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

by Sri Susilawati

Submission date: 04-Apr-2018 01:53PM (UTC+0800)

Submission ID: 940852271

File name: Sri_Susilawati_EDIT_references.docx (52.44K)

Word count: 3358

Character count: 18238

Research Report

Building team agreement on large population survey 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 extent of agreement among examiners or raters.* ***Purpose****: This study aimed to assesses the IRR using consensus and consistency estimates on large population oral health survey.* ***Methods****: A total number of 58 dentist are participated as raters. The benchmarker showed the clinical sample for dental caries and community periodontal index (CPI) score and then the raters were trained to have calibration exercise in dental phantom. Consensus estimate was measured by using percent agreement and Cohen's Kappa statistic. Consistency estimate of IRR was measured by Cronbach's alpha coefficient and intraclass correlation.* ***Results****: The percent agrement of 65.50% for photographic slide of dental caries, 73.13% for photographic slide of CPI, and 78.78% for calibration of dental caries using phantom. There was statistical significant differences between dental caries calibration using photographic slide and phantom (p<0.000) and the consistency of IRR between multiple rater are strong (Cronbrach's Alpha: >0.9).* ***Conclusion:*** *The percent agreement across multiple raters are acceptable for diagnose dental caries. The consistency between multiple raters are reliable to diagnose 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 large population, it might be involve many team member as examiners. Sometimes, the examiner inconsistenly in scoring different level of an oral disease. The question of consistency or agreement among examiners will arises due to variation in oral disease diagnosis between two or more examiners and for the same examiner in two or more occasions. The other factor that influence of consistency is the variability among examiners such as as fatigue or differences in visual acuity and tactile sense.

In order to diagnose oral disease in oral health surveys consistenly, all of the examiner must be standardized and calibrated in training process. It is important to train examiner who will involve in oral health surveys especially for epidemiological studies based on WHO Basic Oral Health Surveys Method (2013).[1]

Oral health surveys are needed to plan and evaluate oral health actions and services. The control of the methodological biases in such surveys must be done. According the WHO methodology, previous training and calibration of the examiners are the initial and essential steps of oral health surveys. The calibration allows standardizing the interpretation of diagnostic criteria among examiners or raters. The general percentage agreement (GPA) and kappa statistics have been proposed for this task.[2]

The extent of agreement among examiners or rater is called "inter-rater reliability *(IRR)*". *IRR* is a concern to one degree or another in most large studies due to the fact that multiple people collecting data may experience and interpret the phenomena of interest differently.[3,4] *IRR* refers to the level of agreement between a particular set of judges on a particular instrument at a particular time.[5,6]

Calibration is needed to ensure that every raters examine to the equal standard. It is recomended that training and calibration process in accordance with the recommended methods for Basic Oral Health Survey of WHO. The purpose of training and calibration process is to minimize the variation between the examiners, to synchronize interpretation, understanding and application of the criteria for oral condition such as dental caries that will be seen and recorded.[1]

The steps of training are theoretical discussions, calibration exercise in dental phantom and practical activities in patient simulation. A benchmarker examiner or gold standard conducted the training process with theoretical and practical activities. In theoretical activities, a benchmarker examiner explain the principles of WHO Basic Oral Health Surveys Method (2013), code and criteria of dental caries and periodontal examination, procedure of data collection, and data management. The study objective is to assess the *IRR* using consensus and consistency estimates on large population oral health survey.

MATERIALS AND METHODS

A total number of 58 dentist from Faculty of Dental Medicine throughout Indonesia are participated in training and calibration of oral health survey. The training was held at Faculty of Dental Medicine, Universitas Airlangga in May 2017. A benchmarker examiner (gold standard) conducted the training processes with theoretical and practical acitivites. The examiner (gold standard) should qualify the requirements: has followed the oral health survey training based on the WHO guidelines and passed the training with a kappa score of at least 0.8, has certificate as calibration trainer and simulation and have experience as a calibration training instructor, and has been involved and have experience in research on oral health survey based on WHO. The training and calibration process have been held in Faculty of Dental Medicine, Universitas Airlangga in conjunction with Dental Public Health Association Meeting. The trainer as a benchmarker examiner was 6 people and comes from Faculty of Dentistry Universitas Padjadjaran, Universitas Jenderal Achmad Yani, Universitas Kristen Maranatha and Universitas Indonesia. The Kappa score among benchmaker examiner from all faculty varies between 0,6-0,7.

First, after theoritical session, the benchmarker showed the clinical sampel of 25 photographic slides for each criterion of healthy teeth and decay teeth. The benchmarker also showed periodontal condition of 13 photographic slide for each scoring of community periodontal index (CPI) using CPI-modified scoring.

The second steps of training is calibration exercise in dental phantom. A total of 36 healthy and decay teeth were mounted in 36 plaster blocks for examination with a ball ended probe using WHO criteria. All of raters examined the clinical diagnosis of healthy teeth, decay teeth and criteria.

