Oral Health Literacy—Measurement Instruments and their Psychometric Properties: A Systematic Review

Edwin Ramos-Pilco, MS*; Marco Antonio Sánchez-Tito, MS†;
Lidia Yileng Tay, PhD‡

Objective: This article aims to provide an evaluation of the psychometric properties of the instruments of oral health literacy in adults.

Methods: An electronic search for instrument studies was performed in the PubMed, PubMed Central, ScienceDirect, Scopus, EMBASE, and PsycINFO databases to find articles published up to June 2021. The risk of bias of the included studies was assessed using the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) Risk of Bias checklist for systematic review.

Results: From an initial sample of 2617 articles, 14 instrument studies were included in the present review. Their sample sizes ranged from 177 to 1405 adults, and the number of items per measurement instrument ranged from 14 to 99. For structural validity, statistical techniques were performed using the classical test theory (exploratory and confirmatory factor analysis) and the item response theory (dichotomous and polytomous models). The Rapid Estimate of Adult Literacy in Dentistry 30, elaborated in the USA, was the measurement instrument that had the greatest number of cultural adaptations, having been validated in such countries as Saudi Arabia, Brazil, Turkey, and Romania. The evaluation of the risk of bias, undertaken using the COSMIN Risk of Bias checklist, showed that 6 of the 10 parameters had been evaluated.

Conclusion: The psychometric properties that were evaluated in the present systematic review were structural validity, internal consistency, reliability (test–retest), and hypothesis testing for construct validity. To date, there is no gold standard measuring instrument to evaluate the criterion validity parameter. [*P R Health Sci J 2023;42(3):187-196*]

Key words: Oral health literacy, Psychometric properties, Validity, COSMIN

Oral diseases (dental caries, periodontal disease, tooth loss, and mouth cancer) are among the main causes of morbidity, worldwide; they have serious health and economic consequences and contribute considerably to reducing the quality of life of those affected by them (1).

Oral health literacy (OHL) has been defined as "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate oral health decisions" (2). Hom et al. (2012) reported that low levels of OHL were related to poor knowledge about the oral health environment (3). Moreover, OHL exists in the contexts of culture/society, educational systems, and individual's interactions whit the public and/or private health systems. It results in costs and achievements in oral health (2).

An instrument that measure OHL could have many practical uses; for example, to screen for individual dental health literacy in clinical settings and to improve the communication between dental health care providers and their patients (2,3). Further, researchers and public health practitioners could use such an instrument to assess the levels of dental health literacy in a group of patients or a community and design interventions to effectively improve oral health and quality of life (1,2,4,5).

Measurement instruments must meet strict scientific standards of quality: The tests cannot make decisions on their own; they are made by health professionals, based on the data obtained by this or another procedure. Therefore, a rigorous evaluation is the basis of an accurate diagnosis, allowing an effective intervention based on empirical evidence; otherwise, there is a serious risk of bias of the results, which could lead to erroneous conclusions (6).

*National University Jorge Basadre Grohmann, School of Dentistry, Tacna, Peru; †Private University of Tacna, School of Dentistry, Tacna, Peru; ‡Peruvian University Cayetano Heredia, Faculty of Stomatology, Lima, Peru

The authors have no conflicts of interest to disclose.

Address correspondence to: Ramos Pilco, Edwin Pascual, MS, Tacna-Perú. Email: edy4208@gmail.com

The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) Risk of Bias checklist was developed exclusively for use in systematic reviews (6,7). This tool was chosen in the present study with the objective of evaluating the methodological quality of the psychometric properties of selected instruments that measure OHL in adults.

Methods

The present review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (8).

Focused question

Which psychometric properties for measuring adult OHL were evaluated by the instruments? This question guided the search strategy.

Eligibility criteria

• Population: Adults over 18 years of age without physical disabilities.

• Intervention: An evaluation of the methodological quality of psychometric properties using the COSMIN Risk of Bias checklist (6).

• Outcomes: The psychometric properties of the measurement instruments reported in the COSMIN Risk of Bias checklist (6).

• Study design: Instrument studies. The development of instruments or tests as well as their psychometric properties (9).

Exclusion criteria

• Studies that did not evaluate the evidence of validity in the internal structure of the test (structural validity).

• Studies that used a sample having fewer than 100 subjects.

• Multiple validation studies conducted by the same (main) author and on the same population.

Search strategy

A systematic review was carried out in the PubMed Central, PubMed, Science Direct, EMBASE, Scopus, and PsycINFO databases. There was no limitation on the initial date of publication, and studies published up to June 2021 were considered. The following search strategy was performed in PubMed using MeSH in combination with the following keywords: ((oral health literacy) OR (health literacy dentistry)) AND ((((validity) OR (scale development)) OR (reliability)) OR (psychometrics properties)) AND (Adult). This search strategy was adapted to the other databases. Also, a complementary exploration of the System for Information on Grey Literature in Europe (http://www.opengrey.eu/) was completed.

Screening, data extraction

The selection of articles by titles, abstracts, and full texts was carried out independently by 2 reviewers (E.R. and M.S.), according to the selection criteria. In case of disagreement, a third reviewer was consulted (Y.T.) and the process of evaluation continued until a consensus was reached. To assess the agreement between the reviewers, the kappa index, which measures inter-rater agreement, was used; a value greater than 0.8 was considered almost perfect (10). All the chosen studies were imported into bibliographic database software (Zotero), and then the references were exported to an Excel spreadsheet.

Assessment of risk of bias

The COSMIN Risk of Bias checklist was used to assess the methodological quality of studies in terms of measurement properties, such as guidance for designing or reporting study measurement properties (6). Ten boxes were evaluated: instrument development, content validity, structural validity, internal consistency, cross-cultural validity\measurement invariance, reliability, measurement error, criterion validity, hypothesis testing for construct validity, and responsiveness. These were classified into 4 categories: very good, adequate, doubtful, and inadequate. For each measurement property in each study, the COSMIN item with the lowest score indicated the overall methodological quality (i.e., worst-score-counts method) (11).

