Original Research
Community Dental Health

# Training and calibration of interviewers for oral health literacy using the BREALD-30 in epidemiological studies

**Karina Duarte VILELLA**[(a)]
**Luciana Reichert da Silva ASSUNÇÃO**[(a)]
**Mônica Carmem JUNKES**[(a)]
**José Vitor Nogara Borges de MENEZES**[(a)]
**Fabian Calixto FRAIZ**[(a)]
**Fernanda de Morais FERREIRA**[(b)]

[(a)]Universidade Federal do Paraná (UFPR), Department of Stomatology, Graduate Program in Dentistry, Curitiba, PR, Brazil.

[(b)]Universidade Federal de Minas Gerais (UFMG), Department of Pediatric Dentistry and Orthodontics, Graduate Program in Dentistry, Belo Horizonte, MG, Brazil.

**Abstract:** The objective of this study was to describe an interviewer training and calibration method to evaluate oral health literacy using the Brazilian Rapid Estimate of Adult Literacy in Dentistry (BREALD-30) in epidemiological studies. An experienced researcher (gold standard) conducted all training sessions. The interviewer training and calibration sessions included three different phases: theoretical training, practical training, and calibration. In the calibration phase, six interviewers (dentists) independently assessed 15 videos of individuals who had different levels of oral health literacy. Accuracy and reproducibility were evaluated using the kappa coefficient and the intraclass correlation coefficient (ICC). The percentage of agreement for each word in the instrument was also calculated. After training, the kappa values were higher than 0.911 and 0.893 for intra- and inter-rater agreement, respectively. When the results were analyzed separately for the different levels of literacy, the lowest agreement rate was found when evaluating the videos of individuals with low literacy (K = 0.871), but still within the range considered to be near-perfect agreement. The ICC values were higher than 0.990 and 0.975 for intra- and inter-rater agreement, respectively. The lowest percentage of agreement was 86.6% for the word "hipoplasia" (hypoplasia). This interviewer training and calibration method proved to be feasible and effective. Therefore, it can be used as a methodological tool in studies assessing oral health literacy using the BREALD-30.

**Keywords:** Health Literacy; Epidemiologic Studies; Reproducibility of Results.

**Declaration of Interests:** The authors certify that they have no commercial or associative interest that represents a conflict of interest in connection with the manuscript.

**Corresponding Author:**
Luciana Reichert da Silva Assunção
E-mail: lurassuncao@yahoo.com.br

**DOI:** 10.1590/1807-3107BOR-2016.vol30.0090

Submitted: Dec 16, 2015
Accepted for publication: Apr 26, 2016
Last revision: May 13, 2016

## Introduction

Health literacy is an individual's ability to obtain, process, and understand basic health and service information, thus supporting appropriate health decisions.[1] This is a broader concept than merely the number of years of formal education and it can be used to develop teaching strategies that are more appropriate to the needs of a specific population.

When this concept is related to dentistry, it has been called oral health literacy. Studies have demonstrated an association of low oral health literacy with worse oral health status[2] and worse outcomes in oral health, such as temporomandibular disorders, prosthetic needs, and periodontal problems.[3] An observational cohort study showed that, in a group of low-income pregnant patients, mother's with low oral health literacy levels can affect health outcomes of both mother and child.[4] Low oral health literacy levels of caregivers were associated with children's worse oral health-related quality of life[5] and failure to show up for dental appointments.[6]

The Rapid Estimate of Adult Literacy in Dentistry (REALD-30) is one of the main instruments used to measure oral health literacy. It evaluates an individual's reading ability based on word recognition.[7] This is the only instrument to measure oral health literacy that has been translated, cross-culturally adapted, and validated into Brazilian Portuguese (BREALD-30).[8]

