Dental Journal
Majalah Kedokteran Gigi

Dental Journal
(Majalah Kedokteran Gigi)
2018 March; 51(1): 42–46

Research Report

Building team agreement on large population surveys through inter-rater reliability among oral health survey examiners

Sri Susilawati,[1] **Grace Monica,**[2] **R. Putri N. Fadilah,**[3] **Taufan Bramantoro,**[4] **Darmawan Setijanto,**[4] **Gilang Rasuna Sadho,**[4] and **Retno Palupi**[4]
[1] Department of Dental Public Health, Universitas Padjadjaran, Bandung – Indonesia
[2] Department of Dental Public Health, Universitas Maranatha Christian, Bandung – Indonesia
[3] Department of Dental Public Health, Universitas Jenderal Achmad Yani, Bandung – Indonesia
[4] Department of Dental Public Health, Universitas Airlangga, Surabaya - Indonesia

ABSTRACT

Background: *Oral health surveys conducted on a very large population involve many examiners who must be consistent in scoring different levels of an oral disease. Prior to the oral health survey implementation, a measurement of inter-rater reliability (IRR) is needed to know the level of agreement among examiners or raters.* ***Purpose:*** *This study aimed to assess the IRR using consensus and consistency estimates in large population oral health surveys.* ***Methods:*** *A total of 58 dentists participated as raters. The benchmarker showed the clinical sample for dental caries and community periodontal index (CPI) score, with the raters being trained to carry out a calibration exercise in dental phantom. The consensus estimate was measured by means of a percent agreement and Cohen's Kappa statistic. The consistency estimate of IRR was measured by Cronbach's alpha coefficient and intraclass correlation.* ***Results:*** *The percent agreement is 65.50% for photographic slides of dental caries, 73.13% for photographic slides of CPI and 78.78% for calibration of dental caries using phantom. There were statistically significant differences between dental caries calibration using photographic slides and phantom (p<0.000), while the consistency of IRR between multiple raters is strong (Cronbrach's Alpha: >0.9).* ***Conclusion:*** *A percent agreement across multiple raters is acceptable for the diagnosis of dental caries. Consistency between multiple raters is reliable when diagnosing dental caries and CPI.*

Keywords: *inter-rater reliability; calibration; training; oral health survey*

Correspondence: Taufan Bramantoro, Department of Dental Public Health, Universitas Airlangga, Jl. Mayjend. Prof. Dr. Moestopo no. 47, Surabaya 60132, Indonesia. E-mail: taufan-b@fkg.unair.ac.id

INTRODUCTION

When an oral health survey is conducted on a large population, it might involve many team members as examiners. At times, these individuals score different levels of oral disease inconsistently. The question of consistency or agreement between examiners will arise due to variations in the diagnosis of oral disease between two or more such individuals or for the same examiner on more than one occasion. The other factor influencing consistency is the variability between examiners due to factors such as fatigue or differences in visual acuity and tactile sensation.

In order to diagnose oral disease consistently in oral health surveys, all examiners must have been trained in standardization and calibration. It is important to train examiners who will be involved in oral health surveys, especially for epidemiological studies based on World Health Organization (WHO) Basic Oral Health Survey Methods (2013).[1]

Oral health surveys are needed to plan and evaluate oral health programs and services, with control of the methodological biases in such surveys being exercised. According to WHO methodology, prior training and calibration of examiners represents the essential initial steps of oral health surveys. The calibration allows standardized interpretation of diagnostic criteria between examiners or raters. The general percentage agreement (GPA) and kappa statistics have been proposed for this task.[2]

The extent of agreement between examiners or raters is referred to as "inter-rater reliability (IRR)". IRR is, to a greater or lesser degree, a concern in most large-scale studies due to the fact that multiple individuals collecting

Dental Journal (Majalah Kedokteran Gigi) p-ISSN: 1978-3728; e-ISSN: 2442-9740. Accredited No. 32a/E/KPT/2017.

