BMJ Open

Validation practice for health literacy assessments: a systematic descriptive literature review using a theoretical validity testing framework

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-035974
Article Type:	Original research
Date Submitted by the Author:	28-Nov-2019
Complete List of Authors:	Hawkins, Melanie; Deakin University, Faculty of Health Elsworth, Gerald; Deakin University School of Health and Social Development, Faculty of Health; Swinburne University of Technology, Centre for Global Health and Equity, Faculty of Health, Arts and Design Hoban, Elizabeth; Deakin University, School of Health and Scial Development, Faculty of Health Osborne, Richard; Swinburne University of Technology, Centre for Global Health and Equity, Faculty of Health, Arts and Design
Keywords:	Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, QUALITATIVE RESEARCH, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title: Validation practice for health literacy assessments: a systematic descriptive literature review using a theoretical validity testing framework

Authors: Melanie Hawkins[1], Gerald Elsworth[1,2], Elizabeth Hoban[1], Richard H Osborne[2]

Institutions: Deakin University[1], Swinburne University[2]

Corresponding author:

Melanie Hawkins
School of Health and Social Development
Faculty of Health, Deakin University, Australia
Email: melanie.hawkins@deakin.edu.au
Phone: +61 439 354 456
Postal address: Deakin University
Faculty of Health
221 Burwood Highway
Melbourne, VIC, 3125
Australia

Co-authors:

Gerald R Elsworth
Honorary Professor (Health Program Evaluation)
School of Health and Social Development
Faculty of Health, Deakin University, Australia
Email: gerald.elsworth@deakin.edu.au

Adjunct Professor
Centre for Global Health and Equity
Faculty of Health, Arts and Design, Swinburne University of Technology, Australia

Elizabeth Hoban
Associate Professor
School of Health and Social Development
Faculty of Health, Deakin University, Australia
Email: elizabeth.hoban@deakin.edu.au

Richard H Osborne
Distinguished Professor of Health Sciences
Centre for Global Health and Equity
Faculty of Health, Arts and Design, Swinburne University of Technology, Australia
Email: rosborne@swin.edu.au

Author contributions: MH and RHO conceptualised the research question and analytical plan. MH led, with all authors contributing to, the development of the search strategy, selection criteria, data extraction criteria, and analysis method. MH conducted the literature search with guidance from EH. MH screened the literature, and extracted and analysed the data with the continuous support of and comprehensive checking by GRE. MH drafted the initial manuscript and led subsequent drafts. GRE, RHO and EH read and provided feedback on manuscript iterations, and approved the final manuscript. RHO is the guarantor.

Funding: MH was funded by a National Health and Medical Research Council (NHMRC) of Australia Postgraduate Scholarship (APP1150679). RHO was funded in part through a National Health and Medical Research Council (NHMRC) of Australia Principal Research Fellowship (APP1155125).

Conflicts of interest: None

Data availability statement: All data relevant to the study are included in the article or uploaded as supplementary information.

Word count: Abstract = 281; Main text = 4942

Acknowledgements: The authors acknowledge and thank Rachel West, Deakin University Liaison Librarian, for her expertise in systematic literature reviews and her patient guidance through the detailed process of searching the literature.

Abstract

Objective

Validity refers to the extent to which evidence and theory support the adequacy and appropriateness of inferences based on score interpretations. The health sector is lacking a theoretically-driven framework for validation practice for the development, testing and use of health assessments. This study used the *Standards for Educational and Psychological Testing* framework of five sources of validity evidence to categorise and count the types of evidence reported for health literacy assessments, and to identify studies that used or made reference to a theoretical validity testing framework.

Methods A systematic descriptive literature review investigated methods and results in peer-reviewed articles and examined theses about health literacy assessment development, application and validity testing studies. Electronic searches were conducted in EBSCOhost, EMBASE, Open Access Theses and Dissertations, and ProQuest Dissertations. Exclusions included studies published and health literacy assessments developed and administered in languages other than English. Data were coded to the *Standards*' five sources of validity evidence, and for direct and indirect reference to a validity testing framework.

Results Forty six studies met the inclusion criteria. Coding resulted in 195 instances of validity evidence across the five sources. Only nine studies directly or indirectly referenced a validity testing framework. Findings show that evidence based on *relations to other variables* is most frequently reported.

Conclusions The validity testing framework of the *Standards* facilitates examination of evidence based on five sources to determine the validity of inferences derived from health assessment data. Findings indicate that theoretical validity testing frameworks are rarely used in validation practice for health literacy assessments. Publication of evidence using the *Standards'* framework supports systematic and transparent reporting of validity testing research for review by other potential users of the health assessment.

Keywords Validity; Validation; Validity Testing Theory; Validity Testing Framework; Health Literacy; Health Assessment; Measurement.

Article summary

Strengths and limitations of this literature review

- This is the first time a theoretical validity testing framework, the five sources of evidence from the *Standards for Educational and Psychological Testing*, has been applied to the examination of validity evidence for health literacy assessments.
- A strength of this study is that validity is clearly defined, in accordance with the authoritative validity testing literature, as the extent to which theory and evidence (quantitative and qualitative) support score interpretation and use.
- A limitation was the restriction of the search to studies and health literacy assessments published or administered in English, which may introduce an English language and culture bias to the sample.
- A further limitation was the lack of clarity in some papers about the methods used and results obtained, leading to difficulties in coding validity evidence and may have led to some misclassification of reported evidence for some papers.

Validation practice for health literacy assessments: a systematic descriptive literature review using a theoretical validity testing framework

Background

It has been argued that the health sector is lacking a theoretically-driven framework of validation practice for the development, testing and use of health assessments. [1-6] Such a framework could guide and strengthen validation planning for the interpretation and use of health assessment data. [2, 3, 7] Interpretations of scores from health assessments are used to make decisions about the design, selection and evaluation of treatments, interventions, and policies. [2-4] To ensure that decisions based on data from health assessments are justified, and lead to equitable outcomes, validation practice must generate information about the degree to which the intended interpretations and use of data are supported by evidence and the theory of the construct being measured. [8-17] Validation research is complex [7, 18] and a theoretical framework would facilitate an evaluation of a range of evidence to determine valid interpretation and use of health assessment data. [2, 4, 16, 18, 19]

Contemporary validity testing theory

The validity testing framework of the 2014 *Standards for Educational and Psychological Testing* (the *Standards*) is the authoritative text for contemporary validity testing theory. [5] It results from about 100 years of the evolution of validity theory. [20, 21] The *Standards* defines validity as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (p.11) and validation as the process of '…accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations' (p.11). The framework describes five types of validity evidence that can be evaluated to justify test score interpretation and use: 1) *test content*; 2) *response processes* of respondents and users; 3) *internal structure* of the assessment test; 4) *relations to other variables*, and 5) *consequences* of testing, as related to validity (Table 1). [5, 6, 22, 23] Evidence from each of these sources may be needed to verify data interpretation and use.

Table 1. The five sources of validity evidence [5, 22]

1.	**Evidence based on test content** The relationship of the item themes, wording and format with the intended construct, including administration process.
2.	**Evidence based on response processes** The cognitive processes and interpretation of items by respondents and users, as measured against the intended construct.
3.	**Evidence based on internal structure** The extent to which item interrelationships conform to the intended construct.
4.	**Evidence based on external variables** The pattern of relationships of test scores to external variables as predicted by the intended construct.
5.	**Evidence based on validity and the consequences of testing** Intended and unintended consequences, as can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant variance.

The expectation of the *Standards* and leading validity theorists is that the validation process consists of an evaluative integration of different types of validity evidence (not types of validity) to support score meaning for a specific use. [2, 4, 5, 11-13, 24-30] Integral to this framework are quantitative methods to evaluate an assessment's statistical properties, but also important is validity evidence based on qualitative research methods. [4, 31-38] Qualitative methods are used to ensure technical evidence for *test content* and *response processes*, and to investigate validity-related *consequences* of testing. [7, 10, 25, 36-42] There are guides to assess quantitative measurement properties [43-45] but still needed are reviews that include qualitative validity evidence, and that place validity evidence for health assessments within a validity testing framework such as the *Standards*. [2, 4, 6, 22]

Health literacy

Health literacy is a relatively new field of research with evolving definitions of the concept [46-49] and advances in the approaches to its measurement. [50-56] Some health literacy assessments measure an observer's (e.g., clinician's) objective observations of a person's health literacy, which often consists of testing a person's numeracy, reading and comprehension. [57, 58] Objective measurement can support a clinician to provide health information in formats and at reading levels that are suited to individual patients but usually these measures do not assess other important dimensions of the health literacy construct. [59] Self-report (subjective) measures of health literacy have become useful with the rise of the patient-centred healthcare movement, and these typically provide individuals' perspectives of a range of aspects of their health and health contexts. [47, 60] This type of measurement can capture the multidimensional aspects of the health literacy construct to look at broader implications of treatment, care and intervention outcomes. [61] Assessments could also combine both objective and subjective measurement of health literacy. Data from health literacy assessments have been used to inform health literacy interventions [17, 62-66] and, increasingly, health policies. [67-71]

Rationale

As a guide to inform and improve the processes used to develop and test health assessments, this review will examine validation practice for health literacy assessments. An assumption underlying this review is that the field of health is not applying contemporary validity testing theory to guide validation practice, and that the focus of validation studies remains on the general psychometric properties of a health assessment rather than on the interpretation and use of scores. This study will provide an example of the application of the *Standards*' theoretical validity testing framework through the review of sources of validity evidence (generated through quantitative and qualitative methods) reported for health literacy assessments.

The aim of this systematic descriptive literature review was to use the validity testing framework of the *Standards* to categorise and count the sources of validity evidence reported for health literacy assessments and to identify studies that used or made reference to a theoretical validity testing framework. Specifically, the review addressed the following questions:

1. What is being reported as validity evidence for health literacy assessment data?
2. Do the studies place the validity evidence within a validity testing framework, such as that offered by the *Standards*?

Methods

King and He situate systematic descriptive literature reviews toward the qualitative end of a continuum of review techniques. [72] Nevertheless, this type of review employs a frequency analysis to categorise qualitative and quantitative research data to reveal interpretable patterns. [56, 72-77] This review will appraise validation practice for health literacy assessments using the *Standards'* framework of five evidence sources. It will not critique nor assess the quality of individual health literacy assessments or studies.

Inclusion and exclusion criteria, information sources, and search strategy

The method for this review was previously reported in a protocol paper. [22] The eligibility and exclusion criteria, information sources, and search terms are summarised in Table 2. Peer reviewed full articles and examined theses were included in the search. Supplementary file 1 shows the MEDLINE database search strategy, and this was modified for the other databases. The review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement. [78] See Supplementary file 2 for the PRISMA checklist.

Table 2. Summary of inclusion and exclusion criteria, information sources, and search terms

Inclusion criteria	**Exclusion criteria**
Not limited by start date: end date March 2019	Systematic reviews and other types of reviews
Development, application and validity testing studies about health literacy assessments	Health literacy assessments designed for specific demographic groups (e.g., children) or health conditions (e.g., kidney disease)
All definitions of health literacy; and objective, subjective, uni- and multi-dimensional health literacy assessments	Predictive, association or other comparative studies that do not claim in the abstract to contribute validity evidence
Studies published and health literacy assessments developed and administered in the English language	Health literacy assessments developed or administered in languages other than English^
Qualitative and quantitative research methods	Translation studies
Information sources: EBSCOhost (MEDLINE Complete, Global Health, CINAHL Complete, PsycINFO, Academic Search Complete); EMBASE; Open Access Theses and Dissertations; ProQuest Dissertations; references of relevant systematic reviews; authors' reference lists	

Search terms: Medical subject headings (MeSH) and text words - valid*, verif*, "patient reported outcome*", questionnaire*, survey*, "self report*", "self rated", assess*, test*, tool*, "health literacy", measure*, psychometric*, interview*, "think aloud", "focus group*", "validation studies", "test validity"

^ See *Results* for exceptions.

Article selection, and data extraction, analysis and synthesis

Duplicates were removed and a title and abstract screening of identified articles was performed in Endnote Reference Manager X9 by one author (MH). Identified full text articles were screened for relevance by MH and corroborated with an independent screening of 10% of the search results by a second author (GRE).

Data extraction from articles for final inclusion was undertaken by one author (MH) and comprehensively and independently checked by a second author (GRE). Both authors then corroborated to achieve categorisation consistency. General characteristics for each study were extracted but of primary interest were the sources of validity evidence reported, as were statements about or references to a theoretical validity testing framework. The validity evidence reported in each article was categorised according to the five sources of validity evidence in the *Standards*, whether or not the authors of the articles reported it that way. When the methods were unclear, the results were interpreted to determine the type of evidence generated by the study. A study was categorised as using or referencing a theoretical validity testing framework if the authors made a statement that referred to a framework and directly cited the framework document or if there was a clear citation path to the framework document.

Descriptive and frequency analyses of the extracted data were conducted to identify patterns in the sources of validity evidence being reported, and for the number of studies that made reference to a validity testing framework.

Patient and public involvement

Patients and the public were not involved in the development or design of this literature review.