Results

The inter-observer agreement was measured by the kappa test and yielded a score of 0.83. The exhaustive search by the authors identified a total of 2617 records, of which 2468 were discarded after an evaluation of the titles and abstracts; the full texts of 31 articles were examined. In the end, 14 articles met the inclusion criteria and were analyzed in the present systematic review (Fig. 1).

The risk of bias assessment is presented in Table 1. The articles included all used the COSMIN Risk of Bias checklist and successfully assessed 6 of the 10 evaluation parameters, which were as follows: instrument development (in the "very good" category, at 100%), content validity (in the "adequate" category, at 57.1%), structural validity (in the "adequate" category, at 64.3%), internal consistency (in the "very good" category, at 100%), reliability (in the "very good" category, at 57.4%), and hypothesis testing for construct validity (in the "very good" category, at 42.9%). The COSMIN checklist consists of 10 boxes that gather data about measurement properties; the following were not included: intercultural validation/ measurement invariance and measurement error were not evaluated by the included articles. The criterion validity and responsiveness parameter was not considered because there was no "gold standard" instrument that could be used to compare the results (Fig. 2), according to the COSMIN Risk of Bias indications (6,7).

The Rapid Estimate of Adult Literacy in Dentistry (REALD) 30, which was elaborated by Lee et al. (4) in the USA, was the measurement instrument that had the most cultural adaptions,

and was used in such countries as Saudi Arabia (12), Brazil (13), Turkey (14), and Romania (15). The sample sizes of the included studies ranged from 177 to 1405 adults, and the items per measurement instrument ranged from 14 to 99; the statistical techniques used in the structural validity evidence (psychometric analysis of the test) were exploratory factor analysis, confirmatory factor analysis, and Rasch analysis. Evidence of convergent, predictive, concurrent, and discriminant validity was obtained.

In the psychometric property of reliability, the area of internal consistency in all the included studies was evaluated with Cronbach's alpha coefficient, and a range of .789 to .91 was obtained. Ho et al. (16) used split-half reliability with the Spearman–Brown coefficient. In the parameter of stability of measurements (test–retest), the statistic used in 8 articles was the intraclass correlation coefficient (ICC), with a range of .73 to .99 (12–15,17–20), likewise, Sfeatcu et al. (15) used the Spearman coefficient. For more information, see Table 2.

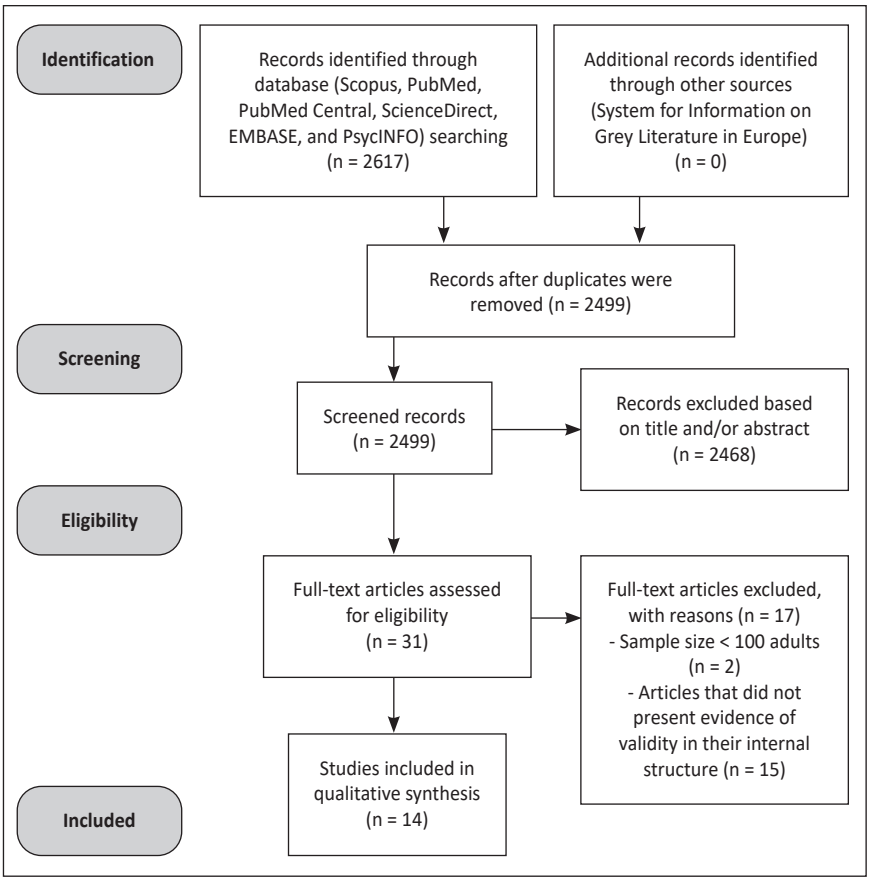

Figure 1. Flowchart of the studies included in the systematic review

Discussion

Sample size

In 6 of the 14 articles analyzed (5,16,19–22), the relationship between the ratio of subject/item was greater than or equal to 10:1, and in just 3 studies, the relationship was greater than 20:1 (5,16,22). In addition, to determine the sample size, only Mialhe et al. (22) indicated that the 10:1 subject/item ratio recommended by Hair et al. (23). Likewise, Peker et al. (14) used sample size to calculate internal consistency with the following parameters: type I error probability (α) equals .05, power ($1-\beta$) equals .8, and expected level for Cronbach's alpha equals .80; the other studies used the calculation of the sample size for convenience, without a design that guides the minimum parameters of the statistical models (24).