DOI: 10.20473/j.djmkg.v51.i1.p42–46

data may experience and interpret the phenomena of interest differently.[3,4] IRR refers to the level of agreement between a specific set of judges on a particular instrument at a particular time.[5,6]

Calibration is needed to ensure that all raters examine to the same standard. It is recommended that the training and calibration processes adhere to the methods propounded by the WHO Basic Oral Health Survey. The purpose of the training and calibration process is to minimize variation between examiners, to synchronize interpretation and to understand and apply the criteria for oral conditions such as dental caries that will be observed and recorded.[1]

The training stages consist of theoretical discussions, calibration exercises on dental phantom models and practical activities involving patient simulation. A benchmarker examiner or gold standard conducted the training with theoretical and practical activities. In theoretical activities, a benchmarker examiner explains WHO Basic Oral Health Survey Methods (2013), principles, code and criteria of dental caries and periodontal examination, the data collection procedure and data management. The study objective is to assess the IRR using consensus and consistency estimates in a large population oral health survey.

MATERIALS AND METHODS

A total of 58 dentists from Faculties of Dental Medicine throughout Indonesia participated in the training and calibration of an oral health survey. The training was held at the Faculty of Dental Medicine, Universitas Airlangga, in May 2017. A benchmarked examiner (gold standard) conducted the training program consisting of theoretical and practical activities. The examiner (gold standard) should meet the following requirements: he/she has followed WHO guideline-based training drawing on the oral health survey and passed with a kappa score of at least 0.8, holds a calibration and simulation trainer qualification, has calibration training instructor experience, and has participated in research on the WHO oral health survey. The training and calibration procedure were delivered at the Faculty of Dental Medicine, Universitas Airlangga, in conjunction with the Dental Public Health Association Meeting. The six trainers as benchmarker examiners were drawn from the Faculties of Dentistry of Universitas Padjadjaran, Universitas Jenderal Achmad Yani, Universitas Kristen Maranatha and Universitas Indonesia. The Kappa scores among the benchmaker examiners from all the faculties varied between 0.6 and 0.7.

First, after the theoretical session, the benchmarker presented the clinical sample of 25 photographic slides for each criterion of healthy and decaying teeth. The benchmarker also displayed 13 photographic slides showing the periodontal condition for each community periodontal index (CPI) score using CPI-modified scoring.

The second stage in the training consisted of a calibration exercise on a dental phantom head. A total of 36 healthy and decaying teeth were mounted in 36 plaster blocks for examination with a ball-ended probe in accordance with WHO criteria. All raters examined the clinical diagnosis of both healthy and decaying teeth and the assessment criteria**.**

The purpose of the first and seconds steps was to determine the inter-rater reliability based on the percent agreement across multiple raters, Cronbach's alpha and the intraclass correlation. To obtain the measure of percent agreement, a matrix in which the columns represented the different raters and the rows represented variables for which the raters had collected data was created. The cells in the matrix contained the rater scores entered for each variable. This technique allowed the researcher to identify variables that may be problematic.[3,7] Percent agreement is useful but, because it does not account for chance agreement, it should not be relied upon as the only measure of inter-rater consensus. In this study, intraclass correlation, as one of the most popular and consistent inter-rater reliability methods for numerous raters has been adopted.

The last step of the training, after parents/teachers had signed an informed consent form, was calibration by examining school children subjects with healthy and decayed teeth, which was granted ethical clearance by the Faculty of Dental Medicine, Universitas Airlangga. The simulation of examination activity involved six school children as standard patients. The participants examined the condition of dental caries. Otherwise, the CPI score examined through slides simulation. Before the raters examined the school children, a gold standard examination of the children's dental condition based on WHO Basic Oral Health Survey Methods was conducted. Each rater was helped during the study by a recorder, but did not discuss their findings with the gold standard.