Results

The PRISMA flow diagram in Figure 1 summarises the results of the search. [78] There were 3,379 records identified through database searches with 4 articles identified through other sources. There were 1,922 records when duplicates were removed. After applying the exclusion and inclusion criteria to all abstracts, with full text screening of 92 articles and theses, 40 articles and 6 theses were included in the review (n=46). Reasons for exclusion were that the health literacy assessment was developed in or administered in a language other than English (n=19); the assessment was specific to a disease or condition (n=8) or to a demographic group (n=2); the article was not a validity

study (n=8); the study was not using a health literacy assessment (n=3) or used an adapted assessment (n=4); the assessment was based on an item-bank, which required a different approach to validity testing (n=1), or was a composite assessment where health literacy data were collected and analysed with another type of data (n=1).

Figure 1. Flow diagram for Preferred Reporting Items for Systematic reviews and Meta-Analyses

Four papers were identified from the broader literature. Two of these were by Davis and colleagues and describe the development of the Rapid Estimate of Adult Literacy in Medicine (REALM) [57] and the shortened version of the REALM. [79] Neither of these papers were detected by the systematic review because Davis *et al*. do not claim these to be measures of health literacy but of literacy in medicine. Rather they state that both versions of the REALM are designed to be used by physicians in public health and primary care settings to identify patients with low reading levels. [57, 79-81] Nevertheless, we included these papers because the REALM and the shortened REALM have been used by clinicians and researchers as measures of health literacy, and are used either as the primary assessment or a comparator assessment in many studies. Two further papers [82, 83] were identified from the references of previous literature reviews.

Three papers identified in the database search were included in this review even though data were collected using translations of assessments originally developed in English. These studies were included because of the frequency of use of these assessments in the field of health literacy measurement, and because at least part of the data were based on English language research. The Test of Functional Health Literacy in Adults (TOFHLA) [84] and the Newest Vital Sign (NVS) [58] both collected data in English and Spanish. The analyses for the European Health Literacy Survey (HLS-EU) study [47] used data from the English (Ireland), as well as Dutch and Greek versions of the HLS-EU.

Of the 46 studies, 34 were conducted in the United States of America (USA), 8 in Australia, 2 in Singapore, and 1 each in Canada and the Netherlands. There were 4 studies published in the decade between 1990 and 1999, 8 studies between 2000 and 2009, and 34 between 2010 and 2019.

Reports of reliability evidence were provided in 33 studies (72%). This resulted in 44 instances of reliability evidence, of which 29 (66% of all instances) were calculated using Cronbach's alpha for internal consistency, 4 (9% of all instances) using test-retest, 4 (9%) using inter-rater reliability calculations, and 7 (16%) using other methods. See Table 3 for country and year of publication, and reliability evidence.

Table 3. Country and year of publication, and reliability evidence

Country of study	**N**	**%**
USA	34	74%
Australia	8	17%
Singapore	2	4%
Canada	1	2%
Netherlands	1	2%
Year of publication by decade		
1990-1999	4	9%
2000-2009	8	17%
2010-2019	34	74%
Reliability		
Cronbach's alpha	29	66%
Test-retest	4	9%
Inter-rater	4	9%
Other methods	7	16%
Total instances of reliability	44	100%

Research question 1. What is being reported as validity evidence for health literacy assessment data?

The data extraction framework was adapted from Hawkins et al (p.1702) [6] and Cox and Owen (p.254). [31] See Supplementary File 3. More detailed sub-coding of the five *Standards*' categories was done and will be drawn on selectively to describe aspects of the results (Supplementary File 4).

Data analysis consisted of coding instances of validity evidence into the five sources of validity evidence of the *Standards*. The results of the review are presented as: 1) the total number of instances of validity evidence for each evidence source reported across all studies; 2) the number of instances reported for objective, subjective and mixed methods health literacy assessments; and 3) the number of instances of evidence within each of the *Standards*' five sources, and a breakdown of the methods used to generate evidence.

Table 4 displays the overall results of the review. For the 46 studies that reported validity evidence for health literacy assessments, we identified 195 instances of validity evidence across the five sources: *test content* (n=52), *response processes* (n=7), *internal structure* (n=28), *relations to other variables* (n=107), and *consequences of testing* (n=1). Across types of health literacy assessments, there were 102 instances of validity evidence reported for health literacy assessments with an objective measurement approach (n=23 studies); 78 instances reported for assessments with a subjective measurement approach (n=20 studies); and 15 instances for assessments with a mixed

methods approach or when multiple types of health literacy assessments were under investigation (n=3 studies).

Table 4. Sources of evidence for all studies, total instances of validity evidence, and for objective, subjective, and multiple/mixed methods health literacy assessments

	Studies (n=46*)	**Instances**** (n=195)	**Objective^** (n=23 studies; n=102 instances)	**Subjective^^** (n=20 studies; n=78 instances)	**Multiple and mixed methods** (n=3 studies; n=15 instances)
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**
1. Test content	22 (48)	52 (27)	27 (26)	22 (28)	3 (20)
2. Response processes	6 (13)	7 (4)	2 (2)	5 (6)	0 (0)
3. Internal structure	15 (33)	28 (14)	11 (11)	15 (19)	2 (13)
4. Relations to other variables	42 (91)	107 (55)	61 (60)	36 (46)	10 (67)
5. Validity and the consequences of testing	1 (2)	1 (1)	1 (1)	0 (0)	0 (0)

*Most studies reported more than one source of validity evidence.
**Each time validity evidence was reported within a study.
^ Measures an observer's (e.g., clinician's) objective observations of a person's health literacy.
^^ Self-report (subjective) measure of health literacy.

Evidence based on test content

Nearly half of all studies (n=22) reported evidence based on test content, which resulted in 52 instances of validity evidence (Table 4 and Supplementary Table 1). Expert review was the most frequently reported method used to generate evidence (n=14 instances; 27% of all evidence based on test content), [47, 57, 58, 60, 79, 80, 85-92] followed by the use of existing measures of the construct (n=8; 15%). [58, 60, 80, 89-91, 93, 94] Analysis of item difficulty was used 5 times (10%), [60, 85, 88, 91, 95] with literature reviews, [47, 89, 92, 96] participant feedback processes about items, [47, 58, 80, 88] and construct descriptions [47, 60, 90, 96] each used 4 times (8% each). Participant concept mapping [47, 60, 87] and examination of administration methods [60, 97, 98] were each used 3 times (6% each), and participant interviews [87, 99] were used twice (4%). Five other methods were each used once in 5 different studies: item intent descriptions; [60] items tested against item intent descriptions; [100] IRT analysis for item selection within domains; [89] item selection based on hospital medical texts; [84] and item selection based on a health literacy conceptual model. [99]

Evidence based on response processes

Only 7 instances based on *response processes* were reported across 6 of the 46 studies (Table 4 and Supplementary Table 2). The methods used were cognitive interviews with respondents (n=3 instances; 43% of all evidence based on *response processes*) [60, 87, 100] and with users (clinicians) (n=1; 14%), [100] as well as recording and timing the response times of respondents (n=3; 43%). [88, 97, 99]

Evidence based on internal structure

There were 15 studies (33% of all studies) that reported evidence based on the *internal structure* of health literacy assessments resulting in 28 instances (Table 4 and Supplementary Table 3). The most frequently reported methods were exploratory factor analysis (EFA) (including principal component analysis (PCA)) (n=7 instances; 25% of all evidence based on *response processes*) [87, 92, 99, 101-104] and confirmatory factor analysis (CFA) (also n=7; 25%). [90, 105, 106] Differential item functioning (DIF) was reported 3 times (11%), [87, 90, 101] and item-remainder correlations twice (7%). [60, 91] There were 9 other methods used to generate evidence for *internal structure*, including a variety of specific item-response theory (IRT) analyses for fit, item selection, and internal consistency. Each method was reported once, with some authors reporting more than one method. [60, 85, 88, 89, 102, 105]

Evidence based on relations to other variables

This was the most commonly reported type of validity evidence across studies (n=42 studies; 91%) (Table 4 and Supplementary Table 4). There were 18 studies that only reported evidence based on *relations to other variables*. [82, 83, 103, 107-121] Evidence within this category was coded, as per the *Standards*, into convergent evidence (i.e., relationships between items and scales of the same or similar structure), discriminant evidence (i.e., assessments measuring different constructs determined to be sufficiently uncorrelated), criterion-referenced evidence (i.e., how accurately scores predict criterion performance), and evidence for group differences (i.e., relationships of scores with background characteristics such as demographic information). The *Standards* also includes evidence for generalisation but states that this relies primarily on studies that conduct research syntheses, and this review excluded studies that conducted meta-analyses. Across all studies, there were 107 instances of validity evidence reported for *relations to other variables*: 57 instances of convergent evidence (53% of all evidence in this category); 3 instances of discriminant evidence (3%); 17 instances of criterion-referenced evidence (16%); and 30 instances of evidence for group differences (28%).

The most frequently-used methods for convergent evidence were Spearman's [82, 84, 93, 95, 98, 104, 107, 109, 115, 117, 121] and Pearson's [57, 58, 79, 80, 89, 92, 103, 111, 112, 119, 122]

correlation coefficients (11 instances and 19% each). These were closely followed by the receiver operating characteristic (ROC) curve and the area under the ROC (AUROC) curve (also n=11 instances; 19%). [83, 96, 98, 102, 109, 110, 116, 119, 122] A further 8 instances (14%) of correlation calculations with similar measures were reported but the types of calculation they performed were unclear. [85, 86, 91, 94, 102, 114, 118, 120]

Harper, Elsworth *et al*., and Osborne *et al*. [60, 89, 105] were the only 3 studies to generate discriminant evidence, as defined by the *Standards*. Harper [89] used the Pearson correlation coefficient to assess the association of components of a new health literacy instrument with the shortened version of the Test of Functional Health Literacy in Adults (S-TOFHLA). Elsworth *et al*. [105] compared the average variance extracted (AVE) and the variance shared between the nine scales of the Health Literacy Questionnaire (HLQ) (discriminant validity evidence between HLQ scales). Similarly, Osborne *et al*. [60] conducted a multi-scale factor analysis to investigate if the nine HLQ scales were conceptually distinct.

Linear regression models were the most common method to generate criterion-referenced evidence (n=6 instances; 35% of all criterion-referenced evidence). [85, 89, 106, 113, 114, 120] The Chi-square test of independence was used by 3 studies (18%), [86, 114, 120] with Spearman's correlation coefficient [109, 114] and logistic regression models [85, 114] each used by 2 studies (12% each).

There were 16 methods used to generate evidence for group differences and these were spread across 19 studies. The most frequently used methods were analysis of variance (ANOVA) (n=5 instances; 17%) [87, 91, 92, 102, 120] and linear regression models (n=4; 13%). [80, 82, 90, 122]

Evidence based on validity and consequences of testing

One study did investigations that led to conclusions about validity and the *consequences* of testing (p.221). [80] Elder *et al*. found that the REALM underrepresented the construct of health literacy when defined as the ability to obtain, interpret, and understand basic health information.

Research question 2. Do the studies place the validity evidence within a validity testing framework, such as that offered by the Standards?

Few studies referred to a validity testing framework or used a framework to structure or guide their work. Of the 46 studies, 9 directly or indirectly referenced a validity testing framework, and made a statement to support the citation. The frameworks directly cited by 3 studies [86, 100, 105] were the 2014 *Standards*; [5] Michael T Kane's argument-based approach to validation; [12] Samuel J Messick's unified theory of validation; [15, 123] and Francis *et al*'s checklist operationalising measurement characteristics of patient-reported outcome measures. [124] There were 6 studies [60, 80, 92, 95, 101, 106] that indirectly cited Messick, Kane, and/or the 1985, 1999 or 2014 versions of the *Standards* [5, 125, 126] through other citations. A 10th study [87] referenced Buchbinder *et al*.

[127], which cites the *Standards*, but there was no clear statement about validity testing to support the citation.

Discussion

This systematic descriptive literature review found that studies in health literacy measurement rarely use or reference a structured theoretical framework for validation planning or testing. Further, this review's use of the *Standards*' framework revealed that validity testing studies for health literacy assessments most frequently, and often only, report evidence based on *relations to other variables*. It is usual and reasonable for a single validity study to not provide comprehensive evidence about a PROM, and this is why an organising framework for evaluating evidence from a range of studies is so important. The findings from this review show that validation practice for health literacy assessments does not use established validity testing criteria and is yet to embrace the structural framework of contemporary validity testing theory. [5, 6]

In this review, evidence based on *relations to other variables* was the most frequent type of validity evidence reported across the 46 studies. It was reported more than twice as frequently as evidence based on *test content*, which was the second most commonly reported source of validity evidence. Evidence based on *internal structure* was reported in almost half the studies. This is not an unexpected result given the propensity for validity testing studies to almost routinely conduct correlation of an assessment with another variable (e.g., a similar or different assessment). [128] In the early 20th Century, the focus of test validation was primarily on predictive validity practices (e.g., prediction of student academic achievement) and so correlation with known criteria was a common validation practice. [21, 129, 130] Development of the theory and practice of validation, and the need to use tests in various contexts with different population groups, has required consideration of the meaning of test scores, and that score interpretations usually lead to decisions or actions that can affect people's lives. [2, 3, 25, 39] As Kane explains, 'ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified' (p.17). [11] A structured theoretical framework, such as the *Standards*, facilitates validation planning, testing, and integration of evidence for decision making. It can also support new users of a health assessment to judge existing evidence and previous rationales for data interpretation and use, and how these might justify the use of the assessment in a new context.