The size of the sample in psychometric studies is generally determined based on the number of items of the instrument, and a participant/item ratio of from 10:1 to 20:1 is considered acceptable to guarantee the quality of the analysis (factor loadings, communalities, and indices of goodness of fit). Hair et al. recommend that the sample should be greater than 100 subjects (23), while other studies suggest that a sample of 300 or more subjects will probably provide a stable factorial structure (25,26). Therefore, it is safe to conclude that researchers who use large samples and make informed decisions about the options available for data analysis are the most likely to achieve their goal: to reach conclusions that will generalize beyond a sample and, in particular, to whatever population is of interest. It has been generally found that small samples tend to be less useful beyond their aptness to the samples themselves and the associated analysis (27).

Psychometric properties of the instruments

Exploratory factor analysis (EFA). This is used to identify the number and composition of common factors (latent variables) necessary to explain the common variance of the set of items; in the present study, the dominant technique was principal component analysis (PCA) (15,18), also known as the EFA of principal components (17,19,21,28), which mostly employs an orthogonal varimax rotation. The PCA is not properly a factor analysis method, nor does it reproduce the EFA model because its objective is to explain the total variance. Therefore, the PCA model is not needed to obtain initial estimates of commonality; however, it is an extended practice as a factor extraction method (24). The EFA has as its main purpose the search for a structure of dimensions or latent variables, based on the correlations between the observed variables from the identification of a set of common factors (27). Ho et al. (16) performed the EFA with the

Table 1. COSMIN Risk of Bias checklist

Reference	Country	Instrument	Instrument development	Content validity	Structural validity	Internal consistency	Cross-cultural validation/ measurement invariance	Reliability (test-retest)	Measurement error	Criterion validity	Hypothesis testing for construct validity	Responsiveness
Lee et al., 2007 (4)	USA	REALD-30	VG	A	A	VG	NI	NI	NI	NA	VG	NA
Stucky et al., 2011 (5)	USA	TS-REALD	VG	A	MB	VG	NI	NI	NI	NA	VG	NA
Wong et al., 2012 (17)	CHINA	HKREALD-30	VG	A	MB	VG	NI	VG	NI	NA	A	NA
Gironda et al., 2013 (21)	USA	REALMD-20	VG	A	A	VG	NI	NI	NI	NA	VG	NA
Jones et al., 2013 (28)	AUSTRALIA	HELD-29	VG	A	A	VG	NI	NI	NI	NA	I	NA
Tadakamadla et al., 2014 (12)	SAU	AREALD-30	VG	VG	VG	VG	NI	VG	NI	NA	VG	NA
Junkes et al., 2015 (13)	BRAZIL	BREALD-30	VG	VG	A	VG	NI	VG	NI	NA	A	NA
Pakpour et al., 2016 (18)	IRAN	IREALD-99	VG	A	A	VG	NI	VG	NI	NA	VG	NA
Cruvinel et al., 2017 (19)	BRAZIL	REALMD-20	VG	A	A	VG	NI	VG	NI	NA	A	NA
Peker et al., 2017 (14)	TURKEY	TREALD-30	VG	VG	VG	VG	NI	VG	NI	NA	VG	NA
Ho et al., 2019 (16)	TAIWAN	MOHL-AQ	VG	VG	VG	VG	NI	NI	NI	NA	NI	NA
Mialhe et al., 2020 (22)	BRAZIL	HELD	VG	VG	A	VG	NI	NI	NI	NA	A	NA
Taoufik et al., 2020 (20)	GREECE	GROHL	VG	VG	A	VG	NI	VG	NI	NA	I	NA
Sfeatcu et al., 2020 (15)	ROMANIA	RREALD-30	VG	A	A	VG	NI	VG	NI	NA	I	NA
Frequency and percentage of categories		VG	n = 14	n = 6	n = 5	n = 14	-	n = 8	-	-	n = 6	-
			100%	42.9%	35.7%	100%	-	57.4%	-	-	42.9%	-
		A	-	n = 8	n = 9	-	-	-	-	-	n = 4	-
				57.1%	64.3%						28.6%	-
		D	-	-	-	-	-	-	-	-	-	-
		I	-	-	-	-	-	-	-	-	n = 3	
											21.4%	

VG: very good; A: adequate; D: doubtful; I: inadequate; NI: no information; NA: not applicable

estimation method of principal axis factoring (part of ordinary least squares) and, to retain the number of factors, used parallel analysis.

Confirmatory factor analysis (CFA). This method of analysis allows the researcher to define how many factors are expected, which factors are related to each other, and which items are related to each factor (7). Five of the retrieved articles carried out the CFA (5,12,14,16,22), with the estimation methods being the maximum likelihood (ML) method (16,22), weighted least squares (WLS) (12), and WLS means and variance adjusted (WLSMV) (5). According to Beauducel et al. (29), ML estimation requires compliance with the assumption of the multivariate normal distribution of the data; in addition, when WLS estimation was used for ordinal data, analysis revealed large amounts of bias, especially in small samples and moderate loads, as this estimator requires large sample sizes (more than 1000 cases). The WLSMV estimation used by Stucky et al. (5) does not require large sample sizes (around 200 cases) compared to WLS and ML estimation, and the magnitude of the loads were accurately estimated when the variables had 2 or 3 categories in comparison with ML (29).

For the model fit, the indices used in the 5 studies were the root mean square error of approximation (RMSEA), which is an absolute measure of fit; the comparative fit index (CFI), for incremental fit measures; and the minimum discrepancy function divided by degrees of freedom for parsimony fit measures (5,12,14,16,22). To a lesser extent, the indices reported were as follows: the goodness of fit index, standardized root mean residual (SRMR), the incremental fit index, and the Tucker–Lewis index (TLI). The indices that were assessed (and the quality of those assessments) using the COSMIN Risk of Bias checklist were the CFI or TLI (> .95), the RMSEA (< .06), and the SRMR (< .08) (7).

