 0.610) in the intraexaminer agreement using the *in-lux* method were retrained and the photographs were re-evaluated in order to obtain new weighted kappa coefficients. The sequence of dental exams according to the groups is presented in Appendix 1.

Appendix 1. Flowchart of the sequence of exams performed by the groups of examiners.

Weighted kappa coefficients were calculated to assess interexaminer and intraexaminer agreement. In the former, records of the dental indices of each examiner were compared with the consensus of their group. Intraexaminer agreement was assessed, contrasting the dental records of the first and second examinations of each examiner using the *in vivo* method. Subsequently, the interexaminer weighted kappa coefficients of the two groups were calculated using the consensus of each group for the *in-lux* method. The intraexaminer kappa coefficients were also calculated using the *in-lux* method. In addition, the kappa coefficients between the consensus of *in vivo* versus *in-lux* of group 1, *in vivo* of group 1 versus *in-lux* of group 2, and *in-lux* of group 1 versus *in-lux* of group 2 were calculated to evaluate the consistency between the consensuses of the different methods (Appendix 2). All analyses were performed using the SPSS statistics software, version 23.0.[22]

Appendix 2. Flowchart of the study analysis.

In summary, the sequence performed by the examiners was: each examiner individually examines the volunteers in the *in vivo* technique and the photos in the *in-lux* technique. Consensus is made by the team of examiners, re-examining the volunteers or the photos in the case of the *in-lux* technique when there are divergences in the answers of the examiners among themselves. The gold standard was defined as the consensus. After reaching the consensus, the individual exams are compared to it, and the weighted kappa of each examiner was calculated.

Results

Table 1 shows the weighted kappa coefficients between the examiners and the consensus of the respective groups, as well as the intraexaminer weighted kappa values according to the *in vivo* and *in-lux* examination methods for the DMFT index. The weighted kappa coefficients obtained after retraining the three examiners who did not reach substantial agreement in the first attempt (kappa ≥ 0.610) are also presented. These values were close to the threshold recommended for an examiner to be considered fit for fieldwork data collection, except for examiners 7, 9, and 10 in the *in-lux* method. These examiners also failed to reach acceptable results in the intraexaminer reliability analysis. After new training, examiner 7 reached a weighted kappa value ≥ 0.610 for the DMFT index and was thus considered calibrated for data collection. Examiner 4 did not reach the minimum kappa value using the *in vivo* method.

Table 2 presents weighted kappa coefficients between examiners and the consensus of the groups, as well as intraexaminer weighted kappa values according to *in* *vivo* and *in-lux* methods for categorical values of DAI. In addition, weighted kappa values after retraining the examiner who did not obtain the minimum kappa coefficient of 0.610 are also reported. Examiner 2 was retrained but the minimum value of 0.610 was not reached, indicating the inconsistency to register DAI in fieldwork data collection.

The comparison of the consensus achieved by examiners 1 to 5 between the *in vivo* method and in the *in-lux* method resulted in weighted kappa values of 0.919 (95% CI, 0.863-0.975) for DMFT and 0.743 (95% CI, 0.539-0.947) for DAI assessed as ordinal data. These values indicate good agreement between the two examination methods for the same group of examiners. The consensus represents the epidemiological classification of the group of examiners on a given disease. A high value of weighted kappa in the comparison between the consensuses in different techniques shows that, for a certain group of examiners, the evaluation in different techniques has a high correspondence.

The weighted kappa coefficients of 0.832 (95% CI, 0.737-0.927) for DMFT and 0.443 (95% CI,0.173-0.712) for DAI categories were obtained between the consensus achieved by examiners 1 to 5 using the *in vivo* method and the consensus of examiners 6 to 10 using the *in-lux* method. The weighted kappa coefficients between the consensus achieved by examiners 1 to 5 using the *in-lux* method and the consensus of examiners 6 to 10 using the *in-lux* method were 0.873 (95% CI, 0.791-0.954) for DMFT and 0.711 (95% CI, 0.525-0.897) for DAI categories.