Reports of evidence based on *response processes* and on *consequences* of testing were negligible in this review. This is the first time this has been observed in the field of health literacy although it has been observed previously in other fields of research. [23, 41, 131] Evidence based on the cognitive (response) processes of respondents (and of assessment users [32, 100]) can be essential to understanding the meanings derived from assessment scores for each new testing purpose. [42] Consequential evidence, although a controversial area of research, [23, 39] can reveal important

outcomes for equitable decision making, such as those discussed by Elder *et al*. [80] regarding the use of the REALM, a word recognition assessment, with non-native speakers of English in a world in which health literacy is understood to be about equitable access to, and understanding and use of health information and services. [67, 132-134] Potential risks for unintended consequences of testing can be lessened through the development of the content of health assessments using comprehensive grounded practices that ensure wide and deep coverage of the lived experiences of intended respondents. [60, 135-137]

The findings of this review are important because institutions and governments around the world are increasingly implementing health literacy as a basis for health policy and practice development and evaluation. [68-71, 138] There needs to be certainty that inferences made from health literacy measurement data are leading to accurate and equitable decision making about health care, interventions, and policies, and that these decisions are as fair for the people with the lowest health literacy as for those with the highest. [9, 17, 25, 71, 139-142] Some types of health interventions are known to widen health inequalities. [142-146] Messick emphasises construct underrepresentation and construct-irrelevant variance as causes for negative testing consequences, as related to validity. [123, 147] For example, if a health assessment is biased by a specific perspective about causes of health disparities then construct underrepresentation can be a threat to the validity of inferences and actions taken from the scores. Likewise, if an assessment reflects a particular social perspective (e.g., middle class values and language embedded in the items) then there is the threat that the responses to the assessment are perfused with irrelevant variance derived from that perspective. Evidence from a range of sources is required to justify the use of measurement data in specific contexts (e.g., socioeconomic, demographic, cultural, language), and to assure decision makers of the absence of validity threats. [4, 24, 27]

This is the first time that a comprehensive review of sources of validity evidence for health literacy assessments has been undertaken within the theoretical validity testing framework of the *Standards*. For some methods, coding into the five sources of validity evidence was not straightforward and, in these cases, the *Standards* were consulted closely for guidance. Coding of studies by Elsworth *et al*. and Osborne *et al*. [60, 105] to *relations to other variables* (discriminant evidence) required some deliberation because the evidence in both studies was for discrimination analyses between independent scales *within* a multi-scale health literacy assessment, rather than between different health literacy assessments. The developers of the HLQ view the nine scales as measuring distinct, albeit related, constructs. [60] The *Standards* (p.16) explain that 'external variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs'. [5] It was on the basis of the last part of this statement about tests measuring related or different

constructs that these two studies were coded in *relations to other variables* as discriminant evidence.

In a few studies, some assessments seemed to be regarded as proxies for health literacy, which suggested that the researchers were thinking of them as measuring similar constructs to health literacy. In these cases, evidence was coded in *relations to other variables* as convergent evidence (i.e., convergence between measures of the same or similar construct) rather than as criterion-referenced evidence (i.e., prediction of other criteria). For example, Curtis *et al*. [85] explored correlations between the Comprehensive Health Activities Scale (CHAS) with the Mini Mental Status Exam (MMSE) as well as with the TOFHLA, the REALM, and the NVS. [85] Driessnack *et al*. [107] looked at correlations between parents' and children's NVS scores with their self-reports of the number of children's books in the home. Dykhuis *et al*. [86] correlated the Brief Medical Numbers Test (BMNT) with the Montreal Cognitive Assessment (MoCA) as well as with two versions of the REALM.

Further to coding for *relations to other variables* are the distinctions between convergent evidence, criterion-referenced evidence, and evidence for group differences. Coding to convergent evidence was based on analyses of assessments of the same or similar construct (e.g., typically, comparisons of one health literacy assessment with another health literacy assessment). Coding to criterion-referenced evidence was based on analyses of prediction (e.g., a health literacy assessment with a disease knowledge survey). Coding for evidence of group differences was based on analyses of relationships with background characteristics such as demographic information.

Reliability was not coded within the five sources of evidence even though it does contribute to understanding the validity of score interpretations and use, especially for purposes of generalisation. [5] The *Standards* (p.33) classifies reliability into reliability/precision (i.e., consistency of scores across different instances of testing) and reliability/generalisability coefficients (i.e., in the way that classical test theory refers to reliability as being correlation between scores on two equivalent forms of a test, with the assumption that there is no effect of the first test instance on the second test instance). The predominant focus in the reviewed papers was on the latter conception of reliability, most often calculated using Cronbach's alpha.

Strengths and limitations

An element of bias is potentially present in this review because of the restriction of the search to studies published and health literacy assessments developed and administered in the English language. Future studies may be improved if other languages were included. The health literacy assessments reviewed are those that are predominant in the field and may well provide a foundation for validity studies of more specifically targeted assessments.

Validation practice is complex and there are many groups publishing validity testing studies that may have limited training and experience in the area. [1-4] There was a lack of clarity in some papers and theses about the methods used and results obtained, which caused difficulties with classifying the evidence within the *Standards* framework, so some misclassification is possible for some papers. Future work in this area would be improved if researchers used clearly defined and structured validity testing frameworks (i.e., the five validity evidence sources of the *Standards*) in which to classify evidence.

The main strength of this study was that validity is clearly defined as the extent to which theory and evidence (quantitative and qualitative) support score interpretation and use. This definition is in accordance with leading authorities in the validity testing literature. [2, 5, 11, 24] A second strength of this study was the use of an established and well-researched theoretical validity testing framework, the *Standards*, to examine sources of evidence for health literacy assessments. Different health literacy assessments have different measurement purposes. Validation planning with a structured framework would help to determine the sources of evidence needed to justify the inferences from data, and to guide potential users. Application of theory to validation practice will provide a scientific basis for the development and testing of health assessments, enable systematic evaluations of validity evidence, and help detect possible threats to the validity of the interpretation and use of data in different contexts.

[2, 3, 13],

Conclusions

The results of this literature review demonstrate that validation practice for health literacy assessments remains largely within the paradigm of correlation of assessments with other variables, and rarely is there reference to a theoretical framework to guide validation practice. Application of the *Standards*' framework will advance validation practice in health to support developers and users of health assessments to clearly outline their measurement purpose, and to define the relevant and appropriate validity evidence needed to ensure evidence-based, valid and equitable decision making for health.

References

1. McClimans, L., *A theoretical framework for patient-reported outcome measures.* Theoretical medicine and bioethics, 2010. **31**(3): p. 225-240.
2. Zumbo, B.D. and E.K. Chan, eds. *Validity and validation in social, behavioral, and health sciences*. Social Indicators Research Series. 2014, Springer International Publishing: Switzerland.
3. Sawatzky, R., et al., *Montreal Accord on patient-reported outcomes (PROs) use series–Paper 7: modern perspectives of measurement validation emphasize justification of inferences based on patient reported outcome scores.* Journal of Clinical Epidemiology, 2017. **89**: p. 154-159.
4. Kwon, J.Y., S. Thorne, and R. Sawatzky, *Interpretation and use of patient-reported outcome measures through a philosophical lens.* Quality of Life Research, 2019: p. 1-8.
5. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing*. 2014, Washington, DC: American Educational Research Association.
6. Hawkins, M., G.R. Elsworth, and R.H. Osborne, *Application of validity theory and methodology to patient-reported outcome measures (PROMs): building an argument for validity.* Quality of Life Research, 2018. **27**(7): p. 1695-1710.
7. O'Leary, T.M., J.A. Hattie, and P. Griffin, *Actual interpretations and use of scores as aspects of validity.* Educational Measurement: Issues and Practice, 2017. **36**(2): p. 16-23.
8. Chapelle, C.A., *The TOEFL validity argument.* Building a validity argument for the Test of English as a Foreign Language, 2008: p. 319-352.
9. Elsworth, G.R., S. Nolte, and R.H. Osborne, *Factor structure and measurement invariance of the Health Education Impact Questionnaire: Does the subjectivity of the response perspective threaten the contextual validity of inferences?* SAGE Open Medicine, 2015. **3**: p. 2050312115585041.
10. Shepard, L.A., *The centrality of test use and consequences for test validity.* Educational Measurement: Issues and Practice, 1997. **16**(2): p. 5-24.
11. Kane, M.T., *Validation*, in *Educational Measurement*, R.L. Brennan, Editor. 2006, Rowman & Littlefield Publishers / Amer Council Ac1 (Pre Acq). p. 17-64.
12. Kane, M.T., *An argument-based approach to validity.* Psychological Bulletin, 1992. **112**(3): p. 527-535.
13. Kane, M.T., *Explicating validity.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 198-211.
14. Messick, S., *Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning.* American Psychologist, 1995. **50**(9): p. 741.
15. Messick, S., *Validity of test interpretation and use.* ETS Research Report Series, 1990. **1990**(1): p. 1487-1495.
16. Moss, P.A., B.J. Girard, and L.C. Haniford, *Validity in educational assessment.* Review of research in education, 2006: p. 109-162.
17. Batterham, R., et al., *Health literacy: applying current concepts to improve health services and reduce health inequalities.* Public health, 2016. **132**: p. 3-12.
18. Shepard, L.A., *Evaluating test validity: Reprise and progress.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 268-280.
19. Hubley, A.M. and B.D. Zumbo, *A dialectic on validity: Where we have been and where we are going.* The Journal of General Psychology, 1996. **123**(3): p. 207-215.
20. Kelley, T.L., *Interpretation of educational measurements*. Measurement and adjustment series. 1927, Yonkers-on-Hudson, N.Y.: World Book. xiii, 363 p.
21. Sireci, S.G., *On the validity of useless tests.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 226-235.

22. Hawkins, M., G.R. Elsworth, and R.H. Osborne, *Questionnaire validation practice: a protocol for a systematic descriptive literature review of health literacy assessments.* BMJ Open, 2019. **9:e030753**(10).
23. Cizek, G.J., S.L. Rosenberg, and H.H. Koons, *Sources of validity evidence for educational and psychological tests.* Educational and psychological measurement, 2008. **68**(3): p. 397-412.
24. Messick, S., *Validity*, in *Educational Measurement*, R. Linn, Editor. 1989, American Council on Education/Macmillan Publishing Company: New York.
25. Messick, S., *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment.* ETS Research Report Series, 1998: p. 3-20.
26. Cronbach, L.J., *Five perspectives on validity argument*, in *Test validity*, H. Wainer and H.I. Braun, Editors. 1988, Lawrence Erlbaum Associates Inc: New Jersey. p. 3-17.
27. House, E., *Evaluating with validity*. 1980, Beverly Hills, California: Sage Publications.
28. Shepard, L.A., *Evaluating test validity*, in *Review of Research in Education*, L. Darling-Hammond, Editor. 1993, American Educational Research Asociation. p. 405-450.
29. Kane, M.T., *Validating the interpretations and uses of test scores.* Journal of Educational Measurement, 2013. **50**(1): p. 1-73.
30. Kane, M.T., *Validity as the evaluation of the claims based on test scores.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 309-311.
31. Cox, D.W. and J.J. Owen, *Validity evidence for a perceived social support measure in a population health context*, in *Validity and validation in social, behavioral, and health sciences*, B.D. Zumbo and E.K. Chan, Editors. 2014, Springer International Publishing: Switzerland.
32. Zumbo, B.D. and A.M. Hubley, eds. *Understanding and investigating response processes in validation research*. Social Indicators Research Series, ed. A.C. Michalos. Vol. 69. 2017, Springer International Publishing: Switzerland.
33. Hubley, A.M. and B.D. Zumbo, *Response processes in the context of validity: Setting the stage*, in *Understanding and investigating response processes in validation research*, B.D. Zumbo and A.M. Hubley, Editors. 2017, Springer International Publishing: Switzerland. p. 1-12.
34. Onwuegbuzie, A.J. and N.L. Leech, *Validity and qualitative research: An oxymoron?* Quality and Quantity, 2007. **41**(2): p. 233-249.
35. Onwuegbuzie, A.J. and N.L. Leech, *On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies.* International journal of social research methodology, 2005. **8**(5): p. 375-387.
36. Castillo-Díaz, M. and J.-L. Padilla, *How cognitive interviewing can provide validity evidence of the response processes to scale items.* Social indicators research, 2013. **114**(3): p. 963-975.
37. Padilla, J.-L. and I. Benítez, *Validity evidence based on response processes.* Psicothema, 2014. **26**(1): p. 136-144.
38. Padilla, J.-L., I. Benítez, and M. Castillo, *Obtaining validity evidence by cognitive interviewing to interpret psychometric results.* Methodology, 2013. **9**(3): p. 113-122.
39. Moss, P.A., *The role of consequences in validity theory.* Educational Measurement: Issues and Practice, 1998. **17**(2): p. 6-12.
40. Hubley, A.M. and B.D. Zumbo, *Validity and the consequences of test interpretation and use.* Social Indicators Research, 2011. **103**(2): p. 219.
41. Zumbo, B.D. and A.M. Hubley, *Bringing consequences and side effects of testing and assessment to the foreground.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 299-303.
42. Kane, M. and R. Mislevy, *Validating score interpretations based on response processes*, in *Validation of score meaning for the next generation of assessments*, K. Ercikan and J.W. Pellegrino, Editors. 2017, Routledge: New York. p. 11-24.
43. Terwee, C.B., et al., *Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist.* Quality of Life Research, 2012. **21**(4): p. 651-657.