The BREALD-30 has adequate psychometric properties and proved to be reliable to measure the levels of oral health literacy in Brazilian Portuguese-speaking adults.[8] This instrument consists of 30 words related to oral health arranged in increasing order of reading difficulty. The BREALD-30 is aimed at providing a rapid estimate in adults and shows a positive and significant correlation with the instruments used to evaluate functional literacy, educational level, and oral health knowledge.[8]

During the administration of the BREALD-30, the interviewee reads each word aloud for the examiner. The words are read only once. Although this instrument is easily administered, its limitation is the great difficulty of calibrating the interviewers because it is expected that the examiners are able to recognize possible pronunciation failures and successes based on a single reading attempt. This is an important aspect, especially in epidemiological studies involving different interviewers. Therefore, it is extremely important to minimize potential measurement variations during data collection by standardizing data collection measurements. This can be achieved through a judicious and well conducted process of training and calibration.[9]

In this sense, the objective of this study was to describe an interviewer training and calibration method to evaluate oral health literacy using the BREALD-30. This is an attempt to develop a feasible calibration process that respects the characteristics of the instrument.

## Methodology

The interviewers training and calibration sessions for evaluation of oral health literacy were held in a classroom at the Universidade Federal do Paraná in November 2014. Six interviewers participated in the sessions. All of them were dentists with no previous experience in literacy research. The coordinator of the BREALD-30 validation study (the gold standard researcher) was responsible for the entire training and calibration process.

### BREALD-30 criteria

The BREALD-30 consists of 30 words related to oral diseases (etiology, anatomy, prevention, and treatment) arranged in increasing order of reading difficulty. The participant is supposed to read these words aloud for the interviewer. Each time the participant pronounces the word correctly, 1 point is assigned; whereas each time the participant fails to read the word correctly, no point is assigned. In the end, the scores for each word are summed up, and the BREALD-30 total score may range from 0 to 30. The higher the score the higher the level of oral health literacy. Some criteria are used to recognize pronunciation errors, namely: a) replacement with similar word (*e.g.*: *"escova"* [tooth brush] instead of *"escovar"* [to brush]); b) irregular words read as regular words (*e.g.*: *"ensaguatório"* instead of *"enxaguatório"* [mouth wash]); c) replacing, omitting, or adding letters (*e.g.*: *"gengiba"* instead of *"gengiva"* [gum]); d) failure to use matching rules (*e.g.*: *"erossão"* instead of *"erosão"* [erosion]); e) failure to recognize the stressed syllable (*e.g.*: *"genetica"* instead of *"genética"* [genetics]). Situations when the participants read the words slowly and without rhythm or when it is necessary to repeat the word or any syllables are also considered pronunciation errors.

### Calibration process

For the calibration process, videos of individuals who had different levels of oral health literacy were selected from a pull of videos of the administration of the BREALD-30. With this purpose, we recruited around 200 adult literate participants whose native language was Brazilian Portuguese and who agreed to be filmed while their oral health literacy was being assessed. The exclusion criteria were: age older than 80 years, any reported or perceived vision or hearing problems, obvious signs of cognitive impairment and/or intoxication with alcohol and drugs at the time of the interview. All participants signed an informed consent form before the word reading session was recorded. The videos were classified according to the BREALD-30 scores as low literacy (≤ 21 points), moderate literacy (22–25 points), and high literacy (≥ 26 points).[10]

The training and calibration of interviewers involved the following phases: training (theoretical and practical) and calibration of interviewers.

### Training of interviewers

First, a theoretical training was provided. It consisted of the presentation of the instrument to the interviewers and the explanation of the criteria for recognition of pronunciation errors. This was a 4-hour phase.

The practical training consisted of watching 10 videos of individuals with varying levels of oral health literacy. The videos were shown only once at the same time to the six interviewers and the gold standard researcher. The interviewers assigned scores to the individuals' performance based on the videos using the BREALD-30, and their results were analyzed and discussed after being compared to the scores assigned by the gold standard researcher. This was a 4-hour phase.