Discussion

The present study evaluated the *in-lux* examination method to assess the reliability of measuring dental indices in population-based epidemiological oral health surveys using the *in vivo* method as a gold standard. The reliability assessment of both indexes proved to be adequate for most examiners. Calibration using the *in-lux* examination method was also able to discriminate the examiners who demonstrated adequate capacity to register the dental indices with consistency and those who did not. The latter included 20% of the examiners for DMFT and 10% for DAI, after retraining.

Even though the kappa coefficient is considered a robust method to assess reliability of measurements, some considerations related to its use and interpretation are needed. The kappa coefficient can be applied to ordinal variables derived from continuous data, as can be seen in the DAI that was analyzed as ordinal data considering the four categories of malocclusion. However, the interpretation of kappa coefficients when ordinal categories derive from continuous data depend on the choice of category limits.

Using the kappa coefficient in such cases might reduce the discriminatory power of the original index and thus influence the judgment of the index's reliability.[23] Another aspect that deserves attention is the caution when interpreting kappa coefficients in some situations. For instance, in the studied sample, the kappa coefficient is influenced by the prevalence of the health condition.[23,24] In the present study, the mean of DMFT was considered low, ranging from 0.98 to 1.68, depending on the examination method and the group of examiners. Similarly, the prevalence of severe and very severe malocclusion ranged from 14% to 18%, according to the DAI classification. Diseases of low and high prevalence are more likely to be over and underestimated, respectively, in health surveys. Thus, the exams can be influenced by the diagnostic behavior of the examiners according to the distribution of the disease in a given population.[23] Finally, a paradox in the kappa coefficient is pointed out in the following situations: (1) low kappa values, despite the high agreement observed in highly symmetric unbalanced marginals and (2) higher kappa values for asymmetric unbalanced marginal distributions when considering the concordance tables. In this study, the tables for calculating the kappa coefficients were asymmetric, and this issue should be considered as a possible source of bias.[25]

Previous studies that adopted the *in vivo* examination method for examiners' training and to assess consistency highlighted the difficulties of the operationalization of the field work due to the country's size, population coverage, and number of examiners involved. In the United States, between 90 and 100 volunteers were examined during the training prior to data collection. The survey involved 17,463 participants who were examined between 2011 and 2014.[11] The UK adult dental survey data collection was conducted by 77 NHS dentists, and 79 individuals participated in the training.[26] In the fieldwork, a total of 6,469 adults were examined.[27] The training for the 2010 Brazilian oral health survey predicted the examination of approximately 100 participants for each group of examiners in different cities. The survey examined 37,519 individuals in 177 cities, involving nearly 2,000 fieldwork professionals.[13] A large number of examiners assessed more individuals in the calibration study than in the field work of the survey.

Calibration using the *in vivo* examination method has been recommended because this method reflects the reality of fieldwork data collection more accurately and simulates the severity of the oral conditions that will be found in the survey. However, previous surveys have criticized the calibration using the *in vivo* examination method and suggested that there was space to improve the calibration process.[28] Calibration using the *in vivo* examination method requires a large number of participants, and discomfort is caused due to repeated dental examinations.[14] Moreover, calibration using the *in vivo* examination method depends upon extensive efforts from volunteers and examiners, as well as a considerable amount of time to perform the exams.

In view of the aforementioned difficulties and limitations, calibration through the *in-lux* examination method was developed and used in the 2010 Brazilian oral health survey to evaluate the reliability of examiners for oral conditions of low prevalence, such as dental trauma and fluorosis.[13] The use of this calibration method was considered innovative and may be useful in forthcoming oral health surveys.[14] The *in-lux* examination method was also successfully used to assess the examiner's reliability to estimate the oral health conditions in another regional population-based survey in Brazil,[17] as well as in the oral health survey of schoolchildren in Pelotas in the South of Brazil. In both surveys, the calibration of examiners involved theoretical training, clinical calibration for common oral diseases, and calibration using the *in-lux* method for oral conditions of low prevalence.[16] The training of examiners that preceded the national children's oral health survey in Italy was conducted by employing photographs of the oral conditions to be evaluated in the survey. During this stage, the examiners also discussed and agreed on specific diagnostic criteria, facilitating the standardization of data collection procedures in the main study.[18] Although the reliability of examiners using the *in-lux* examination method was assessed through kappa coefficients, to date, none of these surveys examined the equivalence between the *in vivo* and *in-lux* examination methods in the calibration exercise. Therefore, the present findings indicate a potential new methodology of dental index calibration of examiners to be used in future oral health surveys.