The purpose of first and seconds steps are to determine the interrater realiability based on the percent agreement across multiple raters, Cronbach's alpha and the intraclass corelation. 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 could be created. The cells in the matrix contained the scores of the raters entered for each variable. This technique allows the researcher to identify variables that may be problematic.[3] Percentage agreement is useful, but because it does not account for chance agreement, it should not be used as the only measure of inter-rater consensus. In this study, intraclass correlation as one of the most popular inter-rater reliability of consistenly method for more raters has been used.

The last step of the training is calibration by examine a subject of school children with a healthy and decay teeth after the parents/teacher signed an informed consent form. The calibration in subject of school children has been through ethical clearance in Faculty of Dental Medicine, Universitas Airlangga. The number of school children examined as a standard patient is 6 students. In school children examined only dental caries score only while the CPI score examination is not done, the CPI score simulation is done only using slides. Before the raters examine the school children, a gold standard has been examine student's dentition condition based on WHO Basic Oral Health Surveys Method. Each raters was helped by a recorder during the study. During this phase, raters did not discuss their findings with gold standard.

The examination were carried out in an indoor setting. The students was lying on a chair or table with the examiner seated behind the student's head and the recorder seat in front of the chair. Dental caries examination is conducted using dental mirror and ball-ended probe with a diameter of 0.5 mm. The result of dental caries examination by raters will be compared with the result of dental caries examination by gold standard. A more reliable way of assessing overall agrement between examiner is the Kappa statistic. The Kappa statistic relates the actual measure of agreement with the degree of agreement which would have occured by chance.[3,7] The Kappa score can be calculated using a 2 x 2 table.[3] The calculation of the Kappa score in examining for dental caries can be seen on Table 1.

Kappa formula:

$$K = \frac{Pa-Pc}{1-Pc}$$

Pa = percentage of assessment that consistent across rater, Pc = percentage of assessment that changes between rater. It can be calculate with formula:

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

$$P_c = \frac{(a+c)x\,(a+b)+(b+d)x\,(c+d)}{(a+b+c+d)^2}$$

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 calculation of percent agrement for each rater about the code of dental caries, CPI use of photographic slides and phantom can be seen on Table 2. Table 2 shows the percent agreement among multiple raters.

Table 2 exhibits the percent agremeent of 65.50% for photographic slide of dental caries, 73.13% for photographic slide of CPI, and 78.78% for calibration of dental caries using phantom.

Table 3 shows variables of photographic slides and phantom. The rater's agreement for each variables can be seen on table 3. Table 3 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. Based on Table 3, the raters achieved 25% until 98.21 % agreement for photographic slide of dental caries, 10.71 % until 100.00% in CPI, and 20.70% until 98.30% for dental caries using phantom.

Table 4 shows us that there was statistical significant differences between dental caries calibration using photographic slide and phantom (p<0.000). Refer to Kolmogorov-Smirnov test, both of percent agremeent were not normal distribution, so that the differences test of non parametric test is used. Based on Wilcoxon test, the difference of mean of percent agremeent between photographic slide and phantom is very significant (p=0.000).

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

Based on Table 5, score of Cronbrach's alpha for all calibration method are >0.9, it means the consistency of IRR between multiple rater are strong. It means all raters reliable in diagnose dental caries using photographic slide and phantom. All raters also reliable in determine the code of CPI using photographic slide.

Table 6 shows the average of the score of the fifty-six raters using photographic slide to diagnose dental caries are reliable (interval of 0.903 to 0.937 with 95% confidence). The average of the score of the fifty-six raters using photographic slide to determine the code of CPI are reliable (interval of 0.946 to 0.9990 with 95% confidence). The average of the score of the fifty-eight raters using phantom to diagnose dental caries are also reliable (interval 0.917 to 0.968 with 95% confidence). Suggesting that despite their apparent differences in diagnosis dental caries and determine CPI using various method, the process was successful in traning the examiner to determine the code of dental caries and CPI based on WHO Oral Health Surveys Method.

In the last session of calibration, the simulation of dental caries examination on student has been done.

In Table 7 shows the result of examination simulation by examiner on student, the Kappa score is 0.23 which is below 0.4 represent fair agreement.

DISCUSSION

The limitation of information about the IRR among examiners in training and calibration of oral health surveys based on WHO method in Indonesia is underlie of this study. The data of IRR in this study is collected by several method using the percentage agreement, Cronbach's alpha, consistency using intraclass correlation and Kappa statistic.[8]

IRR is the degree of agreement between raters. If raters agreed, IRR is 1 (100%) and if raters disagreed, IRR is 0 (0%). It gives a score of how much homogeneity, or consensus, there is in the ratings given by raters. Based on percent agrement, the IRR in this study are in range 60-90%. In general, above 75 % is consider acceptable for diagnose dental caries using photographic slide and phantom and determine CPI using photographic slide.