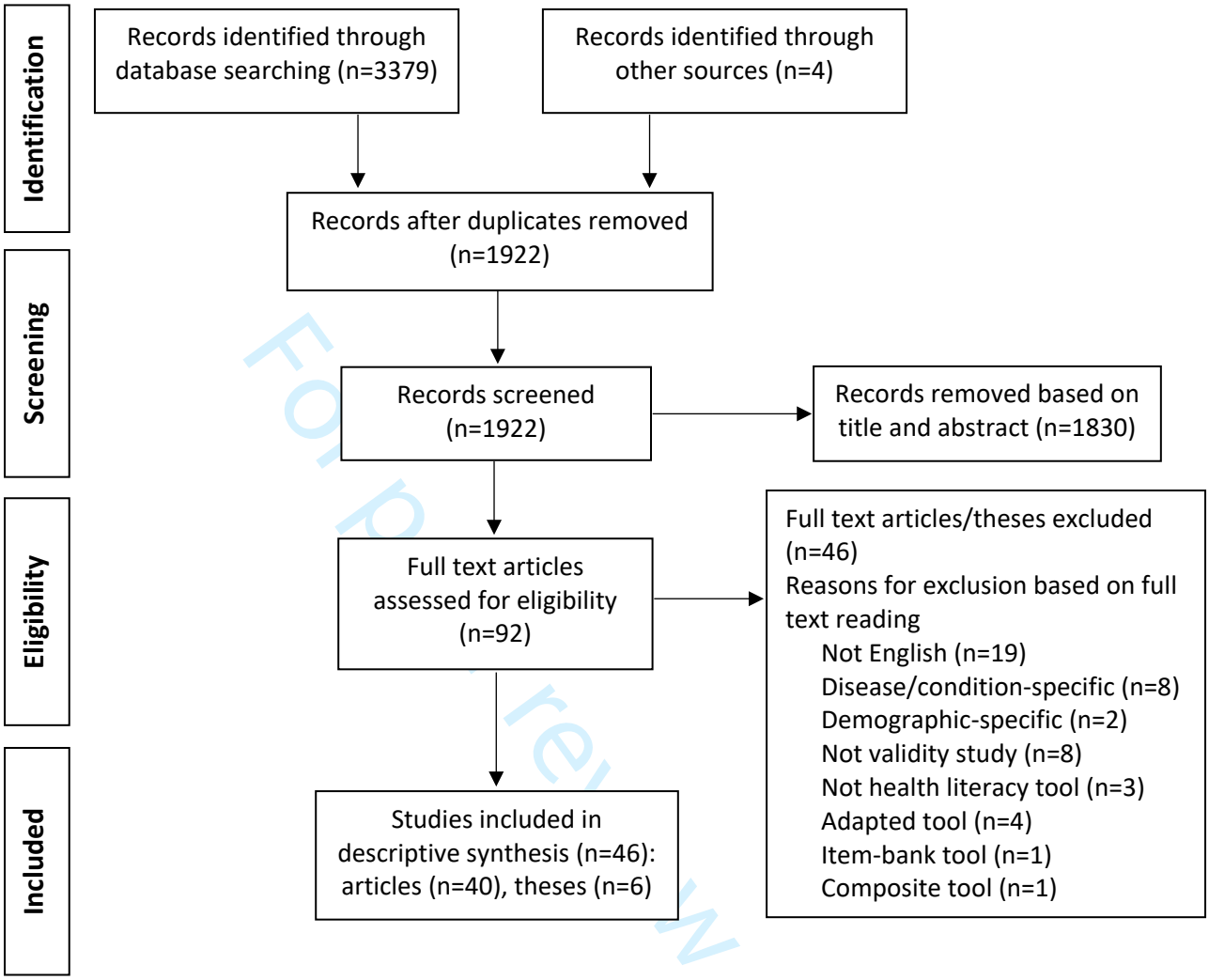
44. Devellis, R.F., *A consumer's guide to finding, evaluating, and reporting on measurement instruments.* Arthritis and Rheumatism, 1996. **9**(3): p. 239-245.
45. Lohr, K.N., *Assessing health status and quality-of-life instruments: attributes and review criteria.* Quality of Life Research, 2002. **11**(3): p. 193-205.
46. Nutbeam, D., *The evolving concept of health literacy.* Soc Sci Med, 2008. **67**(12): p. 2072-8.
47. Sørensen, K., et al., *Health literacy and public health: a systematic review and integration of definitions and models.* BMC Public Health, 2012. **12**: p. 80.
48. Sykes, S., et al., *Understanding critical health literacy: a concept analysis.* BMC Public Health, 2013. **13**(1): p. 150.
49. Pleasant, A., J. McKinney, and R. Rikard, *Health literacy measurement: a proposed research agenda.* Journal of Health Communication, 2011. **16**(sup3): p. 11-21.
50. Jordan, J.E., R.H. Osborne, and R. Buchbinder, *Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses.* J Clin Epidemiol, 2010.
51. Altin, S.V., et al., *The evolution of health literacy assessment tools: a systematic review.* BMC public health, 2014. **14**(1): p. 1207.
52. McCormack, L., et al., *Recommendations for advancing health literacy measurement.* Journal of Health Communication, 2013. **18**(sup1): p. 9-14.
53. Mancuso, J.M., *Assessment and measurement of health literacy: an integrative review of the literature.* Nursing & health sciences, 2009. **11**(1): p. 77-89.
54. Haun, J.N., et al., *Health literacy measurement: an inventory and descriptive summary of 51 instruments.* J Health Commun, 2014. **19 Suppl 2**: p. 302-33.
55. Guzys, D., et al., *A critical review of population health literacy assessment.* BMC Public Health, 2015. **15**(1): p. 215.
56. Barry, A.E., et al., *Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals.* Health Education & Behavior, 2014. **41**(1): p. 12-18.
57. Davis, T.C., et al., *Rapid assessment of literacy levels of adult primary care patients.* Family medicine, 1991. **23**(6): p. 433-435.
58. Weiss, B.D., et al., *Quick assessment of literacy in primary care: the newest vital sign.* The Annals of Family Medicine, 2005. **3**(6): p. 514-522.
59. Jessup, R.L., et al., *Using co-design to develop interventions to address health literacy needs in a hospitalised population.* BMC Health Services Research, 2018. **18**(1): p. 989.
60. Osborne, R.H., et al., *The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ).* BMC Public Health, 2013. **13**: p. 658.
61. Jessup, R.L. and R. Buchbinder, *What if I cannot choose wisely? Addressing suboptimal health literacy in our patients to reduce over-diagnosis and overtreatment.* Internal Medicine Journal, 2018. **48**(9): p. 1154-1157.
62. Roberts, J., *Local action on health inequalities: Improving health literacy to reduce health inequalities*. 2015, UCL Institute of Health Equity: London.
63. Batterham, R.W., et al., *The OPtimising HEalth LIterAcy (Ophelia) process: study protocol for using health literacy profiling and community engagement to create and implement health reform.* BMC Public Health, 2014. **14**(1): p. 694-703.
64. Beauchamp, A., et al., *Systematic development and implementation of interventions to Optimise Health Literacy and Access (Ophelia).* BMC Public Health, 2017. **17**(1): p. 230.
65. Barry, M.M., M. D'Eath, and J. Sixsmith, *Interventions for Improving Population Health Literacy: Insights From a Rapid Review of the Evidence.* Journal of Health Communication, 2013. **18**(12): p. 1507-1522.
66. Bakker, M.M., et al., *Acting together–WHO National Health Literacy Demonstration Projects (NHLDPs) address health literacy needs in the European Region.* Public Health Panorama, 2019. **5**(2-3): p. 233-243.
67. Australian Bureau of Statistics. *National Health Survey: Health Literacy, 2018*. 2019 14 October 2019]; Available from: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4364.0.55.014Main+Features12018?OpenDocument.

68. Trezona, A., G. Rowlands, and D. Nutbeam, *Progress in implementing national policies and strategies for health literacy—What have we learned so far?* International Journal of Environmental Research and Public Health, 2018. **15**(7): p. 1554.
69. WHO Regional Office for Europe, *WHO Health Evidence Network synthesis report 65. What is the evidence on the methods, frameworks and indicators used to evaluate health literacy policies, programmes and interventions at the regional, national and organizational levels?* 2019: Copenhagen.
70. Putoni, S., *Health Literacy in Wales - A Scoping Document for Wales*. 2010, Welsh Assembly Government: Wales.
71. Scottish Government NHS Scotland, *Making it Easier: A Health Literacy Action Plan for Scotland 2017-2025*. 2017: Edinburgh.
72. King, W.R. and J. He, *Understanding the role and methods of meta-analysis in IS research.* Communications of the Association for Information Systems, 2005. **16**(1): p. 32.
73. Yang, H. and M. Tate, *A descriptive literature review and classification of cloud computing research.* Communications of the Association for Information Systems, 2012. **31**: p. 2.
74. Schlagenhaufer, C. and M. Amberg. *A Descriptive Literature Review and Classification Framework for Gamification in Information Systems*. in *European Conference on Information Systems*. 2015. Germany: Gartner.
75. Roter, D.L., J.A. Hall, and N.R. Katz, *Patient-physician communication: a descriptive summary of the literature.* Patient Education and Counseling, 1988. **12**(2): p. 99-119.
76. Guzzo, R.A., S.E. Jackson, and R.A. Katzell, *Meta-analysis analysis*, in *Research in Organizational Behavior*, L.L. Cummings and B.M. Staw, Editors. 1987, JAI Press: Greenwich, CT. p. 407-442.
77. Paré, G., et al., *Synthesizing information systems knowledge: A typology of literature reviews.* Information and Management, 2015. **52**(2): p. 183-199.
78. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS medicine, 2009. **6**(7): p. e1000097.
79. Davis, T.C., et al., *Rapid estimate of adult literacy in medicine: a shortened screening instrument.* 1993. **25**(6): p. 391-395.
80. Elder, C., et al., *Assessing health literacy: A new domain for collaboration between language testers and health professionals.* Language Assessment Quarterly, 2012. **9**(3): p. 205-224.
81. Dumenci, L., et al., *On the Validity of the Shortened Rapid Estimate of Adult Literacy in Medicine (REALM) Scale as a Measure of Health Literacy.* Communication Methods and Measures, 2013. **7**(2): p. 134-143.
82. Barber, M.N., et al., *Up to a quarter of the Australian population may have suboptimal health literacy depending upon the measurement tool: results from a population-based survey.* Health Promot Int, 2009. **24**(3): p. 252-61.
83. Wallace, L.S., et al., *Brief report: screening items to identify patients with limited health literacy skills.* Journal of General Internal Medicine, 2006. **21**(8): p. 874-877.
84. Parker, R.M., et al., *The test of functional health literacy in adults.* Journal of General Internal Medicine, 1995. **10**(10): p. 537-541.
85. Curtis, L.M., et al., *Development and validation of the comprehensive health activities scale: a new approach to health literacy measurement.* Journal of Health Communication, 2015. **20**(2): p. 157-164.
86. Dykhuis, K.E., et al., *A New Measure of Health Numeracy: Brief Medical Numbers Test (BMNT).* Psychosomatics, 2019. **60**(3): p. 271-277.
87. Jordan, J.E., et al., *The Health Literacy Management Scale (HeLMS): A measure of an individual's capacity to seek, understand and use health information within the healthcare setting.* Patient education and counseling, 2013. **91**(2): p. 228-235.
88. Zhang, X.-H., et al., *Development and validation of a functional health literacy test.* The Patient: Patient-Centered Outcomes Research, 2009. **2**(3): p. 169-178.
89. Harper, R., *Comprehensive health literacy assessment for college students*, in *Department of Journalism and Technical Communication*. 2013, Colorado State University: Fort Collins, Colorado.