Calibration using the *in-lux* method has some advantages that should be acknowledged. There are several ethical issues involved in the calibration using *in vivo* examinations that are not implicated when using the *in-lux* method. These include participants' discomfort when undergoing multiple clinical examinations and the need to include several participants when the oral condition is rare. Moreover, the logistics, infrastructure, and time consumption of the *in-lux* examination method are considerably lower than the *in vivo* method. In this study, several groups of examiners performed their training in different locations with different participants, and each group developed and used their own consensus. The *in-lux* method also provides a quick and standardized means of data collection. In addition, the use of photographs offers practicality and comfort to the examiners during training and calibration. Similar to the *in vivo* method, in this study, the *in-lux* method was able to identify examiners that did not reach the minimum kappa coefficient and thus should be retrained or replaced by other examiners.

The limitations of this study include the small number of examiners and participants, as well as the low prevalence of the evaluated oral conditions. The interpretation of the weighted kappa to assess the reliability of examiners, comparing the *in vivo* and *in-lux* examination methods in this study, should also be cautious because of the asymmetrical imbalanced marginal of totals and the arbitrary category limits of the DAI.[23,25] There are some issues related to the use of the *in-lux* method that should be acknowledged. First, this method is not applicable to assess the reliability of periodontal measures and periodontal indices, since periodontal examination is influenced by positioning, angulation, and pressure exerted during periodontal probing. In addition, the *in-lux* method requires specific professionals and photography equipment to obtain good quality photographs of oral conditions. The brightness of the clinical environment where the photos are taken should be similar to those where participants will be examined in the survey.

To continually improve the quality of teaching/training packages prior to calibration exercises, it was suggested that future research address examiners' perceptions of training for epidemiological surveys.

Calibration, using both *in vivo* and *in-lux* examination methods, was able to discriminate the examiners regarding their reliability to reproduce dental indices. The *in-lux* examination method was considered reliable and can replace the *in vivo* examination method. The *in-lux* method might be more feasible to evaluate the reliability of examiners when clinical calibration is unrealistic due to logistic issues and when a large number of examiners is involved in the survey. Future studies involving more examiners and assessing the reliability of examiners through the *in-lux* examination method for other oral conditions should be conducted to confirm the present findings.

References

1. WHO (World Health Organization). Calibration of Examiners for Oral Health Epidemiology Surveys. Technical Report. Geneva: WHO; 1993.

2. Peres MA, Traebert J, Marcenes W. Calibration of examiners for dental caries epidemiologic studies. Cad Saude Publica 2001; 17(1):153-9. Portuguese. doi: 10.1590/s0102-311x2001000100016. PMID: 11241938.
3. Silva AF, Velo MMAC, Pereira AC. Importance of reproducibility of methods for Dentistry diagnosis. RFO UPF vol.21 no.1 Passo Fundo Jan./Abr. 2016. (http://revodonto.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-40122016000100018&lng=pt&nrm=iso).
4. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Secretaria de Vigilância em Saúde. SB Brasil 2010 Pesquisa Nacional de Saúde Bucal: Manual de Calibração de Examinadores. Brasília: Ministério da Saúde, 2009.
5. WHO (World Health Organization), 2013. Oral health surveys: basic methods - 5th edition. Technical Report. Geneva: WHO.
6. Pitts NB, Evans DJ, Pine CM. British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97. Community Dent Health. 1997 Mar;14 Suppl 1:6-9. PMID: 9114553.
7. Assaf AV, de Castro Meneghim M, Zanin L, Tengan C, Pereira AC. Effect of different diagnostic thresholds on dental caries calibration - a 12 month evaluation. Community Dent Oral Epidemiol. 2006b Jun;34(3):213-9. doi: 10.1111/j.1600-0528.2006.00278.x. PMID: 16674753.
8. 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 Dent Health. 1997 Mar;14 Suppl 1:18-29. PMID: 9114555.
9. Declerck D, Leroy R, Martens L, Lesaffre E, Garcia-Zattera MJ, Vanden Broucke S, Debyser M, Hoppenbrouwers K. Factors associated with prevalence and severity of caries experience in preschool children. Community Dent Oral Epidemiol. 2008 Apr;36(2):168-78. doi: 10.1111/j.1600-0528.2007.00385.x. PMID: 18333881.
10. CDC. National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES): Oral Health Examiners Manual, 2016. https://wwwn.cdc.gov/nchs/data/nhanes/2015-2016/manuals/2016_Oral_Health_Examiners_Procedures_Manual.pdf.
11. Dye BA, Nowjack-Raymer R, Barker LK, Nunn JH, Steele JG, Tan S, Lewis BG, Beltran-Aguilar ED. Overview and quality assurance for the oral health component of the