There are some factor that might be influence the low of percent agreement between rater/examiners in this study. First, some of the examiner still not yet familiar with code and criteria of dental caries and CPI based on WHO Basic Oral Health Surveys Method (2013). Second, the quality of photographic slide or phantom are not yet optimal due to unclearly apperance.

The low of percent agreement in diagnose dental caries is found in photographic slide, but the percent agreement are increased when the raters follow the calibration using phantom. Based on Wilcoxon test, the difference of mean of percent agreement between photographic slide and phantom is very significant (p=0.000). It means that perception and understanding of all of the raters about the code of dental caries based on WHO method are increased after the raters follow the second steps of training.

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 easy to explain.[9] The calculation of percent agreement does not take chance agreement into account. That is one of the disadavantage of the percent agreement method. In this study, we use the Kappa statistic as the other method of inter-rater reliability to determine the consensus or agreement among 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 by chance alone based upon the values of the marginal distributions.[10] Kappa statistics is used for the assessment of agreement between two or more raters when the measurement scale is categorical. Kappa agreement is simply adjusted form of percentage agreement that does take into account chance agreement. Kappa is usually expressed as a proportion rather than a percentage, so we don't multiply by 100 as with percent agreement. In this study, the kappa score for one sample is in fair agreement category.[11]

The factors that affect the fair agreement category in this sample of the study are might be the raters is not yet familiar with the code of dental caries using WHO method. The other factors might be related with the condition of mix dentition which is confusing for the raters to be determine coding for deciduos or primary teeth.

In this study, the consistency of raters are 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 hold together to measure a common dimension.[12] The consistencies among raters using various calibration methods are strong in this study.

Intraclass correlation is one of the most popular inter-rater reliability method to measure two or more raters.[13] The result of intraclass correlation method both of dental caries and CPI are realible, it can be seen from the average of the score of raters for all the calibration method. The consistency of multiple raters using various calibration method in this study are strong, both computed by Cronbach's alpha and intraclass correlation. In conclusion, the percent agreement across multiple raters in this study are consider acceptable for diagnose dental caries, but the agreement based on Kappa statistic must be increase with follow the same training especially for the raters with lower of Kappa score. The consistency between multiple raters using Cronbach's Alpha and intraclass correlation in this study was fair agreement and reliable to diagnose dental caries and CPI score in large population oral health survey based on WHO Oral Health Survey Method.

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. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85(3): 257–68.
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. 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.
7. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010; 73(9): 1167–79.
8. 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.
9. 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.
10. 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.
11. Mandrekar JN. Measures of Interrater Agreement. J Thorac Oncol. 2011; 6: 6–7.
12. 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.
13. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012; 8: 23–34.

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

Examiner 2	Examiner 1		
	Healthy	Decay	Total
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 healthy and examiner 2 consider to be decay, c = proportion of teeth examiner 1 considers to be decay and examiner 2 considers to be healthy, d = proportion of teeth both examiners consider to be decay.

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 of percent agreement using 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. Consitency among raters based on intraclass correlation

			95% CI	
Calibration method	n	Average measures	Lower bound	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	Decay	Total
Healthy	11	13	24
Decay	2	9	11
Total	13	21	35

with Kappa formula, we obtain the Pc score is 0.23 (fair agreement)

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

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

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

ORIGINALITY REPORT

20%	16%	10%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pareonline.net Internet Source	4%
2	biochemia-medica.com Internet Source	3%
3	locus.ufv.br Internet Source	3%
4	Assaf, AndrÃ©a Videira, Elaine Pereira da Silva Tagliaferro, Marcelo de Castro Meneghim, Cristiana Tengan, Antonio Carlos Pereira, GlÃ¡ucia Maria Bovi Ambrosano, and FÃ¡bio Luiz Mialhe. "A new approach for interexaminer reliability data analysis on dental caries calibration", Journal of Applied Oral Science, 2007. Publication	2%
5	aperito.org Internet Source	1%
6	en.wikipedia.org Internet Source	1%

7	Submitted to iGroup Student Paper	1%
8	linknovate.com Internet Source	1%
9	www.science.gov Internet Source	1%
10	www.scribd.com Internet Source	1%
11	Submitted to Savitribai Phule Pune University Student Paper	<1%
12	Submitted to Laureate Higher Education Group Student Paper	<1%
13	Mohsen Anvaari, Carl-Fredrik Sørensen, Olaf Zimmermann. "Associating architectural issues with quality attributes", Proccedings of the 10th European Conference on Software Architecture Workshops - ECSAW '16, 2016 Publication	<1%
14	journal.unair.ac.id Internet Source	<1%
15	www.lifescied.org Internet Source	<1%
16	www.lppm.its.ac.id Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		