The examination was carried out indoors, the school children lying on a chair or table with the examiner seated at their heads and the recorder sitting in front of the chair. The examination of dental caries was conducted using a dental mirror and a 0.5 mm diameter ball-ended probe.

Table 1. Calculation of the Kappa score for dental caries examination

Examiner 1	Examiner 2		Total
	Healthy	Decay	
Healthy	a	c	a + c
Decay	b	d	b + d
Total	a + b	c + d	a + b + c + d

a = proportion of teeth both examiners consider to be healthy, b = proportion of teeth examiner 1 considers to be decayed and examiner 2 consider to be healthy, c = proportion of teeth examiner 1 considers to be healthy and examiner 2 considers to be decayed, d = proportion of teeth both examiners consider to be decayed.

DOI: 10.20473/j.djmkg.v51.i1.p42–46

The results of such examinations carried out by raters were compared with those conducted by the gold standard. A more reliable means of assessing the overall agreement between examiners is the Kappa statistic which relates the actual measure of agreement with the degree of agreement which would have occurred by chance.[3,8,9] The Kappa score in examining for dental caries can be calculated using a 2 x 2 table[3,8] as seen in Table 1.

Kappa formula:

$$K = \frac{Pa–Pc}{1–Pc}$$

Pa = the percentage of assessments that are consistent across raters,

Pc = the percentage of assessments that vary between raters. This figure can be calculated using the following formula:

$$Pa = \frac{(a+d)}{(a+b+c+d)}$$

$$Pc = \frac{(a+c)\times(a+b)+(b+d)\times(c+d)}{(a+b+c+d)^2}$$

The Kappa score is interpreted as follows: <0.20 poor agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, 0.81-1.00 almost perfect agreement.

RESULTS

The IRR, based on a calculation of the percent agreement for each rater regarding the code of dental caries, CPI use of photographic slides and phantom, can be seen in Table 2 which shows the percent agreement between multiple raters.

Table 2 exhibits a percent agreemeent of 65.50% for photographic slide of dental caries, 73.13% for photographic slides of CPI and 78.78% for calibration of dental caries using a phantom head.

Table 3 shows the variables of photographic slides and phantom heads. Rater agreement for each variable can be seen which shows that the raters achieved 65.79% agreement for all variables of photographic slide of dental caries, 73.76% for CPI and 79.27% for dental caries variables using phantom heads. According to the contents of Table 3, the raters achieved between 25% and 98.21 % agreement for photographic slide of dental caries, between 10.71% and 100.00% in CPI and between 20.70% and 98.30% for dental caries using phantom heads.

Table 4 shows that statistically significant differences existed between dental caries calibration using photographic slides and phantom head (p<0.000). With reference to the Kolmogorov-Smirnov test, neither percent agreement was normally distributed, so that a differences test of non parametric tests was employed. Based on the results of a Wilcoxon test, the difference between the mean of percent agreement between photographic slides and phantom heads is very significant (p=0.000).

The method of calculating the percent agreement does not account for chance agreement. In this study, the inter-rater reliability was based on Cronbach's Alpha and an intraclass correlation method to analyze the consistency and agreement among raters

Based on the contents of Table 5, the Cronbrach's alpha score for all calibration methods was >0.9, indicating that the consistency of IRR between multiple rater was strong. This means that all raters were reliable in diagnosing dental caries using photographic slides and phantom heads. All raters were also reliable in determining the CPI code using photographic slides.