### Calibration of interviewers

For the calibration phase, 15 videos of individuals with varying levels of functional oral health literacy were shown to the interviewers. One of these videos showed an individual with high literacy, two videos showed individuals with moderate literacy, and 12 videos featured individuals with low literacy (showing more pronunciation errors, thus generating more doubts in terms of classification). The examiners evaluated the videos individually and repeated the evaluation after one week in order to analyze the intra-rater agreement. At this phase, the interviewers could not talk to each other. Each session was 2 hours long.

### Analysis of results

The scores assigned during the interviewer calibration phase were compared to the scores assigned by the gold standard researcher and then discussed by the members of the research team.

The results were statistically analyzed using three different methods. We used the *kappa* coefficient to calculate the intra- and inter-rater agreement based on each word of the BREALD-30. We considered that the words could be classified as correctly pronounced and incorrectly pronounced and, therefore, are dichotomous variables. This index was also used to evaluate the intra- and inter-rater agreement according to the different levels of literacy, because the assessment was performed on a word-to-word basis as well. The evaluation of the intra- and inter-rater agreement based on the BREALD-30 total score assigned to each video was performed using the ICC, as this is a numeric variable. Total score was the sum of the points assigned to each word in the instrument (0 or 1). Inter- and intra-rater agreement rates higher than 0.85 were considered acceptable levels of agreement[11] (according to the literature, *kappa* coefficient > 0.80 is considered to be near-perfect agreement[12] and ICC > 0.75 is an excellent agreement[13]).

Furthermore, the percentage of agreement for each word of the BREALD-30 was calculated considering all interviewers in relation to the gold standard researcher. The reliability parameters used for assessing the percentage of agreement followed the WHO criteria,[14] where values above 85% indicate acceptable agreement.

All statistical analyses were conducted using the Statistical Package for the Social Sciences (SPSS™ for Windows™, version 20.0, SPSS Inc., Chicago, USA).

### Ethical aspects

This study was approved by the Research Ethics Committee of the Universidade Federal do Paraná (protocol no. 0171.0.091.000-11). All videos were recorded with the consent of the participants after they signed the informed consent form.

## Results

The *kappa* intra-rater agreement rates ranged from 0.911 to 0.938, whereas the inter-rater agreement rates varied from 0.893 to 0.920. The ICC intra-rater agreement rates ranged from 0.990 to 0.993, whereas the inter-rater agreement rates varied from 0.975 to 0.991.

The *kappa* intra-rater agreement rates based on the levels of oral health literacy were 1.00 for the video showing the individual with high literacy; between 0.915 and 1.00 for the videos showing individuals with moderate literacy; and between 0.893 and 0.921 for the videos of individuals with low literacy (Table 1). Based on the same parameter, the inter-rater agreement rates calculated using the *kappa* coefficient were 1.00 for the video of the individual with high literacy; between 0.915 and 0.957 for the videos showing individuals with moderate literacy; and 0.871 to 0.904 for those with low literacy (Table 1).

**Table 1.** Intra- and inter-rater *kappa* coefficient based on each word of the BREALD-30 according to the different levels of oral health literacy.

| Literacy interviewer | Kappa intra | | | Kappa inter | | |
|---|---|---|---|---|---|---|
| | High | Moderate | Low | High | Moderate | Low |
| A | 1.00 | 1.00 | 0.905 | 1.00 | 0.957 | 0.871 |
| B | 1.00 | 1.00 | 0.921 | 1.00 | 0.955 | 0.904 |
| C | 1.00 | 1.00 | 0.905 | 1.00 | 0.955 | 0.883 |
| D | 1.00 | 0.958 | 0.899 | 1.00 | 0.915 | 0.899 |
| E | 1.00 | 0.915 | 0.911 | 1.00 | 0.915 | 0.877 |
| F | 1.00 | 0.957 | 0.893 | 1.00 | 0.957 | 0.882 |

The results of the agreement rates for each word of the BREALD-30 considering all interviewers in comparison with the gold standard researcher are shown in Table 2. Eight words reached 100% agreement. The words showing the lowest agreement rates in descending order were "*restauração*" (restoration) and "*dentição*" (dentition), both with 88% and "*hipoplasia*" (hypoplasia) with 86.6% of agreement.