Reliability

Internal consistency. All the studies used Cronbach's alpha coefficient as a measure of internal consistency, with the values considered acceptable; this coefficient depends on the magnitude of the correlation between items and the number of items in the instrument (30). Furthermore, Ho et al. (16) used the method of split-half with the Spearman–Brown statistic.

Table 2. Psychometric properties of the articles included

Instrument OHL	Number of items	Evidence of validity			Reliability		Sample size
		Based on content	Structural validity (Based on the internal structure)	Hypothesis testing for construct validity	Internal consistency	Measurement stability (Test–retest)	
REALD-30 (Lee, 2007) USA	30 items (d)	Only by the authors *Pilot test:* 15	*Factor analysis* Tetrachoric correlations matrix. 2 Factors: 1st = 8.78 eigenvalue, 70.7% variance 2nd = 2.10 eigenvalue, 16.9% variance	*Convergent validity evidence:* -with REALM (r = 0.86)/TOFHLA (r = 0.64); *P* < .05 *Predictive validity evidence:* multivariate regression analysis -with OHIP 14 (-0.14); *P* < .05 -with SDHS (0.35); *P* > .05	Cronbach's alpha = .87	NI	200 adults
TS-REALD (Stucky, 2011) USA	26 items (d)	Instrument REALD-30 (Lee, 2007) USA *Pilot test:* NI	*Confirmatory factor analysis* Polychoric correlation matrix. Estimation: WLSMV Fit indices: X2(114) = 613; CFI = 0.95; TLI = 0.97; RMSEA = 0.056 *Item response theory (2PL)* Slope parameters (a) = 2.91 to 1.09 (mean = 2.12; SD = 0.44) Difficulty parameters (b) = 1.97 to -2.93 (mean = -0.01; SD = 1.42)	*Convergent validity evidence:* -with REALD-30 = 0.96 -with NVS = 0.51; *P* < .05 *Predictive validity evidence:* Multiple regression model -with OHIP-14 (β = 0.10; SD = 0.04); *P* < 0.05	> .85	NI	1405 women
HKREALD-30 (Wong, 2012) CHINA	30 items (d)	Cultural adaptation (forward–backward translation, REALD-99) *Pilot test:* NI	*Principal component factor analysis* Promax rotation (<0.3) *Partial credit Rasch model* infit/outfit ZSTD (-2 and 2) infit/outfit MNSQ (0.70 and 1.30)	*Convergent validity evidence:* -with TOFHLiD (rho = 0.693); *P* < .01	Cronbach's alpha = .84	ICC: 0.78	200 adults
REALMD-20 (Gironda, 2013) USA	20 items (d)	Only by the authors *Pilot test:* NI	*Principal component factor analysis* Varimax rotation = 2 Factors: 1st = 7.14 eigenvalue, 45% variance 2nd = 1.77 eigenvalue	*Convergent validity evidence:* -with REALM (rho = 0.894); *P* < .001	Cronbach's alpha = .86	NI	200 adults
HELD-29 (Jones, 2013) AUSTRALIA	29 items (p)	*Experts*: Australian Research Centre for Population Oral Health *Pilot test:* NI	*Principal component factor analysis* Varimax rotation = 7 Factors	*Convergent validity evidence:* -with oral health questions (r = 0.13–0.22); *P* < .05	Cronbach's alpha = .91	NI	209 adults
AREALD-30 (Tadakamadla, 2014) SAU	30 items (d)	*Cultural adaptation* (forward–backward translation, REALD-30) *Experts* 2 Bilingual dentists 1 Bilingual dentist *Pilot test:* 20	*Confirmatory factor analysis* Estimation: WLS with asymptomatic covariance matrix. Model 2: X2 = 1803.87/df = 405/CFI: 0.89/NNFI: 0.88/PNFI = 079/RMSEA: 0.14 *Rasch analysis/partial credit model:* 1. Infit MNSQ: 0.50 – 2.0 2. Person and item reliability estimates: 0.86 and 0.98 3. Person separation index: 2.45–2.80 4. Amount of variance by Rasch: 50.9%	*Convergent validity evidence:* -with REALD-99 (rho = 0.959); *P* < .01 *Predictive validity evidence:* -with OHIP-14 (rho = -0.105) -with SDHS (rho = 0.136)	Cronbach's alpha = .89	Retest: 20 ICC = 0.99	177 adults