National Health and Nutrition Examination Survey (NHANES), 2003-04. J Public Health Dent. 2008 Fall;68(4):218-26. doi: 10.1111/j.1752-7325.2007.00076.x. PMID: 18248340.
12. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. Coordenação de Saúde Bucal. Projeto SB Brasil 2003. Condições de saúde bucal da população brasileira: 2002-2003: resultados principais. Brasília: Ministério da Saúde, 2004.
13. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Secretaria de Vigilância em Saúde. SB Brasil 2010: Pesquisa Nacional de Saúde Bucal: resultados principais. Brasília: Ministério da Saúde, 2012.
14. Andrade FR, Narvai PC, Montagner MÂ. The ethics of in vivo calibrations in oral health surveys. Rev Bras Epidemiol. 2016 Oct-Dec;19(4):812-821. Portuguese, English. doi: 10.1590/1980-5497201600040011. PMID: 28146170.
15. Pinto RS, Lopes DL, Santos JS, Roncalli AG. Projeto SB Minas Gerais 2012: Pesquisa das Condições de Saúde Bucal da População Mineira – Métodos e Resultados Principais. Arquivos em Odontologia, v. 54, 24 dez. 2018. Available: https://periodicos.ufmg.br/index.php/arquivosemodontologia/article/view/3733/9847
16. Goettems ML, Correa MB, Vargas-Ferreira F, Torriani DD, Marques M, Domingues MR, Hallal PC, Demarco FF. Methods and logistics of a multidisciplinary survey of schoolchildren from Pelotas, in the Southern Region of Brazil. Cad Saude Publica. 2013 May;29(5):867-78. PMID: 23702993.
17. Minas Gerais. Secretaria de Estado de Saúde. Subsecretaria de Políticas e Ações de Saúde. Superintendência de Redes de Atenção à Saúde. Diretoria de Saúde Bucal. [Internet] SB Minas Gerais: pesquisa das condições de saúde bucal da população mineira: resultados principais. Belo Horizonte: SES-MG, 2013. [Acces: 26 nov. 2019]. Available: www.saude.mg.gov.br/images/documentos/SBMinas_Relatorio_Final.pdf
18. Castiglia P, Campus G, Solinas G, Maida C, Strohmenger L. Children's oral health in Italy: training and clinical calibration of examiners for the National Pathfinder about caries disease. Oral Health Prev Dent. 2007;5(4):255-61. PMID: 18173085.
19. Jenny J, Cons NC. Comparing and contrasting two orthodontic indices, the index of orthodontic treatment need and the dental aesthetic index. Am J Orthod Dentofacial Orthop 1996; 110: 410–416.

20. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to their Development and Use. 3rd ed. Oxford, England: Oxford University Press; 2003.

21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74. PMID: 843571.

22. IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

23. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. 2005 Mar;85(3):257-68. PMID: 15733050.

24. Tonello AS, Silva RP, Assaf AV, Ambrosano GM, Peres SH, Pereira AC, Meneghim M de C. Interexaminer agreement dental caries epidemiological surveys: the importance of disease prevalence in the sample. Rev Bras Epidemiol. 2016 Apr-Jun;19(2):272-9. doi: 10.1590/1980-5497201600020005. PMID: 27532751.

25. Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. BMC Med Res Methodol. 2014 Aug 28;14:100. doi: 10.1186/1471-2288-14-100. PMID: 25168681; PMCID: PMC4236536.

26. NHS. Foundation Report: Adult Dental Health Survey 2009 (Technical information). 2011a Available: https://files.digital.nhs.uk/publicationimport/pub01xxx/pub01086/adul-dent-heal-surv-summ-them-foun-2009-re14.pdf> Access: 28 april 2020.

27. NHS. Executive Summary: Adult Dental Health Survey 2009 2011b. Available: https://files.digital.nhs.uk/publicationimport/pub01xxx/pub01086/adul-dent-heal-surv-summ-them-exec-2009-rep2.pdf> Access: 28 april 2020.

28. Soares FF, Freire MCM, Reis SCGB. The 2010 Brazilian Oral Health Survey (SBBrasil 2010 Project): what do the coordinators propose for future surveys? Interface (Botucatu) vol.21 no.63 Botucatu Oct./Dec. 2017 Epub May 25, 2017. http://dx.doi.org/10.1590/1807-57622015.0963

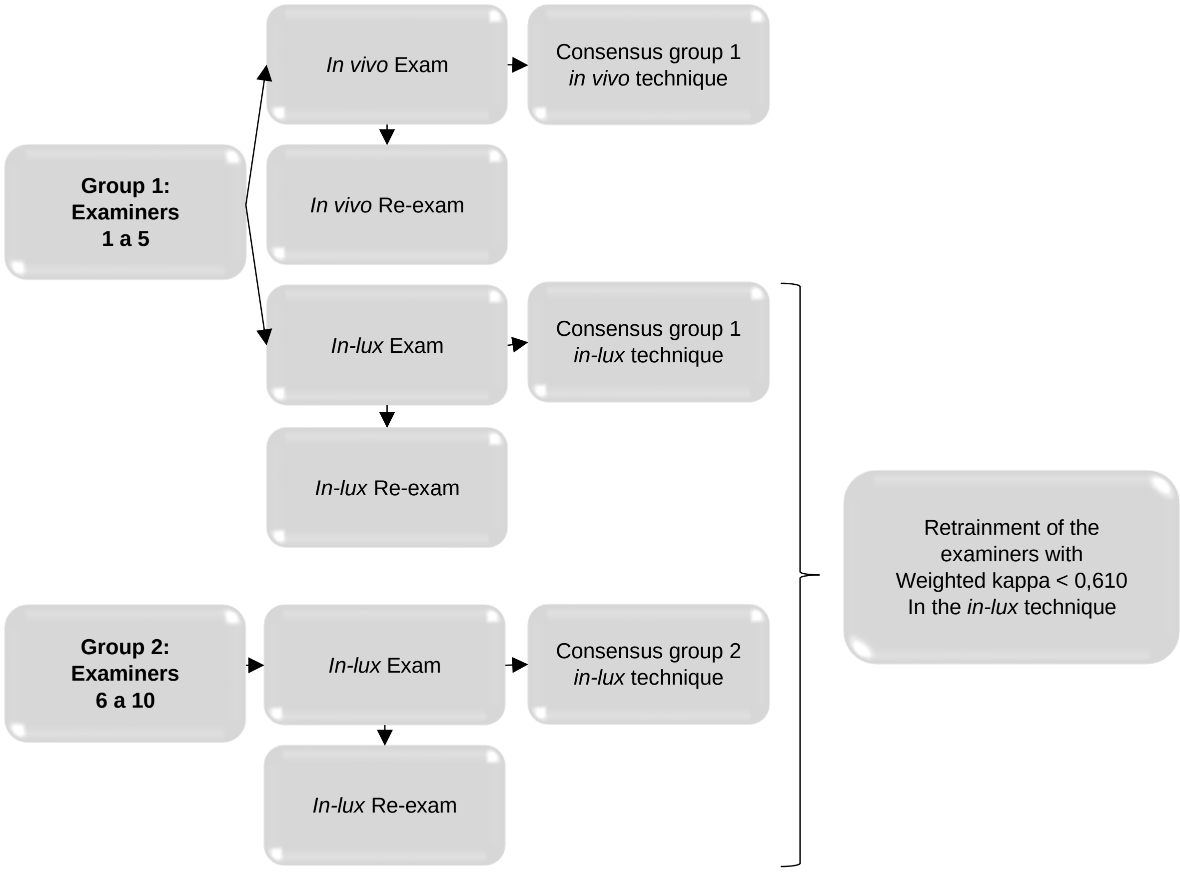
Table 1. Weighted kappa coefficients for DMFT in the *in vivo* and *in-lux* techniques.