90. Bann, C.M., et al., *The health literacy skills instrument: a 10-item short form.* Journal of Health Communication, 2012. **17**(sup3): p. 191-202.
91. McCormack, L., et al., *Measuring health literacy: a pilot study of a new skills-based instrument.* Journal of Health Communication, 2010. **15**(S2): p. 51-71.
92. DeBello, M.C., *The development and psychometric testing of the health literacy knowledge, application, and confidence scale (HLKACS)*, in *College of Education and College of Nursing*. 2016, Eastern Michigan University: Michigan.
93. Baker, D.W., et al., *Development of a brief test to measure functional health literacy.* Patient Education and Counseling, 1999. **38**(1): p. 33-42.
94. Shaw, T.C., *Uncovering health literacy: Developing a remotely administered questionnaire for determining health literacy levels in health disparate populations.* Journal of Hospital Administration, 2014. **3**(4): p. 140.
95. Begoray, D.L. and B. Kwan, *A Canadian exploratory study to define a measure of health literacy.* Health Promotion International, 2011. **27**(1): p. 23-32.
96. Chew, L.D., K.A. Bradley, and E.J. Boyko, *Brief questions to identify patients with inadequate health literacy.* Family Medicine, 2004. **11**: p. 12.
97. Chesser, A.K., et al., *Health literacy assessment of the STOFHLA: paper versus electronic administration continuation study.* Health Education & Behavior, 2014. **41**(1): p. 19-24.
98. Dageforde, L.A., et al., *Validation of the written administration of the short literacy survey.* Journal of Health Communication, 2015. **20**(7): p. 835-842.
99. Sørensen, K., et al., *Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q).* BMC Public Health, 2013. **13**(1): p. 1-22.
100. Hawkins, M., et al., *The Health Literacy Questionnaire (HLQ) at the patient-clinician interface: a qualitative study of what patients and clinicians mean by their HLQ scores.* BMC Health Services Research, 2017. **17**(1): p. 309.
101. Morris, R.L., et al., *Measurement properties of the Health Literacy Questionnaire (HLQ) among older adults who present to the emergency department after a fall: a Rasch analysis.* BMC health services research, 2017. **17**(1): p. 605.
102. Sand-Jecklin, K. and S. Coyle, *Efficiently assessing patient health literacy: the BHLS instrument.* Clinical Nursing Research, 2014. **23**(6): p. 581-600.
103. Haun, J., et al., *Measurement variation across health literacy assessments: implications for assessment selection in research and practice.* Journal of Health Communication, 2012. **17**(sup3): p. 141-159.
104. Miller, B., *Investigating the Construct of Health Literacy Assessment: A Cross-Validation Approach*, in *Graduate School, the College of Education and Psychology and the Department of Educational Research & Administration*. 2018, The University of Southern Mississippi: Ann Arbor, MI.
105. Elsworth, G.R., A. Beauchamp, and R.H. Osborne, *Measuring health literacy in community agencies: a Bayesian study of the factor structure and measurement invariance of the health literacy questionnaire (HLQ).* BMC Health Serv Res, 2016. **16**(1): p. 508.
106. Goodwin, B.C., et al., *Health literacy and the health status of men with prostate cancer.* Psycho-Oncology, 2018. **27**(10): p. 2374-2381.
107. Driessnack, M., et al., *Using the "Newest Vital Sign" to assess health literacy in children.* Journal of Pediatric Health Care, 2014. **28**(2): p. 165-171.
108. Goodman, M.S., et al., *Do subjective measures improve the ability to identify limited health literacy in a clinical setting?* Journal of the American Board of Family Medicine, 2015. **28**(5): p. 584-594.
109. Cavanaugh, K.L., et al., *Performance of a brief survey to assess health literacy in patients receiving hemodialysis.* Clinical Kidney Journal, 2015. **8**(4): p. 462-468.
110. Chew, L.D., et al., *Validation of screening questions for limited health literacy in a large VA outpatient population.* Journal of General Internal Medicine, 2008. **23**(5): p. 561-566.
111. Housten, A.J., et al., *Limitations of the S-TOFHLA in measuring poor numeracy: a cross-sectional study.* BMC Public Health, 2018. **18**(1): p. 405.

112. Kirk, J.K., et al., *Performance of health literacy tests among older adults with diabetes.* Journal of General Internal Medicine, 2012. **27**(5): p. 534-540.
113. Ko, Y., et al., *Development and validation of a general health literacy test in Singapore.* Health Promotion International, 2011. **27**(1): p. 45-51.
114. Kordovski, V.M., et al., *Is the newest vital sign a useful measure of health literacy in HIV disease?* Journal of the International Association of Providers of AIDS Care, 2017. **16**(6): p. 595-602.
115. McNaughton, C., et al., *Short, subjective measures of numeracy and general health literacy in an adult emergency department.* Academic Emergency Medicine, 2011. **18**(11): p. 1148-1155.
116. Morris, N.S., et al., *The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability.* BMC Family Practice, 2006. **7**(1): p. 21.
117. Quinzanos, I., et al., *Cross-sectional correlation of single-item health literacy screening questions with established measures of health literacy in patients with rheumatoid arthritis.* Rheumatology International, 2015. **35**(9): p. 1497-1502.
118. Rawson, K.A., et al., *The METER: a brief, self-administered measure of health literacy.* Journal of General Internal Medicine, 2010. **25**(1): p. 67-71.
119. Wallston, K.A., et al., *Psychometric properties of the brief health literacy screen in clinical practice.* J Gen Intern Med, 2014. **29**(1): p. 119-26.
120. Hadden, K.B., *Health Literacy and Pregnancy: Validation of a New Measure and Relationships of Health Literacy to Pregnancy Risk Factors*. 2012, University of Arkansas for Medical Sciences: Arkansas.
121. Soelberg, J., *Determining the reliability and validity of the newest vitalsign in the inpatient setting*. 2015, Rush University: Chicago, Illinois.
122. Haun, J.N., *Health Literacy: The Validation of a Short Form Health Literacy Screening Assessment in an Ambulatory Care Setting*. 2007, University of Florida: Florida.
123. Messick, S., *Foundations of validity: Meaning and consequences in psychological assessment.* ETS Research Report Series, 1993. **1993**(2): p. i-18.
124. Francis, D.O., et al., *Checklist to operationalize measurement characteristics of patient-reported outcome measures.* Systematic Reviews, 2016. **5**(1): p. 129.
125. American Educational Research Association, et al., *Standards for educational and psychological testing*. 1999, Washington, DC: American Educational Research Association.
126. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing*. 1985: American Educational Research Association.
127. Buchbinder, R., et al., *A validity-driven approach to the understanding of the personal and societal burden of low back pain: development of a conceptual and measurement model.* Arthritis Research & Therapy, 2011. **13**(5): p. R152.
128. McClimans, L., *Interpretability, validity, and the minimum important difference.* Theoretical Medicine and Bioethics 2011. **32**(6): p. 389-401.
129. Kane, M. and B. Bridgeman, *Research on Validity Theory and Practice at ETS*, in *Advancing Human Assessment: The Methodological, Psychological and Policy Contribution of ETS*, R.E. Bennett and M. von Davier, Editors. 2017, Springer Nature: Cham, Switzerland. p. 489-552.
130. Landy, F.J., *Stamp collecting versus science: Validation as hypothesis testing.* American Psychologist, 1986. **41**(11): p. 1183.
131. Spurgeon, S.L., *Evaluating the unintended consequences of assessment practices: Construct irrelevance and construct underrepresentation.* Measurement and Evaluation in Counseling and Development, 2017. **50**(4): p. 275-281.
132. Nutbeam, D., *Health promotion glossary.* Health Promotion International, 1998. **13**(4): p. 349-364.
133. New Zealand Ministry of Health, *Content Guide 2017/18: New Zealand Health Survey*. 2018, NZ Ministry of Health: Wellington, New Zealand.

134. Bo, A., et al., *National indicators of health literacy: ability to understand health information and to engage actively with healthcare providers - a population-based survey among Danish adults.* BMC Public Health, 2014. **14**(1): p. 1095.
135. Busija, L., R. Buchbinder, and R.H. Osborne, *A grounded patient-centered approach generated the personal and societal burden of osteoarthritis model.* Journal of Clinical Epidemiology, 2013. **66**(9): p. 994-1005.
136. Rosas, S.R. and J.W. Ridings, *The use of concept mapping in measurement development and evaluation: application and future directions.* Evaluation and Program Planning, 2017. **60**: p. 265-276.
137. Soellner, R., N. Lenartz, and G. Rudinger, *Concept mapping as an approach for expert-guided model building: The example of health literacy.* Evaluation and Program Planning, 2017. **60**: p. 245-253.
138. WHO Regional Office for Europe. *Health literacy in action*. 2019 [cited 2019 18 October]; Available from: http://www.euro.who.int/en/health-topics/disease-prevention/health-literacy/health-literacy-in-action.
139. Nguyen, T.H., et al., *State of the science of health literacy measures: Validity implications for minority populations.* Patient Educ Couns, 2015.
140. Marmot, M., *Fair society, healthy lives: the Marmot Review: strategic review of health inequalities in England Post-2010.* 2010.
141. Kane, M., *Validity and fairness.* Language testing, 2010. **27**(2): p. 177-182.
142. Carey, G., B. Crammond, and E. De Leeuw, *Towards health equity: a framework for the application of proportionate universalism.* International Journal for Equity in Health, 2015. **14**(1): p. 81.
143. Beauchamp, A., et al., *The effect of obesity prevention interventions according to socioeconomic position: a systematic review.* Obes Rev, 2014. **15**(7): p. 541-54.
144. Beeston, C., et al., *Health inequalities policy review for the Scottish Ministerial Task Force on health inequalities*. 2014: Edinburgh.
145. Addison, M., et al., *Equal North: how can we reduce health inequalities in the North of England? A prioritization exercise with researchers, policymakers and practitioners.* Journal of Public Health, 2018: p. 1-13.
146. Capewell, S. and H. Graham, *Will cardiovascular disease prevention widen health inequalities?* PLoS Medicine, 2010. **7**(8): p. e1000320.
147. Messick, S., *Test validity: A matter of consequence.* Social Indicators Research, 1998. **45**(1-3): p. 35-44.



MY

EBSCOhost

Monday, March 11, 2019 7:57:11 PM

#	Query	Limiters/Expanders	Last Run Via	Results
S28	S24 AND S25 AND S26 AND S27	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	1,036
S27	S12 OR S13 OR S14 OR S15 OR S16 OR S22 OR S23	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	3,396,491
S26	S11 OR S21	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	68,560
S25	S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S18 OR S19 OR S20	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	5,965,966
S24	S1 OR S2 OR S17	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	813,727
S23	(MH "Focus Groups") OR (MH "Interviews as Topic") OR (MH "Data Accuracy")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	79,811
S22	(MH "Psychometrics")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced	69,650

			Search Database - MEDLINE Complete	
S21	(MH "Health Literacy") OR (MH "Health Education") OR (MH "Consumer Health Information")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	65,418
S20	(MH "Self-Assessment")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	11,920
S19	(MH "Health Surveys") OR (MH "Surveys and Questionnaires") OR (MH "Health Care Surveys")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	486,718
S18	(MH "Patient Outcome Assessment") OR (MH "Self Report") OR (MH "Patient Reported Outcome Measures")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	31,421
S17	(MH "Validation Studies as Topic")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	1,977
S16	TI “focus group*” OR AB “focus group*”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	36,147
S15	TI “think aloud” OR AB “think aloud”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	1,091

S14	TI interview* OR AB interview*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	433,189
S13	TI Psychometric* OR AB Psychometric*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	32,322
S12	TI measur* OR AB measur*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	2,602,779
S11	TI “health literacy” OR AB “health literacy”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	4,293
S10	TI tool* OR AB tool*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	682,713
S9	TI test* OR AB test*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	2,261,528
S8	TI assess* OR AB assess*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	1,782,492
S7	TI “self rated" OR AB “self rated"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced	7,786

			Search Database - Academic Search Complete	
S6	TI “self report*” OR AB “self report*”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	92,555
S5	TI survey* OR AB survey*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	590,730
S4	TI questionnaire* OR AB questionnaire*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	287,547
S3	TI “patient reported outcome*” OR AB “patient reported outcome*”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	7,191
S2	TI Verif* OR AB Verif*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	276,503
S1	TI valid* OR AB valid*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	639,560

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1, 4
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2. This systematic review is not registered
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	5-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Protocol paper: https://bmjopen.bmj.com/content/9/10/e030753
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6-7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6-7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Supplementary file 1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6-8 including Figure 1.
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	7-8
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7-8
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	3, 16
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7 (types of validity evidence)

PRISMA 2009 Checklist

Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	7-8

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	3, 16
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	7-8
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7-8
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	8-9
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	3, 16
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	8-13
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	8-13
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	3, 16
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8-13
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	13-16
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16-17
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	17
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

PRISMA 2009 Checklist

Supplementary file 3: Data extraction details

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Baker et al (1999)	USA	Develop and test S-TOFHLA	REALM	x	Cronbach's alpha	1	x	x	1	x
Bann et al (2012)	USA	Develop and test HLSI-SF	S-TOFHLA, self-report questions	x	Cronbach's alpha	3	x	2	1	x
Barber et al (2009)	Australia	Test REALM, TOFHLA, NVS	AQOL	x	x	x	x	x	3	x
Begoray (2012)	Canada	Develop and test health literacy assessment (no name) (9 s-r items, 2 Cloze)	REALM	Indirect to Standards [Hubley and Zumbo 1996]	Cronbach's alpha	1	x	x	2	x
Cavanaugh et al (2015)	USA	Test BHLS (3 items)	REALM, S-TOFHLA, MMSE / CHeKS, PiKS	x	Cronbach's alpha	x	x	x	6	x
Chesser et al (2014)	USA	Test S-TOFHLA	x	x	x	1	1	x	x	x
Chew et al 2004	USA	Develop and test 3 screening questions	S-TOFHLA	x	x	2	x	x	1	x