Instrument OHL	Number of items	Evidence of validity: Based on content	Evidence of validity: Structural validity (Based on the internal structure)	Evidence of validity: Hypothesis testing for construct validity	Reliability: Internal consistency	Reliability: Measurement stability (Test–retest)	Sample size
BREALD-30 (Junkes, 2015) BRAZIL	30 items (d)	*Cultural adaptation* (forward–backward translation, REALD-30) *Experts* Committee of experts *Pilot test:* 10	*Exploratory factor analysis* = 1 Factor: 1st = 7.36 eigenvalue, 24.5% variance	*Convergent validity evidence:* -with NFLI (rho = 0.593); *P* < .001 -with educational attainment (rho = 0.541); *P* < .001 *Predictive validity evidence:* -with SDHS; *P* = .003 -with OHIP-14 (rho = -0.080); *P* = .198	Cronbach's alpha = .88–.89	ICC = 0.983	258 adults
IREALD-99 (Pakpour, 2016) IRAN	99 items (d)	*Cultural adaptation* (forward–backward translation, REALD-99) *Pilot test:* 12	*Principal component factor analysis* Varimax rotation 1st = 47% variance 2nd = 42% variance 3rd = 11% variance *Rasch's polytomous model*	*Convergent validity evidence:* -with TOFHLiD (rho = 0.72); *P* < .01 -with SDHS (rho = 0.31); *P* < .01	Cronbach's alpha = .98	ICC = 0.97	421 adults
REALMD-20 (Cruvinel, 2017) BRAZIL	20 items (d)	*Cultural adaptation* (forward–backward translation, REALMD-20, Gironda, 2013) *Pilot test:* 10	*Exploratory factor analysis, principal component analysis* Varimax rotation 1st = 4.53 eigenvalue, 25.18% variance 2nd = 1.88 eigenvalue, 10.46% variance	Convergent validity evidence: -with BNFLI (rho = 0.60); *P* < .001 -with REALD-30 (rho = 0.73); *P* < .001	Cronbach's alpha = .789	ICC = 0.73	200 adults
TREALD-30 (Peker, 2017) TURKEY	30 items (d)	*Cultural adaptation* (forward–backward translation, REALD-30) *Experts:* 8 experts 4 translators *Pilot test:* 35	*Principal component analysis* Varimax rotation = 2 factors 1st = 8.31 eigenvalue, 27.72% variance 2nd = 2.26 eigenvalue, 7.53% variance *Confirmatory factor analysis* X^2/df = 1.34/CFI: 0.89/IFI: 0.90/TLI: 0.89/RMSEA: 0.052 (2 factors) *Rasch analysis/partial credit model:* 19 items (63.3%) = infit/outfit 0.7 and 1.3 Amount of variance by Rasch = 37.9 variance	*Convergent validity evidence:* -with REALM (rho = 0.73); *P* < .01 *Predictive validity evidence:* -with SDHS (rho = 0.34); *P* < .01 -with OHIP-14 (rho = -0.28); *P* < .01	Cronbach's alpha = .91	ICC = 0.99	127 adults
MOHL-AQ (Ho, 2019) TAIWAN	17 items (d)	*Cultural adaptation* (forward–backward translation, OHL-AQ) Experts: Committee of 5 experts in public health *Pilot test:* 30	*Exploratory factor analysis* Estimation: principal axis Number of factors: parallel analysis 1st = 3.85 eigenvalue, 22.62% variance *Confirmatory factor analysis* Estimation: maximum likelihood GFI = 0.93/AGFI = 0.92/SRMR = 0.94/RMSEA: 0.04/X^2/df = 1.86/CFI: 0.90/NNFI = 0.87/IFI: 0.89/PGFI = 0.73/PNFI = 0.78	NI	Cronbach's alpha = .77 split-half method, Spearman–Brown = 0.78	NI	402 adults
HELD (Mialhe, 2020) BRAZIL	14 and 29 items (p)	*Cultural adaptation* (forward–backward translation, HELD-14 and 29) *Pilot test:* 25	*Confirmatory factor analysis HELD-14* Estimation: maximum likelihood X^2/df = 1.8–2.3; CFI = 0.97–0.98; GFI/NFI = 0.98–0.99; RMSEA = 0.05; SRMR = 0.03 *Confirmatory factor analysis HELD-29* Estimation: maximum likelihood X^2/df = 4.3; CFI = 0.84; GFI/NFI = 0.88; RMSEA = 0.12–0.13; SRMR = 0.07	*Convergent validity evidence:* -with average variance extracted (AVE) ≥ 0.50 *Discriminant validity evidence:* -with composite reliability (CR) ≥ 0.70	Cronbach's alpha ≥ .87	NI	535 adults

Instrument OHL	Number of items	Evidence of validity			Reliability		Sample size
		Based on content	Structural validity (Based on the internal structure)	Hypothesis testing for construct validity	Internal consistency	Measurement stability (Test–retest)	
GROHL (Taoufik, 2020) GREECE	20 items (d)	*Cultural adaptation* (forward–backward translation, READ-30, READ-99) *Experts:* 2 dentists	*Item response theory (2PL)* (reduction of 44 items) -excluded 12 = insufficient invariance -excluded 10 = item-test correlation < 40 -excluded 02 = difficulty > 0.05	*Convergent validity evidence:* positively correlated -with oral hygiene behaviors -with dental attendance -with education level *Predictive validity evidence:* with OHIP-14 (rho = 0.10); *P* = .11	Cronbach's alpha = .80	Retest: 20 ICC = 0.95	282 adults
REALD-30 (Sfeatcu, 2020) ROMANIA	30 items (d)	*Cultural adaptation* (forward–backward translation, REALD-30) *Pilot test:* 20	*Principal component analysis* = 1 factor *Item response theory/Rasch model* evidenced discrimination capacity of the items	*Validity evidence:* -with OHIP-14; *P* = .004 -with SDHS; *P* < .001	Cronbach's alpha = .88	Retest: 20 ICC = 0.90 rho = 0.98	224 adults

d: dichotomous; p: polytomous; WLSMV: Weighted Least Squares Mean and Variance adjusted; WLS: Weighted Least Squares; 2PL: 2-parameter logistic model; SD: standard deviation; SDHS: self-perceived dental health status; NVS: newest vital sign; NFLI: National Functional Literacy Index; ICC: intraclass correlation coefficient; NI: no information; BNFLI: Brazilian National Functional Literacy Index; df: degrees of freedom; MNSQ: mean square; TOFHLID: Test of Functional Health Literacy in Dentistry; REALM: Rapid Estimate of Adult Literacy in Medicine; ZSTD: standardized residuals; OHIP-14: Oral Health Impact Profile; REALD: Rapid Estimate of Adult Literacy in Dentistry; REALMD: dental/medical health literacy screen; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index; IFI: Incremental Fit Index; GFI: Goodness of Fit Index; SRMR: Standarized Root Mean Square Residual; RMSEA: Root Mean Square Error of Approximation; NNFI: Non-Normed Fit Index; PGFI: Parsimony Goodness of Fit Index; PNFI: Parsimony Normed Fit Index

Here, the test is divided into 2 halves (which must be equivalent) to show adequate internal consistency (30).