Table 2. The percent agreement between multiple raters

Calibration Method	n	Mean (%)	SD	Min (%)	Max (%)
Dental caries (slide)	56	65.50	73.29	12.00	84.00
CPI (slide)	56	73.13	10.16	46.00	92.00
Dental caries (phantom)	58	78.78	12.61	16.67	94.44

Table 3. The percent agreement among multiple raters for each variables

Item	Total Variables	Mean (%)	SD	Min (%)	Max (%)
Dental caries (slide)	25	65.79	22.25	25.00	98.21
CPI (slide)	13	73.76	26.09	10.71	100.00
Dental caries (phantom)	36	79.27	19.21	20.70	98.30

Table 4. Differences in percent agreement between photographic slide and phantom

	Mean (%)	SD	Z	p
Photographic slide	65.81	12.37	-5.635	0.000*
Phantom	80.66	8.5		
n=53				

* Significant

Table 5. Inter-rater reliability based on Cronbach's alpha

Calibration Method	n	Cronbrach alpha
Dental caries (slide)	56	0.942*
CPI (slide)	56	0.973*
Dental caries (phantom)	58	0.946*

*) p=0.000

Table 6. Consistency among raters based on intraclass correlation

Calibration method	n	Average measures	95% CI Lower bound	95% CI Upper Bound
Dental caries (slide)	56	0.942*	0.903	0.937
CPI (slide)	56	0.973*	0.946	0.990
Dental caries (phantom)	58	0.946*	0.917	0.968

*) p = 0.000

Table 7. Calculation of the Kappa score for dental caries examination

Examiner 1	Examiner 2 Healthy	Examiner 2 Decay	Total
Healthy	11	13	24
Decay	2	9	11
Total	13	22	35

Using the Kappa formula, a Pc score of 0.23 (fair agreement) is obtained

Table 6 shows the average score of the 56 raters using photographic slides to diagnose dental caries to be reliable (an interval of 0.903 to 0.937 with 95% confidence). The average score of the 56 raters using photographic slides to determine the CPI code was reliable (an interval of 0.946 to 0.9990 with 95% confidence). The average score of the 56 raters using phantom heads to diagnose dental caries was also reliable (an interval of 0.917 to 0.968 with 95% confidence). This suggests that, despite their apparent differences in diagnosis dental caries and determining CPI using various methods, the process was successful in training the examiner to determine the code of dental caries and CPI based on WHO Oral Health Survey methods.

In the final calibration session, the simulation of dental caries examination was carried out on the students. Table 7 shows the results of the examination simulation performed by the examiners on students. The Kappa score is 0.23 which, being below 0.4, represents fair agreement.

DISCUSSION

The limitation of information about the IRR among examiners in training and the calibration of oral health surveys based on WHO methods in Indonesia underlies this study. The data of IRR in this study was collected by several methods using the percent agreement, Cronbach's alpha, consistency using intraclass correlation and Kappa statistics.[10,11]

IRR constitutes the degree of agreement between raters. If raters agree, IRR is 1 (100%), whereas if they disagree, the IRR is 0 (0%). This produces a score of how much homogeneity or consensus exists between the ratings awarded by raters. Based on the percent agreement, the IRR in this study fell within the range of 60-90%. In general, above 75 % is considered acceptable for diagnosing dental caries using photographic slides and phantom heads and determining CPI using photographic slides.

There are some factors potentially influencing the low percent agreement between rater/examiners in this study. First, some of the examiners might not yet have been familiar with the code and criteria of dental caries and CPI based on WHO Basic Oral Health Survey Methods (2013). Second, the quality of photographic slides or phantom heads remains in question due to their unclear appearance.

The low percent agreement in diagnosing dental caries is found in photographic slides, but increases when the raters follow the calibration using phantom heads. Based on a Wilcoxon test, the difference in the mean of the percent agreement between photographic slides and phantom heads is very significant (p=0.000). It means that the perception and understanding of all raters about the code of dental caries based on WHO methods increased after they had followed the second training stage.

The most popular method for computing a consensus estimate of inter-rater reliability is through the use of the percent agreement between multiple raters. Percent agreement is easy to calculate and explain.[12] The calculation of percent agreement does not take chance agreement into account. That is one of the disadvantages of the percent agreement method. In this study, the Kappa statistic was used as the other method of IRR to determine the consensus or agreement between two raters.