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Chew et al 2008	USA	Test 3 screening questions	S-TOFHLA, REALM	x	x	x	x	x	1	x
Curtis et al (2015)	USA	Develop and test CHAS	S-TOFHLA, REALM, NVS, MMSE / self-reported health status, SF-36, PROMIS short form emotional health	x	Cronbach's alpha; IRT TIF; Omega analysis	2	x	3	6	x
Dageforde et al (2015)	USA	Test SLS (3 items)	REALM, S-TOFHLA	x	Cronbach's alpha; Wilcoxen signed rank test	1	x	x	2	x
Davis et al (1991)	USA	Test REALM	PIAT-R, SORT	x	Test-retest; Inter-rater	1	x	x	1	x
Davis et al (1993)	USA	Develop and test REALM-SF	PIAT-R, SORT-R, WRAT-R	x	x	1	x	x	1	x

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
DeBello (2016) thesis	USA	Test HKACS	x	Indirect to Standards [Waltz, Strictland and Lenz 2010]	Cronbach's alpha; Test-retest	2	x	1	4	x
Driessnack et al (2014)	USA	Test NVS	N of children's books	x	Cronbach's alpha	x	x	x	3	x
Dykhuis et al (2019)	USA	Test BMNT	REALM-R, REALM-SF	Direct to Francis et al (2016) checklist	Cronbach's alpha	1	x	x	2	x
Elder et al (2012)	Australia	Test REALM (13 items)	TOFHLA, NVS, Definition scores / AQOL	Indirect to Standards [p.206: Abedi et al (2004) and LaCelle-Peterson & Rivera (1994)]	Cronbach's alpha; Inter-rater	3	x	x	4	1
Elsworth et al (2016)	Australia	Test HLQ	x	Direct to Messick 1992 In Alkin MC	Cronbach's alpha, Composite reliability	x	x	2	2	x
Goodman et al (2015)	USA	Test BHLS (3 items)	REALM-R, NVS	x	x	x	x	x	1	x

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Goodwin et al (2018)	Australia	HLQ	x	Indirect to Standards [Buchbinder et al 2011 & Hawkins et al 2018]	Cronbach's alpha	x	x	1	1	x
Hadden (2012) thesis	USA	Test HLSI-SF	S-TOFHLA, Perceptions of Difficulty with Health Literacy Skills	x	x	x	x	x	5	x
Harper (2013) thesis	USA	Develop and test a new health literacy assessment [no name]	S-TOFHLA	x	Cronbach's alpha	4	x	2	3	x
Haun (2012)	USA	Test S-TOFHLA, REALM, BRIEF (4 items)	x	x	Cronbach's alpha	x	x	1	4	x
Haun (2007) thesis	USA	Test BRIEF (4 items)	S-TOFHLA and REALM	x	Cronbach's alpha	x	x	x	3	x
Hawkins et al (2017)	Australia	Test HLQ	x	Direct ref to Standards 2014, Kane 1992, Messick 1993	Inter-rater	1	2	x	x	x

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Housten et al (2018)	USA	Test S-TOFHLA	SNS, GL (GL1, GL2, GL3)	x	x	x	x	x	1	x
Jordan et al (2013)	Australia	Develop and test HeLMS	x	x	Cronbach's alpha; Test-retest	3	1	3	2	x
Kirk et al (2012)	USA	Test REALM-SF, NVS	S-TOFHLA	x	x	x	x	x	2	x
Ko et al (2012)	Singapore	Develop and test HLTS	NVS	x	Cronbach's alpha	x	x	x	4	x
Kordovski et al (2017)	USA	Test NVS	REALM, SILS	x	Cronbach's alpha	x	x	x	7	x
McCormack et al (2010)	USA	Develop and test HLSI	S-TOFHLA, self-report questions	x	Cronbach's alpha	3	x	2	3	x
McNaughton et al (2011)	USA	Test SLS (3 items) and SNS (8 items)	S-TOFHLA, REALM, WRAT4	x	Cronbach's alpha	x	x	x	3	x
Miller (2018) thesis	USA	Test HLSI (Cloze only), NVS, S-TOFHLA	x	x	Cronbach's alpha	x	x	1	1	x
Morris et al (2006)	USA	Test SILS 1	S-TOFHLA	x	x	x	x	x	1	x
Morris et al (2017)	Australia	Test HLQ	x	Indirectly to Standards [Hawkins et al 2017]	Person separation index (PSI) in IRT	x	x	3	x	x

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Osborne et al (2013)	Australia	Develop and test HLQ	x	Indirectly to Standards [Buchbinder et al 2011]	Composite reliability	7	1	3	1	x
Parker et al (1995)	USA	Develop and test TOFHLA	WRAT-R, REALM	x	Cronbach's alpha; Split halves coefficient	1	x	x	1	x
Quinzanos et al (2015)	USA	test SILS 1 and SILS 2	REALM and S-TOFHLA	x	x	x	x	x	1	x
Rawson et al (2010)	USA	Develop and test METER	REALM	x	Cronbach's alpha	x	x	x	2	x
Sand-Jecklin et al (2014)	USA	Develop and test BHLS (5 items)	S-TOFHLA	x	Cronbach's alpha	x	x	2	4	x
Shaw et al (2014)	USA	Develop and test remote admin health literacy assessment	S-TOFHLA	x	x	1	x	x	1	x
Soelberg (2015)	USA	Test NVS	S-TOFHLA	x	Cronbach's alpha	x	x	x	2	x
Sørensen et al (2013)	Netherlands	Develop and test HLS-EU-Q	x	x	Cronbach's alpha	7	1	1	x	x
Wallace et al (2006)	USA	Test 3 screening questions	REALM	x	x	x	x	x	2	x

Author	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reference to validity testing framework	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Wallston et al (2014)	USA	Test BHLS (3 items)	S-TOFHLA	x	Cronbach's alpha; Inter-rater	x	x	x	4	x
Weiss et al (2005)	USA	Develop and test NVS	TOFHLA	x	Cronbach's alpha	3	x	x	3	x
Zhang et al (2009)	Singapore	Develop and test FHLT (21 items)	REALM	x	Cronbach's alpha; Test-retest	3	1	1	5	x
Totals						**52**	**7**	**28**	**107**	**1**

Supplementary file 4: Detail of data extraction framework

Data were extracted in Excel. These are the data extraction category headings from the Excel spreadsheet.

1. Evidence based on test content

1. Test content evaluated: yes/no/unclear

1. Test content: literature review

1. Test content: prior existing measures of the construct

1. Test content: expert review

1. Test content: participant involvement in construct / item development - structured workshops, concept mapping

1. Test content: participant involvement in construct / item development - interviews

1. Test content: participant feedback processes about items

1. Test content: construct description (incl. high/low descriptors)

1. Test content: item intent descriptions

1. Test content: examination of administration methods

1. Test content: other method (e.g., Item difficulty)

2. Evidence based on response processes

2. Response processes evaluated: yes/no/unclear

2. Response processes - respondents: cognitive interviews

2. Response processes - respondents: think aloud protocols

2. Response processes - respondents: recording and timing responses to items

2. Response processes - users: cognitive interviews

2. Response processes - users: think aloud protocols

2. Response processes - users: recording and timing responses to items

2. Response processes: other method (e.g., determining construct irrelevant factors and construct underrepresentation)

3. Evidence based on internal structure

3. Internal structure evaluated: yes/no/unclear

3. Internal structure: exploratory factor analysis (EFA)

3. Internal structure: confirmatory factor analysis (CFA)

3. Internal structure: multi-group factor analysis (MGFA) (SEM, measurement invariance)

3. Internal structure: correlation patterns and multi-trait scaling analysis (inter-item, item-total and item-remainder correlations)

3. Internal structure: differential item functioning (DIF)

3. Internal structure: other method

4. Evidence based on relations to other variables

4. Relations to other variables evaluated: yes/no/unclear

4. Relations to other variables: convergent validity (between measures of the same or similar construct)

4. Relations to other variables: discriminant validity

4. Relations to other variables: test-criterion relationships (how accurately test scores predict criterion performance)

4. Relations to other variables: group differences (relationships with background characteristics such as demographics information)

4. Relations to other variables: validity generalisation (e.g., meta-analyses / statistical summaries of past studies; cumulative databases)

4. Relations to other variables: nomological networks

4. Relations to other variables: other method

5. Evidence based on validity and the consequences of testing

5. Consequences of testing evaluated: yes/no/unclear

5. Consequences of testing: methods to test for consequential validity (intended consequences e.g., benefits)

5. Consequences of testing: methods to test for consequential validity (unintended consequences e.g., negative effects)

5. Consequences of testing: methods to test for construct underrepresentation

5. Consequences of testing: methods to test for construct-irrelevant components

5. Consequences of testing: methods to test for claims made beyond the intended score interpretation

5. Consequences of testing: methods to test for consequences for clinical implications

5. Consequences of testing: other methods to test consequential validity

5. Consequences of testing: other method (e.g., fairness - low/high-stakes consequences)

Supplementary file 5 – Supplementary Tables 1 to 4

Supplementary Table 1. Evidence based on test content

Number of instances of evidence based on test content across all studies		
Method to generate evidence		
Literature review	4	8%
Existing measures of the construct	8	15%
Expert review	14	27%
Participant involvement:		
Concept mapping	3	6%
Interviews	2	4%
Participant feedback processes about items	4	8%
Construct descriptions (e.g., high/low)	4	8%
Item intent descriptions	1	2%
Examination of administration methods	3	6%
Other method (e.g., item difficulty):		
Item difficulty	5	10%
Items tested against item intents	1	2%
IRT analysis for item selection within domains	1	2%
Item selection based on hospital medical texts	1	2%
Item selection based on HL conceptual model	1	2%
Total instances of evidence based on test content	*52*	*100%*

Supplementary Table 2. Evidence based on response processes

Number of instances of evidence based on response processes across all studies		
Method to generate evidence		
With respondents:		
Cognitive interviews	3	43%
Recording and timing responses to items	3	43%
With users:		
Cognitive interviews	1	14%
Total instances of evidence based on response processes	*7*	*100%*

Supplementary Table 3. Evidence based on internal structure

Number of instances of evidence based on internal structure across all studies		
Method to generate evidence		
Exploratory factor analysis (incl. PCA*)	7	25%
Confirmatory factory analysis (incl. IRT** item discriminations)	7	25%
Multi-group factor analysis	1	4%
Correlation patterns / multi-trait scaling analysis:		
Tetrachoric correlations	1	4%

Inter-item correlations	1	4%
Item-total correlations	1	4%
Item-remainder correlations	2	7%
Differential item functioning	3	11%
Other method:		
Very Simple Structure	1	4%
Velicer's Minimum Average partial criterion	1	4%
Rasch analysis (overall fit, individual person/item fit)	1	4%
Intra-factor correlations	1	4%
IRT for item discriminations	1	4%
Total instances of evidence based on response processes	*28*	*100%*

*PCA = principal component analysis; **IRT = item response theory

Supplementary Table 4. Evidence based on relations to other variables

Summary of number of instances of evidence based on relations to other variables across all studies		
Type of evidence		
Convergent evidence	57	53%
Discriminant evidence	3	3%
Criterion-referenced evidence	17	16%
Evidence for group differences	30	28%
Evidence for generalisation	0	0%
Total instances of evidence based on relations to other variables	107	100%
Number of instances of evidence based on relations to other variables across all studies		
Convergent evidence (relationships between items and scales of the same or similar structure) (n=38 studies):		
Spearman's correlation coefficient	11	19%
Pearson correlation coefficient	11	19%
Linear regression models	5	9%
Logistic regression models	2	4%
Receiver operating characteristic / Area under the ROC (AUROC)	11	19%
Wilcoxen signed rank test	2	4%
Cross tabulations / calculated agreement and disagreement	2	4%
Goodman-Kruskal gamma correlation	1	2%
Bland-Altman plots	1	2%
Cohen's Kappa	1	2%
Sensitivity and specificity	1	2%
Stratum-specific likelihood ratios	1	2%
Unnamed / unclear correlation calculations with similar measures	8	14%
Total instances of convergent evidence	*57*	*100%*
Discriminant evidence (measures of different constructs are sufficiently uncorrelated) (n=2 studies)		
Comparison of AVE and shared variance between HLQ scales	1	33%
Pearson correlation coefficient	1	33%%
Multiscale factor analysis	1	33%
Total instances of discriminant evidence	*3*	*100%*

Criterion-referenced evidence (how accurately test scores predict criterion performance) (n=9 studies):		
Spearman's correlation coefficient	2	12%
Pearson correlation coefficient	1	6%
Linear regression models	6	35%
Logistic regression models	2	12%
ROC/AUROC	1	6%
Chi-squared test of independence	3	18%
ANOVA	1	6%
Cohen's d	1	6%
Total instances of criterion-referenced evidence	*17*	*100%*
Evidence for group differences (relationships of test scores with background characteristics such as demographic information) (n=19 studies):	**N**	**%**
Linear regression models	4	13%
Logistic regression models	3	10%
Univariate associations	1	3%
Spearman's correlation coefficient	1	3%
Chi-squared test	3	10%
Analysis of variance (ANOVA)	5	17%
Analysis of covariance (ANCOVA)	1	3%
Cross tabulations	1	3%
Area under the ROC (AUROC)	1	3%
Kruskal-Wallis test	1	3%
Mann-Whitney U test	2	7%
Goodman-Kruskal gamma correlation	1	3%
Independent sample t-test	3	10%
Exploratory partial correlation analysis	1	3%
Bayesian fit statistics	1	3%
Descriptive statistics (sub-group differences)	1	3%
Total instances of evidence of group differences	*30*	*100%*
Evidence for generalisation (degree to which evidence can be generalised to a new situation) (n=0 studies):	**N**	**%**
Only research synthesis-type studies - see validity generalisation in the *Standards*.	0	0%