There is extensive literature that criticizes the use of the Cronbach's alpha coefficient without considering the data distribution and the sample size. The requirements for using this measure are demanding and require the presence of tau-equivalence (unlikely to obtain), the absence of correlation between errors, and the presence of data normality (31,32). According to Trizano-Hermosilla et al. (33), who simulated the Cronbach's alpha coefficient, McDonald's omega, and greatest lower bound (GLB) in a 1-dimensional model in terms of skewness and not tau-equivalence, the results showed that the omega coefficient is a better option than Cronbach's alpha, and that in the presence of skew items, it is preferable to use the omega coefficient and GLB, even in small samples.

Measurement stability. When it comes to the statistical methods used to assess the test–retest reliability, all the studies that performed this test (12–15,17–20) selected appropriate statistical methods based on the recommendation of the ICC (for continuous scores) and the kappa statistic (for categorical scores) (11).

Item response theory

Item Response Theory (IRT) is a set of model-based psychometric techniques used to examine the relationship between item responses and the underlying latent ability; the relationship between an item response and the latent ability is represented by an item characteristic curve (34). Fifty percent of the articles included (5,12,14,15,17,18,20) that in turn used dichotomous items were evaluated with the IRT; of the included articles, 4 used the partial credit model (PCM) of GN Masters (35), which is an extension of the model developed by Rasch (36). In addition, Pakpour et al. (18) used a Rasch polytomous model. The PCM is a model designed for polytomous items, that is, those with K response categories (where K>2) (35). The models used to evaluate dichotomous items are those called 1-, 2-, 3-, and 4-parameter models, both in their normal and logistic versions (37); only Stucky et al. (5) and Taoufik et al. (20) used the 2-parameter logistic model (2PL), which evaluates the discrimination index (a) and the difficulty of the item (b). Likewise, Sfeatcu et al. (15) used the Rasch model, also called the 1PL, in which the parameter "b" was evaluated (37).

Some important considerations in the choice of the model are the characteristics of the items (dichotomous or polytomous) and the sample-size requirements of each of them. In this sense, the 1PL model can work with considerably reduced sample sizes (minimum 200 participants) in relation to those of the other models (minimum 500). It should also be noted that a simpler model is always preferable (34). The rapid acceptance and expansion of IRT over the last decade suggests that the methodology has become a mainstay of measurement instrument validation (34,37).

Hypothesis testing for construct validity

Unlike the internal structure of the test, these measurement properties mainly assess the quality of the scale or subscale as a whole, rather than the items (6). The COSMIN Risk of Bias checklist was used to assess a set of hypotheses that concerned the expected

Table 3. Guide for the minimum validation procedures of the OHL construct

	CTT	IRT
Sample size	300–500 subjects Ratio: subject/item 10:1–20:1	*Dichotomous items* 1 PL ≥ 200 subjects 2 PL ≥ 1000 subjects Ratio: subject/item 10:1–20:1 *Polytomous items* Polytomous models are extensions of dichotomous models, considered according to the model to be used.
Structural validity	*CFA* -Estimation method: WLSMV -Fit indices: CFI or TLI > 0.95 RMSEA < 0.06 SRMR < 0.08 *EFA* -Association matrix: Matrix of tetrachoric (dichotomous items) or polychoric (polytomous items) correlations -Factor estimation method: Unweighted least square -Number of factors to retain: According to Parallel Analysis -Rotation method factors: Oblique rotation	*Rasch analysis/1PL* -No violation of unidimensionality: CFI or TLI > 0.95; RMSEA < 0.06; SRMR < 0.08 -No violation of local independence: Residual correlations among the items after controlling for the dominant factor < 0.20 or Q3's < 0.37 -No violation of monotonicity: -Adequate looking graphs OR item scalability > 0.30 Adequate model fit: IRT: X^2 > 0.01 Rasch: infit/outfit mean squares: ≥ 0.5 and ≤ 1.5 Z-standardized values: > -2 and < 2
Internal consistency	McDonald's omega coefficient ≥ 0.70 Cronbach's alpha coefficient ≥ 0.70	
Test–retest	ICC ≥ 0.70	
Hypothesis testing for construct validity	Expected relationships with other well-defined and high-quality instruments (convergent validity). The result agrees with the hypothesis.	

OHL: Oral Health Literacy; CTT: Classical Test Theory; IRT: Item Response Theory; PL: Parameter Logistic; CFA: Confirmatory Factor Analysis; EFA: Exploratory Factor Analysis; WLSMV: Weighted Least Squares Means and Aariance adjusted; CFI: Comparative Fit Index; TLI: Tucker–Lewis Index; RMSEA: Root Mean Square Error of Approximation; SRMR: Standardized Root Mean Residuals; ICC: Intraclass Correlation Coefficient

relationships between the instrument under review and other well-defined and high-quality instruments (6,11). In this regard, 3 studies (4,14,21) made the comparisons with measurement instruments that evaluate literacy in general health (convergent validity), specifically the Rapid Estimate of Adult Literacy in Medicine (38); in addition, 4 of the included articles (5,12,17,18) made the comparison with other instruments that measure the same construct, which instruments were as follows: the Test of Functional Health Literacy in Dentistry (39), the READ-30 (4), and the READ-99 (40); their use resulted in direct and significant correlations. Seven studies hypothesized (4,5,12–15,20) an inverse and significant relationship (predictive validity) with the Oral Health Impact Profile-14 (41), obtaining uneven results.