Examiner	***In vivo***		***In-lux***		***In-lux (retraining)***
	Weighted Kappa (95% CI)		**Weighted Kappa (95% CI)**		**Weighted Kappa (95% CI)**
	examiner x consensus *in vivo*	**intraexaminer *in vivo***	**examiner x consensus *in-lux***	**intraexaminer *in-lux***	**examiner x consensus *in-lux***
1	0.650 (0.506-0.795)	0.769 (0.632-0.905)	0.798 (0.709-0.887)	0.780 (0.688-0.873)	-
2	0.851 (0.768-0.934)	0.851 (0.750-0.952)	0.847 (0.765-0.929)	0.831 (0.738-0.923)	-
3	0.740 (0.619-0.862)	0.726 (0.595-0.858)	0.773 (0.660-0.886)	0.604 (0.415-0.793)	-
4	0.597 (0.441-0.753)	0.569 (0.418-0.721)	0.679 (0.552-0.806)	0.684 (0.568-0.801)	-
5	0.728 (0.604-0.851)	0.652 (0.526-0.772)	0.698 (0.575-0.821)	0.715 (0.594-0.837)	-
6	-	-	0.856 (0.769-0.943)	0.728 (0.610-0.847)	-
7	-	-	0.549 (0.431-0.666)	0.593 (0.487-0,700)	0.713 (0.595-0.830)
8	-	-	0.659 (0.514-0.805)	0.714 (0.597-0.832)	-
9	-	-	0.440 (0.304-0.576)	0.591 (0.476-0.706)	0.487 (0.359-0.615)
10	-	-	0.463 (0.330-0.596)	0.426 (0.281-0.570)	0.316 (0.152-0.480)

Table 2. Weighted kappa coefficients for the DAI in the *in vivo* and *in-lux* techniques.

Examiner	***In vivo***		***In-lux***		***In lux (retraining)***
	Weighted Kappa (95% CI)		**Weighted Kappa (95% CI)**		**Weighted Kappa (95% CI)**
	examiner x consensus *in vivo*	**intraexaminer *in vivo***	**examiner x consensus *in-lux***	**intraexaminer *in-lux***	**examiner x consensus *in-lux***
1	0.844 (0.701-0.988)	0.834 (0.647-1.000)	0.643 (0.426-0.860)	0.700 (0,467-0,933)	-
2	0.689 (0.440-0.900)	0.644 (0.412-0.877)	0.524 (0.208-0.839)	0.559 (0.242-0.876)	0.554 (0.318-0.789)
3	0.741 (0.583-0.900)	0.746 (0.544-0.948)	0.783 (0.619-0.948)	0.776 (0.639-0.912)	-
4	0.574 (0.346-0.802)	0.670 (0.434-0.906)	0.632 (0.411-0.854)	0.656 (0.459-0.853)	-
5	0.588 (0.255-0.922)	0.648 (0.342-0.954)	0.753 (0.583-0.923)	0.644 (0.459-0.829)	-
6	-	-	0.701 (0.437-0.966)	0.596 (0.345-0.846)	-
7	-	-	0.648 (0.382-0.915)	0.551 (0.282-0.819)	-
8	-	-	0.665 (0.424-0.906)	0.838 (0.673-1.000)	-
9	-	-	0.695 (0.495-0.895)	0.683 (0.482-0.884)	-
10	-	-	0.667 (0.415-0.918)	0.341 (0.028-0.655)	-

Appendix 1. Flowchart for the sequence of exams performed by the groups of examiners.



Appendix 2. Flowchart of the study analysis.