BMJ Open

Questionnaire validation practice within a theoretical framework: a systematic descriptive literature review of health literacy assessments

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-035974.R1
Article Type:	Original research
Date Submitted by the Author:	09-Mar-2020
Complete List of Authors:	Hawkins, Melanie; Deakin University, Faculty of Health Elsworth, Gerald; Deakin University School of Health and Social Development, Faculty of Health; Swinburne University of Technology, Centre for Global Health and Equity, Faculty of Health, Arts and Design Hoban, Elizabeth; Deakin University, School of Health and Scial Development, Faculty of Health Osborne, Richard; Swinburne University of Technology, Centre for Global Health and Equity, Faculty of Health, Arts and Design
Primary Subject Heading:	Public health
Secondary Subject Heading:	Health policy, Health services research, Qualitative research, Research methods
Keywords:	PUBLIC HEALTH, QUALITATIVE RESEARCH, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title: Questionnaire validation practice within a theoretical framework: a systematic descriptive literature review of health literacy assessments

Authors: Melanie Hawkins[1], Gerald R Elsworth[1,2], Elizabeth Hoban[1], Richard H Osborne[2]

Institutions: Deakin University, Swinburne University of Technology

[1] School of Health and Social Development
Faculty of Health, Deakin University, Australia

[2] Centre for Global Health and Equity
Faculty of Health, Arts and Design, Swinburne University of Technology, Australia

Corresponding author:

Melanie Hawkins
School of Health and Social Development
Faculty of Health, Deakin University, Australia

Research Fellow
Postal address: Centre for Global Health and Equity
AMDC building, Level 9, Room 907
Faculty of Health, Arts and Design, Swinburne University of Technology
453/469-477 Burwood Road, Hawthorn, Australia
Phone: +61 439 354 456
Correspondence email: melaniehawkins@swin.edu.au

Co-authors:

Gerald R Elsworth
Honorary Professor (Health Program Evaluation)
School of Health and Social Development
Faculty of Health, Deakin University, Australia

Adjunct Professor
Centre for Global Health and Equity
Faculty of Health, Arts and Design, Swinburne University of Technology, Australia
Email: gelsworth@swin.edu.au

Elizabeth Hoban
Associate Professor
School of Health and Social Development
Faculty of Health, Deakin University, Australia
Email: elizabeth.hoban@deakin.edu.au

Richard H Osborne
Distinguished Professor of Health Sciences
Centre for Global Health and Equity
Faculty of Health, Arts and Design, Swinburne University of Technology, Australia
Email: rosborne@swin.edu.au

Author contributions: MH and RHO conceptualised the research question and analytical plan. MH led, with all authors contributing to, the development of the search strategy, selection criteria, data extraction criteria, and analysis method. MH conducted the literature search with guidance from EH. MH screened the literature, and extracted and analysed the data with the continuous support of and comprehensive checking by GRE. MH drafted the initial manuscript and led subsequent drafts. GRE, RHO and EH read and provided feedback on manuscript iterations, and approved the final manuscript. RHO is the guarantor.

Funding: MH was funded by a National Health and Medical Research Council (NHMRC) of Australia Postgraduate Scholarship (APP1150679). RHO was funded in part through a National Health and Medical Research Council (NHMRC) of Australia Principal Research Fellowship (APP1155125).

Conflicts of interest: None

Data availability statement: All data relevant to the study are included in the article or uploaded as supplementary information.

Word count: Abstract = 281; Main text = 4942

Acknowledgements: The authors acknowledge and thank Rachel West, Deakin University Liaison Librarian, for her expertise in systematic literature reviews and her patient guidance through the detailed process of searching the literature.

Abstract

Objective Validity refers to the extent to which evidence and theory support the adequacy and appropriateness of inferences based on score interpretations. The health sector is lacking a theoretically-driven framework for the development, testing and use of health assessments. This study used the *Standards for Educational and Psychological Testing* framework of five sources of validity evidence to assess the types of evidence reported for health literacy assessments, and to identify studies that referred to a theoretical validity testing framework.

Methods A systematic descriptive literature review investigated methods and results in health literacy assessment development, application and validity testing studies. Electronic searches were conducted in EBSCOhost, EMBASE, Open Access Theses and Dissertations, and ProQuest Dissertations. Data were coded to the *Standards*' five sources of validity evidence, and for reference to a validity testing framework.

Results Coding on 46 studies resulted in 195 instances of validity evidence across the five sources. Only nine studies directly or indirectly referenced a validity testing framework. Evidence based on *relations to other variables* is most frequently reported.

Conclusions The health and health equity of individuals and populations are increasingly dependent on decisions based on data collected through health assessments. An evidence-based theoretical framework provides structure and coherence to existing evidence and stipulates where further evidence is required to evaluate the extent to which data are valid for an intended purpose. This review demonstrates the use of the *Standards'* theoretical validity testing framework to evaluate

sources of evidence reported for health literacy assessments. Findings indicate that theoretical validity testing frameworks are rarely used to collate and evaluate evidence in validation practice for health literacy assessments. Use of the *Standards'* theoretical validity testing framework would improve evaluation of the evidence for inferences derived from health assessment data on which public health and health equity decisions are based.

Keywords Validity; Validation; Validity Testing Theory; Validity Testing Framework; Health Literacy; Health Assessment; Measurement.

Article summary

Strengths and limitations of this literature review

- This is the first time a theoretical validity testing framework, the five sources of evidence from the *Standards for Educational and Psychological Testing*, has been applied to the examination of validity evidence for health literacy assessments.
- A strength of this study is that validity is clearly defined, in accordance with the authoritative validity testing literature, as the extent to which theory and evidence (quantitative and qualitative) support score interpretation and use.
- A limitation was the restriction of the search to studies and health literacy assessments published or administered in English, which may introduce an English language and culture bias to the sample.
- A further limitation was the lack of clarity in some papers about the methods used and results obtained, leading to difficulties in coding validity evidence and may have led to some misclassification of reported evidence for some papers.

Questionnaire validation practice within a theoretical framework: a systematic descriptive literature review of health literacy assessments

Background

It has been argued that the health sector is lacking a theoretically-driven framework of validation practice for the development, testing and use of health assessments. [1-6] Such a framework could guide and strengthen validation planning for the interpretation and use of health assessment data. [2, 3, 7] Interpretations of scores from health literacy, assessments are increasingly being used to make decisions about the design, selection and evaluation of interventions and policies to improve health equity for individuals, communities and populations. [2-4, 8, 9] To ensure that decisions based on data from all health assessments are justified, and lead to equitable outcomes, validation practice must generate information about the degree to which the intended interpretations and use of data are supported by evidence and the theory of the construct being measured. [10-19] Validation research is complex [7, 20] and a theoretical framework would facilitate an evaluation of a range of evidence to determine valid interpretation and use of health assessment data. [2, 4, 18, 20, 21]

Health literacy

Health literacy is a relatively new field of research with a range of definitions for different settings [22-25] and advances in the approaches to its measurement. [26-32] Some health literacy assessments measure an observer's (e.g., clinician's or researcher's) observations of a person's health literacy, which often consists of testing a person's health-related numeracy, reading and comprehension. [33, 34] Objective measurement can support a clinician to provide health information in formats and at reading levels that are suited to individual patients but usually these measures do not assess other important dimensions of the health literacy construct. [35] Self-report measures of health literacy have become useful with the rise of the patient-centred healthcare movement, and these typically provide individuals' perspectives of a range of aspects of their health and health contexts. [23, 36] This type of measurement can capture the multidimensional aspects of the health literacy construct to look at broader implications of treatment, care and intervention outcomes. [37] Assessments could also combine both objective and self-report measurement of health literacy. Data from health literacy assessments have been used to inform health literacy interventions [8, 19, 38-41] and, increasingly, health policies. [42-46] However, despite the different definitions that health literacy assessments are based on (and thus, necessarily, the different score interpretations and uses), the data are often correlated and compared as if the interpretation of the scores have the same meaning. [27] A theoretical validity testing framework would help researchers, clinicians and policy makers to differentiate between the meanings of data from different health

literacy assessments, and evaluate existing evidence to support data interpretations, to enable them to choose the assessment that is most appropriate for their intended clinical or research purpose.

Contemporary validity testing theory

The validity testing framework of the 2014 *Standards for Educational and Psychological Testing* (the *Standards*) is the authoritative text for contemporary validity testing theory. [5] It results from about 100 years of the evolution of validity theory. [47, 48] The *Standards* defines validity as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (p.11) and validation as the process of '…accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations' (p.11). The framework describes five types of validity evidence that can be evaluated to justify test score interpretation and use: 1) *test content*; 2) *response processes* of respondents and users; 3) *internal structure* of the assessment test; 4) *relations to other variables*, and 5) *consequences* of testing, as related to validity (Table 1). [5, 6, 49, 50] Evidence from each of these sources may be needed to verify data interpretation and use.

Table 1. The five sources of validity evidence [5, 49]

1.	**Evidence based on test content** The relationship of the item themes, wording and format with the intended construct, including administration process.
2.	**Evidence based on response processes** The cognitive processes and interpretation of items by respondents and users, as measured against the intended construct.
3.	**Evidence based on internal structure** The extent to which item interrelationships conform to the intended construct.
4.	**Evidence based on external variables** The pattern of relationships of test scores to external variables as predicted by the intended construct.
5.	**Evidence based on validity and the consequences of testing** Intended and unintended consequences, as can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant variance.

The expectation of the *Standards* and leading validity theorists is that the validation process consists of an evaluative integration of different types of validity evidence (not types of validity) to support score meaning for a specific use. [2, 4, 5, 13-15, 51-57] Integral to this framework are quantitative methods to evaluate an assessment's statistical properties, but also important is validity evidence based on qualitative research methods. [4, 58-65] Qualitative methods are used to ensure technical evidence for *test content* and *response processes*, and to investigate validity-related *consequences* of testing. [7, 12, 52, 63-69] There are guides to assess quantitative measurement properties [70-72] but still needed are reviews that include qualitative validity evidence, and that place validity evidence for health assessments within a validity testing framework such as the *Standards*. [2, 4, 6, 49]

Rationale

As a guide to inform and improve the processes used to develop and test health assessments, this review will examine validation practice for health literacy assessments. Health literacy is a relatively new area of research that appears to have proceeded with the 'types of validity' paradigm of early validation practice in education, and so it is ideally poised to embrace advancements in validity testing practices. Thus, an assumption underlying this review is that the field of health is not applying contemporary validity testing theory to guide validation practice, and that the focus of validation studies remains on the general psychometric properties of a health assessment rather than on the interpretation and use of scores. This study will provide an example of the application of the *Standards*' theoretical validity testing framework through the review of sources of validity evidence (generated through quantitative and qualitative methods) reported for health literacy assessments.

The aim of this systematic descriptive literature review was to use the validity testing framework of the *Standards* to categorise and count the sources of validity evidence reported for health literacy assessments and to identify studies that used or made reference to a theoretical validity testing framework. Specifically, the review addressed the following questions:

1. What is being reported as validity evidence for health literacy assessment data?
2. Is the validity evidence currently provided for health literacy assessments placed within a validity testing framework, such as that offered by the *Standards*?

Methods

King and He situate systematic descriptive literature reviews toward the qualitative end of a continuum of review techniques. [73] Nevertheless, this type of review employs a frequency analysis to categorise qualitative and quantitative research data to reveal interpretable patterns. [32, 73-78] This review will appraise validation practice for health literacy assessments using the *Standards'* framework of five evidence sources. It will not critique nor assess the quality of individual health literacy assessments or studies.

Inclusion and exclusion criteria, information sources, and search strategy

The method for this review was previously reported in a protocol paper. [49] The eligibility and exclusion criteria, information sources, and search terms are summarised in Table 2. Peer reviewed full articles and examined theses were included in the search. Supplementary file 1 shows the MEDLINE database search strategy, and this was modified for the other databases. The review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement. [79] See Supplementary file 2 for the PRISMA checklist.

Table 2. Summary of inclusion and exclusion criteria, information sources, and search terms

Inclusion criteria	**Exclusion criteria**
Not limited by start date: end date March 2019	Systematic reviews and other types of reviews, grey literature (i.e., any studies or reports not published in a peer reviewed journal)
Development, application and validity testing studies and examined theses about health literacy assessments	Health literacy assessments designed for specific demographic groups (e.g., children) or health conditions (e.g., kidney disease)
All definitions of health literacy; and objective, subjective, uni- and multi-dimensional health literacy assessments	Predictive, association or other comparative studies that do not claim in the abstract to contribute validity evidence
Studies published and health literacy assessments developed and administered in the English language	Health literacy assessments developed or administered in languages other than English^
Qualitative and quantitative research methods	Translation studies
Information sources: EBSCOhost (MEDLINE Complete, Global Health, CINAHL Complete, PsycINFO, Academic Search Complete); EMBASE; Open Access Theses and Dissertations; ProQuest Dissertations; references of relevant systematic reviews; authors' reference lists	
Search terms: Medical subject headings (MeSH) and text words - valid*, verif*, "patient reported outcome*", questionnaire*, survey*, "self report*", "self rated", assess*, test*, tool*, "health literacy", measure*, psychometric*, interview*, "think aloud", "focus group*", "validation studies", "test validity"	

^ See *Results* for exceptions.