Based on the present systematic review results, the authors propose a guide for the validation procedures—preserving the minimum psychometric properties—to be used in the evaluation of the Oral health literacy construct. As has been stated regarding the included articles that presented an instrumental design, once the content validity evidence (boxes 1 and 2) had been obtained using the COSMIN checklist (7), the dimensionality of the test was studied to acquire evidence of the validity of its internal structure (box 3). According to the classical test theory, the EFA and the CFA are the most known and used techniques to examine the internal structure that underlies the scores of an evaluation instrument. According to the IRT, the 1PL and its variant in polytomous models is the recommended technique due to its lower requirement in terms of sample size; once the dimensionality of the scores was determined, the reliability estimation was carried out (boxes 4 and 6). Subsequently,

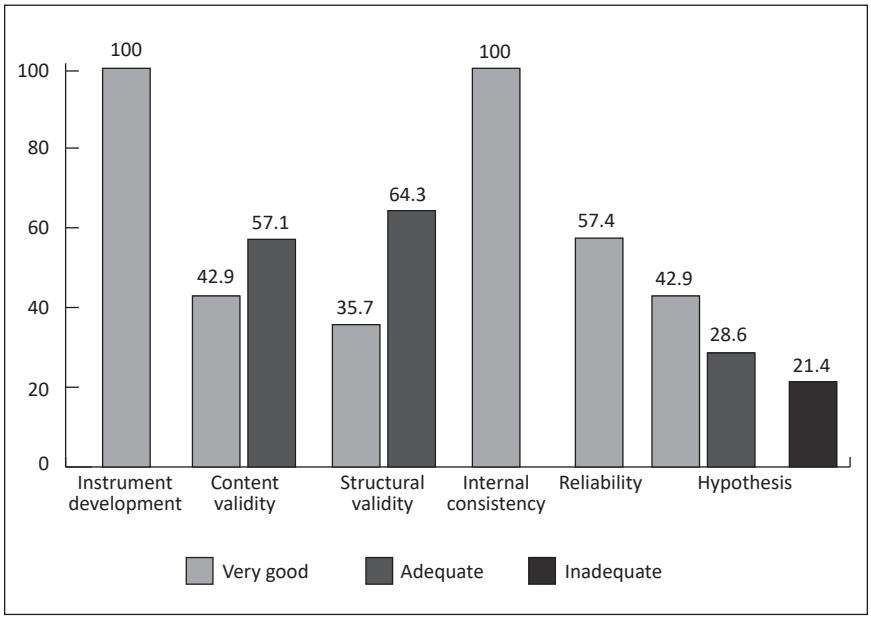

Figure 2. Scoring of metric properties

to obtain evidence of external validity, the relationship of the measurement instrument with other evaluation instruments had to be observed (box 9). For more details, see Table 3.

Study limitations

The instruments evaluated were not validated with any clinical result (That is, oral health status).

Conclusion

The psychometric properties that were evaluated (using the COSMIN Risk of Bias checklist) in this systematic review were structural validity, internal consistency, reliability (test–retest), and hypothesis testing for construct validity. To date, there is no "gold standard" measuring instrument to evaluate criterion validity or responsiveness. The other parameters used with the COSMIN checklist were instrument development and content validity. The measurement instrument most frequently culturally adapted was REALD-30 (USA), and it was validated in such countries as Saudi Arabia, Brazil, Turkey, and Romania.

Resumen

Objetivo: Evaluar mediante una revisión sistemática las propiedades psicométricas de los instrumentos para la medición de la alfabetización en salud bucal de adultos. Material y métodos: Se realizó una búsqueda electrónica de estudios instrumentales en bases de datos PubMed, PubMed Central, ScienceDirect, Scopus, EMBASE y PsycINFO, para hallar artículos publicados hasta junio del 2021. El riesgo de sesgo de los estudios incluidos se evaluó mediante la lista de chequeo COSMIN ("COnsensus-based Standards for the selection of health Measurement Instruments") Risk of Bias. Resultados: De una muestra inicial de 2617 artículos, 14 estudios instrumentales fueron incluidos en la presente revisión. Sus tamaños de muestra oscilaron entre 177 y 1405 adultos y el número de ítems por instrumento de medición variaron de 14 a 99. Para la validez estructural, las técnicas estadísticas se realizaron según la Teoría clásica de los test (análisis factorial exploratorio y confirmatorio) y Teoría respuesta al ítem (modelos dicotómicos y politómicos). El "Rapid Estimate of Adult Literacy in Dentistry 30" elaborado en USA, fue el instrumento de medida que tuvo mayores adaptaciones culturales, se validaron en países como: Arabia Saudita, Brasil, Turquía y Rumania. La evaluación de riesgo de sesgo según la lista de chequeo "COSMIN Risk of Bias", evidenció que se evaluaron seis de los diez parámetros. Conclusiones: Las propiedades psicométricas que se evaluaron en la presente revisión sistemática fueron: validez estructural, consistencia interna, fiabilidad ("test–retest") y prueba de hipótesis para validez de constructo. Hasta el momento no se tiene un instrumento de medición "gold standard" para evaluar el parámetro validez de criterio.