## Discussion

Studies involving a larger number of examiners or interviewers require training to make sure that the variability of results is as low as possible, resulting in greater data reliability. Considering the characteristics of the BREALD-30 and the inherent variability of each interviewer, achieving good reproducibility of observations is an essential condition so that the results are reliable. Our study demonstrated that this method can be safely used in studies using the BREALD-30 because it showed excellent results according to the three statistical methods used.

In order to achieve valid measurements with good reproducibility, clear definitions of the event to be evaluated are required, including measurement

**Table 2.** Percentage of agreement for each word of the BREALD-30 considering the total number of interviewers in relation to the gold standard researcher.

| Word | Agreement (%) | Word | Agreement (%) | Word | Agreement (%) |
|---|---|---|---|---|---|
| *Açúcar* (sugar) | 100.0 | *Biópsia* (biopsy) | 100.0 | *Endodontia* (endodontics) | 92.2 |
| *Dentadura* (denture) | 94.4 | *Enxaguatório* (mouthrinse) | 100.0 | *Maloclusão* (malocclusion) | 90.0 |
| *Fumante* (smoker) | 100.0 | *Bruxismo* (bruxism) | 100.0 | *Abcesso* (abscess) | 96.6 |
| *Esmalte* (enamel) | 91.1 | *Escovar* (to brush) | 100.0 | *Biofilme* (biofilm) | 98.8 |
| *Dentição* (dentition) | 88.8 | *Hemorragia* (bleeding) | 94.4 | *Fístula* (fistula) | 93.3 |
| *Erosão* (erosion) | 96.6 | *Radiografia* (radiograph) | 97.7 | *Hiperemia* (redness) | 92.2 |
| *Genética* (genetics) | 100.0 | *Película* (film) | 90.0 | *Ortodontia* (orthodontics) | 90.0 |
| *Incipiente* (incipient) | 97.7 | *Halitose* (halitosis) | 96.6 | *Temporomandibular* (temporomandibular) | 96.6 |
| *Gengiva* (gum) | 90.0 | *Periodontal* (periodontal) | 92.2 | *Hipoplasia* (hypoplasia) | 86.6 |
| *Restauração* (restoration) | 88.8 | *Analgesia* (analgesia) | 100.0 | *Apicectomia* (apicoectomy) | 90.0 |

and classification standards.[15] Accordingly, the first phase of the calibration and training process of interviewers is a fundamental step so that this process can be successful. In terms of the use of the BREALD-30, this training allowed for a better standardization of the evaluations according to the criteria to be considered, such as pronunciation errors during the instrument administration. In addition, it is worth mentioning that the proposed method included only one training session and this was sufficient to achieve great agreement results. Although we did not aim to establish the optimal number of interviewers, the inclusion of a larger number of participants should not be seen as a limitation of the method.

According to the World Health Organization,[14] the process of calibration of examiners in epidemiological surveys allows each examiner to provide consistent evaluation and reduces the variability between examiners. The use of a gold standard researcher provides the calibration process with greater validity,[16] and consists of one of the major goals of this method.[14] Although the gold standard examiner is also subject to misclassification,[17] he/she is considered "error-free" and regarded as a benchmark that reflects the truth.[18]

The *kappa* index is considered to be the measure of choice for the calculation of intra- and inter-rater reproducibility because it provides more conclusive and higher quality information.[9] The lowest *kappa* values found for intra- and inter-rater agreement were 0.911 and 0.893, respectively. These values indicate the effectiveness of the proposed interviewer training and calibration method in studies using the BREALD-30. The near-perfect agreement rates[12] were higher than 0.85, which is considered acceptable by the WHO in calibration processes.[11]