Cohen's Kappa was designed to estimate the degree of consensus between two raters after correcting the percent agreement figure for the amount of agreement that could be expected due to chance alone based upon the values of the marginal distributions.[13] Kappa statistics are used for the assessment of agreement between two or more raters when the measurement scale is categorical. Kappa agreement is simply an adjusted form of percent agreement that takes chance agreement into account. A Kappa score is usually expressed as a proportion, rather than a percentage, and is not multiplied by 100 as with percent agreement. In this study, the kappa score for one sample falls within the fair agreement category.[4,5]

The factors that affect the fair agreement category in this sample of the study might be that the raters lack familiarity with the dental caries code based on the WHO method. The other factors might be related to the condition of mixed dentition which is confusing for raters to determine the coding for deciduous or primary teeth.

In this study, the consistency of raters is assessed by Cronbach's Alpha and intraclass correlation method. Cronbach's Alpha coefficient is a measure of internal consistency reliability and is useful for understanding the extent to which the ratings from a group of raters can be taken together to measure a common dimension.[13,14] The consistency between raters using various calibration methods in this study is rigorous.

DOI: 10.20473/j.djmkg.v51.i1.p42–46

Intraclass correlation is one of the most popular IRR methods to measure two or more raters.[14,15] The results of an intraclass correlation method of both dental caries and CPI are reliable, as can be seen from the average score of raters for all the calibration methods. The consistency of multiple raters using various calibration methods in this study was strong, when computed by both Cronbach's alpha and intraclass correlation. In conclusion, the percent agreement across multiple raters in this study are considered acceptable for diagnosing dental caries, but the agreement based on Kappa statistics must increase if raters, particularly those with low Kappa scores, follow the same training. The consistency between multiple raters using Cronbach's Alpha and intraclass correlation in this study was fair and reliable in diagnosing dental caries and CPI score in large population oral health surveys based on WHO Oral Health Survey Methods.

REFERENCES

1. World Health Organization. Oral health surveys : basic methods. 5th ed. France: World Health Organization; 2013. p. 25-7.
2. Tonello AS, Silva RP da, Assaf AV, Ambrosano GMB, Peres SH de CS, Pereira AC, Meneghim M de C. Interexaminer agreement dental caries epidemiological surveys: the importance of disease prevalence in the sample. Rev Bras Epidemiol. 2016; 19(2): 272–9.
3. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012; 22(3): 276–82.
4. Mandrekar JN. Measures of Interrater Agreement. J Thorac Oncol. 2011; 6: 6–7.
5. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Pract Assess Res Eval. 2004; 9(4): 1–11.
6. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Soc Adm Pharm. 2013; 9(3): 330–8.
7. Lebreton JM, Burgess JRD, Kaiser RB, Atchley EK, James LR. The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? Organ Res Methods. 2003; 6: 80–2.
8. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85(3): 257–68.
9. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010; 73(9): 1167–79.
10. Marusteri M, Bacarea V. Comparing groups for statistical differences: how to choose the right statistical test? Biochem Medica. 2010; 20: 15–32.
11. Pieper D, Jacobs A, Weikert B, Fishta A, Wegewitz U. Inter-rater reliability of AMSTAR is dependent on the pair of reviewers. BMC Med Res Methodol. 2017; 17: 98.
12. Stolarova M, Wolf C, Rinker T, Brielmann A. How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. Front Psychol. 2014; 5: 509.
13. McCrae RR, Kurtz JE, Yamagata S, Terracciano A. Internal consistency, retest reliability, and their implications for personality scale validity. Personal Soc Psychol Rev. 2011; 15: 28–50.
14. Vilella KD, Assunção LR da S, Junkes MC, de Menezes JVNB, Fraiz FC, Ferreira F de M. Training and calibration of interviewers for oral health literacy using the BREALD-30 in epidemiological studies. Braz Oral Res. 2016; 30: e90.
15. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012; 8: 23–34.