References

1. Peres MA, Macpherson LMD, Weyant RJ, et al. Oral diseases: a global public health challenge [published correction appears in Lancet. 2019 Sep 21;394(10203):1010]. Lancet. 2019;394(10194):249-260. doi:10.1016/S0140-6736(19)31146-8
2. Institute of Medicine. Oral health literacy: Workshop summary. Washington, DC: The National Academies Press; 2013. https://doi.org/10.17226/13484
3. 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):972-980. doi:10.14219/jada.archive.2012.0322
4. 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):94-98. doi:10.1111/j.1752-7325.2007.00021.x
5. Stucky BD, Lee JY, Lee SY, Rozier RG. Development of the two-stage rapid estimate of adult literacy in dentistry. Community Dent Oral Epidemiol. 2011;39(5):474-480. doi:10.1111/j.1600-0528.2011.00619.x
6. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;27(5):1171-1179. doi:10.1007/s11136-017-1765-4
7. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147-1157. doi:10.1007/s11136-018-1798-3
8. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097
9. Montero I, Leon OG. A guide for naming research studies in psychology. Int J Clin Health Psychol. 2007;7(3):847-862.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174.
11. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012
12. Tadakamadla SK, Quadri MF, Pakpour AH, et al. Reliability and validity of Arabic Rapid Estimate of Adult Literacy in Dentistry (AREALD-30) in Saudi Arabia. BMC Oral Health. 2014;14:120. Published 2014 Sep 29. doi:10.1186/1472-6831-14-120
13. 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. Published 2015 Jul 9. doi:10.1371/journal.pone.0131600
14. Peker K, Köse TE, Güray B, Uysal Ö, Erdem TL. Reliability and validity of the Turkish version of the Rapid Estimate of Adult Literacy in Dentistry (TREALD-30). Acta Odontol Scand. 2017;75(3):198-207. doi:10.1080/00016357.2016.1278079
15. Sfeatcu R, Lie SA, Funieru C, Åström AN, Virtanen JI. The reliability and validity of the Romanian rapid estimate of adult literacy in dentistry (RREALD-30). Acta Odontol Scand. 2021;79(2):132-138. doi:10.1080/00016357.2020.1814405
16. Ho MH, Montayre J, Chang HR, et al. Validation and evaluation of the Mandarin version of the oral health literacy adult questionnaire in Taiwan. Public Health Nurs. 2020;37(2):303-309. doi:10.1111/phn.12688
17. Wong HM, Bridges SM, Yiu CK, McGrath CP, Au TK, Parthasarathy DS. Development and validation of Hong Kong Rapid Estimate of Adult Literacy in Dentistry. J Investig Clin Dent. 2012;3(2):118-127. doi:10.1111/j.2041-1626.2012.00113.x
18. Pakpour AH, Lawson DM, Tadakamadla SK, Fridlund B. Validation of Persian rapid estimate of adult literacy in dentistry. J Investig Clin Dent. 2016;7(2):198-206. doi:10.1111/jicd.12135
19. Cruvinel AFP, Méndez DAC, Oliveira JG, et al. The Brazilian version of the 20-item rapid estimate of adult literacy in medicine and dentistry. PeerJ. 2017;5:e3744. Published 2017 Aug 29. doi:10.7717/peerj.3744
20. Taoufik K, Divaris K, Kavvadia K, Koletsi-Kounari H, Polychronopoulou A. Development of a Greek Oral health literacy measurement instru-

ment: GROHL. BMC Oral Health. 2020;20(1):14. Published 2020 Jan 15. doi:10.1186/s12903-020-1000-5
21. Gironda M, Der-Martirosian C, Messadi D, Holtzman J, Atchison K. A brief 20-item dental/medical health literacy screen (REALMD-20). J Public Health Dent. 2013;73(1):50-55. doi:10.1111/jphd.12005
22. Mialhe FL, Bado FMR, Ju X, Brennan DS, Jamieson L. Validation of the Health Literacy in Dentistry scale in Brazilian adults. Int Dent J. 2020;70(2):116-126. doi:10.1111/idj.12531
23. Hair JF Jr, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7th ed. Pearson Prentice Hall; 2010.
24. Howard M. A Review of Exploratory Factor Analysis (EFA) Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? Int J Hum-Comput Interact. 2015;32(1):51-62 doi:10.1080/10447318.2015.1087664.
25. Comrey A, Lee H. A First Course in Factor Analysis. 2nd ed. Lawrence Erlbaum Associates, Inc; 1992.
26. Field A. Discovering statistics using IBM SPSS statistics. 4th ed. SAGE Publications; 2013.
27. Costello AB, Osborne J. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. Pract Assess Res Eval. 2005;10:1-9. doi: https://doi.org/10.7275/jyj1-4868
28. Jones K, Parker E, Mills H, Brennan D, Jamieson LM. Development and psychometric validation of a Health Literacy in Dentistry scale (HeLD). Community Dent Health. 2014;31(1):37-43.
29. Beauducel A, Herzberg PY. On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. Struct Equ Modeling. 2006;13(2):186-203. https://doi.org/10.1207/s15328007sem1302_2
30. Streiner D, Norman G, Cairney J. Health measurement scales: a practical guide to their development and use. 5th ed. Oxford University Press; 2015. doi:10.1093/med/9780199685219.001.0001
31. Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika. 2005;70(1):123-133. https://doi.org/10.1007/s11336-003-0974-7
32. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. Psychometrika. 2009;74(1):107-120. doi:10.1007/s11336-008-9101-0
33. Trizano-Hermosilla I, Alvarado JM. Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. Front Psychol. 2016;7:769. Published 2016 May 26. doi:10.3389/fpsyg.2016.00769
34. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014;36(5):648-662. doi:10.1016/j.clinthera.2014.04.006
35. Masters GN. A rasch model for partial credit scoring. Psychometrika. 1982;47(2):149-174. https://doi.org/10.1007/BF02296272
36. Rasch G. Probabilistic models for some intelligence and attainment tests. Danmarks Paedagogiske Institut, Copenhagen; 1960.
37. Thomas ML. Advances in applications of item response theory to clinical assessment. Psychol Assess. 2019;31(12):1442-1455. doi:10.1037/pas0000597
38. Davis TC, Long SW, Jackson RH, et al. Rapid estimate of adult literacy in medicine: a shortened screening instrument. Fam Med. 1993;25(6):391-395.
39. Gong DA, Lee JY, Rozier RG, Pahel BT, Richman JA, Vann WF Jr. Development and testing of the Test of Functional Health Literacy in Dentistry (TOFHLiD). J Public Health Dent. 2007;67(2):105-112. doi:10.1111/j.1752-7325.2007.00023.x
40. Richman JA, Lee JY, Rozier RG, Gong DA, Pahel BT, Vann WF Jr. Evaluation of a word recognition instrument to test health literacy in dentistry: the REALD-99. J Public Health Dent. 2007;67(2):99-104. doi:10.1111/j.1752-7325.2007.00022.x
41. Slade GD. Derivation and validation of a short-form oral health impact profile. Community Dent Oral Epidemiol. 1997;25(4):284-290. doi:10.1111/j.1600-0528.1997.tb00941.x