The ICC is a better option to measure the reliability when a numerical measurement is used for assessment.[19] This coefficient (ICC) is considered a measure of reliability, correlation, and consistency.[20] In the present study, the minimum ICC values for intra-rater agreement of 0.990 and inter-rater agreement of 0.975 are excellent.[13] Although the more heterogeneous the individuals' measures the higher the ICC, this characteristic has been seen as an advantage because it may reduce the disagreement in relation to the magnitude of the measure.[21]

The intra rater agreement is necessary because of the subjectivity of some evaluation methods and because of the examiner's specific conditions, such as fatigue, which may lead to inconsistent results.[11] Considering the BREALD-30, it is important to make sure that the criteria for mispronounced words are fully understood by the interviewer so that there are standardized evaluations at different times and situations. The results of the proposed method showed intra-rater agreement rates that are considered near-perfect for the *kappa* coefficient[12] and excellent for the ICC.[13]

The results of the percentage of agreement showed that eight out of the 30 words in the BREALD-30 were assigned the same score by the interviewers and the gold standard researcher in all evaluations (100%). The lowest rate of percentage of agreement was found in the word "*hipoplasia*" (86.6%). The words of the BREALD-30 are arranged in increasing order of reading difficulty based on the average word length, number of syllables, and difficulty of combining sounds.[8] Not surprisingly, therefore, the lowest agreement rates were associated with the last words of the instrument, such as "*hipoplasia*". For this particular case, there was probably more frequent mispronunciation and, consequently, the interviewers were more insecure about the score assignment. However, the lowest agreement rate found in the present study is higher than the rate expected as acceptable[14] and certainly does not cause any harm to the proposed method.

Watching videos as a methodological proposal in a calibration process involving the application of the BREALD-30 is extremely important and ensures the standardization of the interviewers' conditions of evaluation. Calibration based on direct observation could have operational difficulties because repeated readings of the instrument would be required, with obvious change in the individual's reading ability. Furthermore, the simultaneous evaluation by all interviewers of a single reading session could embarrass the individuals being tested. The use of videos eliminates this problem.

Also, the method suggests the use of videos with individuals who have different levels of literacy, including a larger number of individuals with low literacy. This proportion ensures greater validity of the method because pronunciation errors are more frequent in individuals with lower levels of literacy, which, in theory, could result in greater difficulty for the interviewers to assign scores. However, when the intra- and inter-rater agreement rates were assessed according to the different levels of literacy, the results of the *kappa* coefficient were higher than or equal to 0.871, and this agreement is "near-perfect"[12] even among the videos of individuals with lower levels of literacy.

The limitations of this study should be addressed. One of the limitations is that the respondents were selected from a sample of patients seen at an oral health center. Therefore, they are more likely to be familiar with the terms of the instrument. Future studies should be carried out with a more heterogeneous group of respondents. Another limitation is the number of interviewers. Although this study showed minimal and acceptable variations on the study group, it was not possible to define the maximum number of interviewers that can be trained and calibrated simultaneously. Based on the conclusions of the present study, further studies should be designed with that goal.

## Conclusion

The consistency of our results shows that the proposed method for training and calibration of interviewers is feasible and effective; therefore, it can be used as a methodological tool in studies aimed at evaluating oral health literacy using the BREALD-30.

## Acknowledgements

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES) for the financial support.

## References

1. American Medical Association. Health literacy: report of Council on Scientific Affairs. JAMA. 1999;281(6):552-7. doi:10.1001/jama.281.6.552
2. Lee JY, Divaris K, Baker AD, Rozier RG, Vann JF Jr. The relationship of oral health literacy and self-efficacy with oral health status and dental neglect. Am J Public Health. 2012;102(5):923-9. doi:10.2105/AJPH.2011.300291
3. Haridas R S, Ajagannanavar SL, Tikare S, Maliyil MJ, Kalappa AA. Oral health literacy and oral health status among adults attending dental college hospital in India. J Int Oral Health. 2014;6(6):61-6.
4. Hom JM, Lee JY, Divaris K, Baker AD, Vann WF Jr. Oral health literacy and knowledge among patients who are pregnant for the first time. J Am Dent Assoc. 2012;143(9):972-80. doi:10.14219/jada.archive.2012.0322
5. Divaris K, Lee JY, Baker AD, Vann WF Jr. Caregivers' oral health literacy and their young children's oral health-related quality-of-life. Acta Odontol Scand. 2015;70(5): 390-7. doi:10.3109/00016357.2011.629627
6. Hotzman JS, Atchison KA, Gironda MW, Radbod R, Gornbein J. The association between oral health literacy and failed appointments in adults attending a university-based general dental clinic. Community Dent Oral Epidemiol. 2014;42(3):263-70. doi:10.1111/cdoe.12089
7. Lee JY, Rozier RG, Lee SY, Bender D, Ruiz RE. Development of a word recognition instrument to test health literacy in dentistry: the REALD-30: a brief communication. J Public Health Dent. 2007;67(2):94-8. doi:10.1111/j.1752-7325.2007.00021.x
8. Junkes MC, Fraiz FC, Sardenberg F, Lee JY, Paiva SM, Ferreira FM. Validity and reliability of the Brazilian version of the Rapid Estimate of Adult Literacy in Dentistry – BREALD-30. PLoS One. 2015;10(7):e0131600. doi:10.1371/journal.pone.0131600
9. Assaf AV, Zanin L, Meneghim MC, Pereira AC, Ambrosano GMB. [Comparison of reproducibility measurements for calibration of dental caries epidemiological studies]. Cad Saúde Pública. 2006;22(9):1901-7. Portuguese. doi:10.1590/S0102-311X2006000900021
10. Wehmeyer MM, Corwin CL, Guthmiller JM, Lee JY. The impact of oral health literacy on periodontal health status. J Public Health Dent. 2014;74(1):80-7. doi:10.1111/j.1752-7325.2012.00375.x
11. World Health Organization. Oral health surveys: basic methods. Geneva: World Health Organization; 1997.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74. doi:10.2307/2529310
13. Oremus M, Oremus C, Hall GB, McKinnon MC. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. BMJ Open. 2012;2(4):e001368. doi:10.1136/bmjopen-2012-001368

14. World Health Organization. Calibration of examiners for oral health epidemiological surveys. Geneva: World Health Organization; 1993.
15. Peres MA, Traebert J, Marcenes W. [Calibration of examiners for dental caries epidemiologic studies]. Cad Saúde Pública. 2001;17(1):153-9. Portuguese. doi:10.1590/S0102-311X2001000100016
16. Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. J Dent Educ. 2006;70(4):428-33.
17. Klein CH, Costa EA. Os erros de classificação e os resultados de estudos epidemiológicos. Cad Saúde Pública. 1987;3(3):236-49. doi:10.1590/S0102-311X1987000300003
18. Agbaje JO, Mutsvari T, Lesaffre E, Declerck D. Measurement, analysis and interpretation of examiner reliability in caries experience surveys: some methodological thoughts. Clin Oral Investig. 2012;16(1):117-27. doi:10.1007/s00784-010-0475-x
19. Fleiss JL, Chilton NW, Park MH. Inter- and intra-examiner variability in scoring supragingival plaque: II. Statistical analysis. Pharmacol Ther Dent. 1980;5(1-2):5-9.
20. Dobbyn LM, Weir JT, Macfarlane TV, Mossey PA. Calibration of the modified Huddart and Bodenham scoring system against the GOSLON/5-year-olds' index for unilateral cleft lip and palate. Eur J Orthod. 2012;34(6):762-7. doi:10.1093/ejo/cjr092
21. Streiner DL, Norman GR. Health measurements scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.