Article selection, and data extraction, analysis and synthesis

Duplicates were removed and a title and abstract screening of identified articles was performed in Endnote Reference Manager X9 by one author (MH). Identified full text articles (n=92) were screened for relevance by MH and corroborated with an independent screening of 10% (n=9) of the search results by a second author (GRE). Additionally, MH consulted with GRE when a query arose about inclusion of an article in the review.

Data extraction from articles for final inclusion was undertaken by one author (MH) with all data extraction comprehensively and independently checked by a second author (GRE). Both authors then corroborated to achieve categorisation consistency. General characteristics for each study were extracted but of primary interest were the sources of validity evidence reported, as were statements about or references to a theoretical validity testing framework. The validity evidence reported in each article was categorised according to the five sources of validity evidence in the *Standards*, whether or not the authors of the articles reported it that way. When the methods were unclear, the results were interpreted to determine the type of evidence generated by the study. A study was categorised as using or referencing a theoretical validity testing framework if the authors made a statement that referred to a framework and directly cited the framework document or if there was a clear citation path to the framework document.

Descriptive and frequency analyses of the extracted data were conducted to identify patterns in the sources of validity evidence being reported, and for the number of studies that made reference to a validity testing framework.

Patient and public involvement

Patients and the public were not involved in the development or design of this literature review.

Results

Overall, 46 articles were identified for the review. The PRISMA flow diagram in Figure 1 summarises the results of the search. [79] There were 3,379 records identified through database searches with 4 articles identified through other sources. There were 1,922 records remaining after 1457 duplicates were removed. After applying the exclusion and inclusion criteria to all abstracts, with full text screening of 92 articles and theses, 40 articles and 6 theses were included in the review (n=46). Reasons for exclusion were that the health literacy assessment was developed in or administered in a language other than English (n=19); the assessment was specific to a disease or condition (n=8) or to a demographic group (n=2); the article was not a validity study (n=8); the study was not using a health literacy assessment (n=3) or used an adapted assessment (n=4); the assessment was based on an item-bank, which required a different approach to validity testing (n=1), or was a composite assessment where health literacy data were collected and analysed with another type of data (n=1).

Figure 1. Flow diagram for Preferred Reporting Items for Systematic reviews and Meta-Analyses

Four papers were identified from the broader literature. Two papers were identified from the references of previous literature reviews [80, 81]. The other two papers were known to the authors and were in their personal reference lists. These two papers were by Davis and colleagues and describe the development of the Rapid Estimate of Adult Literacy in Medicine (REALM) [33] and the shortened version of the REALM. [82] Neither of these papers were detected by the systematic review because Davis *et al*. do not claim these to be measures of health literacy but of literacy in medicine. Rather they state that both versions of the REALM are designed to be used by physicians in public health and primary care settings to identify patients with low reading levels. [33, 82-84] Nevertheless, we included these papers because the REALM and the shortened REALM have been used by clinicians and researchers as measures of health literacy, and are used either as the primary assessment or a comparator assessment in many studies.

Three papers identified in the database search were included in this review even though data were collected using translations of assessments originally developed in English. These studies were

included because of the frequency of use of these assessments in the field of health literacy measurement, and because at least part of the data were based on English language research. The Test of Functional Health Literacy in Adults (TOFHLA) [85] and the Newest Vital Sign (NVS) [34] both collected data in English and Spanish. The analyses for the European Health Literacy Survey (HLS-EU) study [23] used data from the English (Ireland), as well as Dutch and Greek versions of the HLS-EU.

Of the 46 studies, 34 were conducted in the United States of America (USA), 8 in Australia, 2 in Singapore, and 1 each in Canada and the Netherlands. There were 4 studies published in the decade between 1990 and 1999, 8 studies between 2000 and 2009, and 34 between 2010 and 2019.

Reports of reliability evidence were provided in 33 studies (72%). This resulted in 44 instances of reliability evidence, of which 29 (66% of all instances) were calculated using Cronbach's alpha for internal consistency, 4 (9% of all instances) using test-retest, 4 (9%) using inter-rater reliability calculations, and 7 (16%) using other methods. See Table 3 for country and year of publication, and reliability evidence.

Table 3. Country and year of publication, and reliability evidence

Country of study	**N**	**%**
USA	34	74%
Australia	8	17%
Singapore	2	4%
Canada	1	2%
Netherlands	1	2%
Year of publication by decade		
1990-1999	4	9%
2000-2009	8	17%
2010-2019	34	74%
Reliability		
Cronbach's alpha	29	66%
Test-retest	4	9%
Inter-rater	4	9%
Other methods	7	16%
Total instances of reliability	44	100%

Validity evidence for health literacy assessment data

The data extraction framework (Supplementary File 3) was adapted from Hawkins et al (p.1702) [6] and Cox and Owen (p.254). [58] More detailed sub-coding of the five *Standards*' categories was done and will be drawn on selectively to describe aspects of the results (Supplementary File 4).

Data analysis consisted of coding instances of validity evidence into the five sources of validity evidence of the *Standards*. The results of the review are presented as: 1) the total number of

instances of validity evidence for each evidence source reported across all studies; 2) the number of instances reported for objective, subjective and mixed methods health literacy assessments; and 3) the number of instances of evidence within each of the *Standards*' five sources, and a breakdown of the methods used to generate evidence.

Table 4 displays the overall results of the review. For the 46 studies that reported validity evidence for health literacy assessments, we identified 195 instances of validity evidence across the five sources: *test content* (n=52), *response processes* (n=7), *internal structure* (n=28), *relations to other variables* (n=107), and *consequences of testing* (n=1). Across types of health literacy assessments, there were 102 instances of validity evidence reported for health literacy assessments with an objective measurement approach (n=23 studies); 78 instances reported for assessments with a subjective measurement approach (n=20 studies); and 15 instances for assessments with a mixed methods approach or when multiple types of health literacy assessments were under investigation (n=3 studies).

Table 4. Sources of evidence for all studies, total instances of validity evidence, and for objective, subjective, and multiple/mixed methods health literacy assessments

	Studies (n=46*)	**Instances**** (n=195)	**Objective^** (n=23 studies; n=102 instances)	**Subjective^^** (n=20 studies; n=78 instances)	**Multiple and mixed methods** (n=3 studies; n=15 instances)
	N (%)	**N (%)**	**N (%)**	**N (%)**	**N (%)**
1. Test content	22 (48)	52 (27)	27 (26)	22 (28)	3 (20)
2. Response processes	6 (13)	7 (4)	2 (2)	5 (6)	0 (0)
3. Internal structure	15 (33)	28 (14)	11 (11)	15 (19)	2 (13)
4. Relations to other variables	42 (91)	107 (55)	61 (60)	36 (46)	10 (67)
5. Validity and the consequences of testing	1 (2)	1 (1)	1 (1)	0 (0)	0 (0)

*Most studies reported more than one source of validity evidence.
**Each time validity evidence was reported within a study.
^ Measures an observer's (e.g., clinician's) objective observations of a person's health literacy.
^^ Self-report (subjective) measure of health literacy.

1. Evidence based on test content

Nearly half of all studies (n=22) reported evidence based on test content, which resulted in 52 instances of validity evidence (Table 4 and Supplementary Table 1). Expert review was the most frequently reported method used to generate evidence (n=14 instances; 27% of all evidence based

on test content), [23, 33, 34, 36, 82, 83, 86-93] followed by the use of existing measures of the construct (n=8; 15%). [34, 36, 83, 90-92, 94, 95] Analysis of item difficulty was used 5 times (10%), [36, 86, 89, 92, 96] with literature reviews, [23, 90, 93, 97] participant feedback processes about items, [23, 34, 83, 89] and construct descriptions [23, 36, 91, 97] each used 4 times (8% each). Participant concept mapping [23, 36, 88] and examination of administration methods [36, 98, 99] were each used 3 times (6% each), and participant interviews [88, 100] were used twice (4%). Five other methods were each used once in 5 different studies: item intent descriptions; [36] items tested against item intent descriptions; [101] IRT analysis for item selection within domains; [90] item selection based on hospital medical texts; [85] and item selection based on a health literacy conceptual model. [100]

2. Evidence based on response processes

Only 7 instances based on *response processes* were reported across 6 of the 46 studies (Table 4 and Supplementary Table 2). The methods used were cognitive interviews with respondents (n=3 instances; 43% of all evidence based on *response processes*) [36, 88, 101] and with users (clinicians) (n=1; 14%), [101] as well as recording and timing the response times of respondents (n=3; 43%). [89, 98, 100]

3. Evidence based on internal structure

There were 15 studies (33% of all studies) that reported evidence based on the *internal structure* of health literacy assessments resulting in 28 instances (Table 4 and Supplementary Table 3). The most frequently reported methods were exploratory factor analysis (EFA) (including principal component analysis (PCA)) (n=7 instances; 25% of all evidence based on *response processes*) [88, 93, 100, 102-105] and confirmatory factor analysis (CFA) (also n=7; 25%). [91, 106, 107] Differential item functioning (DIF) was reported 3 times (11%), [88, 91, 102] and item-remainder correlations twice (7%). [36, 92] There were 9 other methods used to generate evidence for *internal structure*, including a variety of specific item-response theory (IRT) analyses for fit, item selection, and internal consistency. Each method was reported once, with some authors reporting more than one method. [36, 86, 89, 90, 103, 106]

4. Evidence based on relations to other variables

This was the most commonly reported type of validity evidence across studies (n=42 studies; 91%) (Table 4 and Supplementary Table 4). There were 18 studies that only reported evidence based on *relations to other variables*. [80, 81, 104, 108-122] Evidence within this category was coded, as per the *Standards*, into convergent evidence (i.e., relationships between items and scales of the same or similar structure), discriminant evidence (i.e., assessments measuring different constructs determined to be sufficiently uncorrelated), criterion-referenced evidence (i.e., how accurately

scores predict criterion performance), and evidence for group differences (i.e., relationships of scores with background characteristics such as demographic information). The *Standards* also includes evidence for generalisation but states that this relies primarily on studies that conduct research syntheses, and this review excluded studies that conducted meta-analyses. Across all studies, there were 107 instances of validity evidence reported for *relations to other variables*: 57 instances of convergent evidence (53% of all evidence in this category); 3 instances of discriminant evidence (3%); 17 instances of criterion-referenced evidence (16%); and 30 instances of evidence for group differences (28%).

The most frequently-used methods for convergent evidence were Spearman's [80, 85, 94, 96, 99, 105, 108, 110, 116, 118, 122] and Pearson's [33, 34, 82, 83, 90, 93, 104, 112, 113, 120, 123] correlation coefficients (11 instances and 19% each). These were closely followed by the receiver operating characteristic (ROC) curve and the area under the ROC (AUROC) curve (also n=11 instances; 19%). [81, 97, 99, 103, 110, 111, 117, 120, 123] A further 8 instances (14%) of correlation calculations with similar measures were reported but the types of calculation they performed were unclear. [86, 87, 92, 95, 103, 115, 119, 121]

Harper, Elsworth *et al*., and Osborne *et al*. [36, 90, 106] were the only 3 studies to generate discriminant evidence, as defined by the *Standards*. Harper [90] used the Pearson correlation coefficient to assess the association of components of a new health literacy instrument with the shortened version of the Test of Functional Health Literacy in Adults (S-TOFHLA). Elsworth *et al*. [106] compared the average variance extracted (AVE) and the variance shared between the nine scales of the Health Literacy Questionnaire (HLQ) (discriminant validity evidence between HLQ scales). Similarly, Osborne *et al*. [36] conducted a multi-scale factor analysis to investigate if the nine HLQ scales were conceptually distinct.

Linear regression models were the most common method to generate criterion-referenced evidence (n=6 instances; 35% of all criterion-referenced evidence). [86, 90, 107, 114, 115, 121] The Chi-square test of independence was used by 3 studies (18%), [87, 115, 121] with Spearman's correlation coefficient [110, 115] and logistic regression models [86, 115] each used by 2 studies (12% each).

There were 16 methods used to generate evidence for group differences and these were spread across 19 studies. The most frequently used methods were analysis of variance (ANOVA) (n=5 instances; 17%) [88, 92, 93, 103, 121] and linear regression models (n=4; 13%). [80, 83, 91, 123]

5. Evidence based on validity and consequences of testing

One study did investigations that led to conclusions about validity and the *consequences* of testing (p.221). [83] Elder *et al*. found that the REALM underrepresented the construct of health literacy when defined as the ability to obtain, interpret, and understand basic health information.

Use of a validity testing framework when reporting validity evidence for health literacy assessments

Few studies referred to a validity testing framework or used a framework to structure or guide their work. Of the 46 studies, 9 directly or indirectly referenced a validity testing framework, and made a statement to support the citation (see Supplementary File 3). The frameworks directly cited by 3 studies [87, 101, 106] were the 2014 *Standards*; [5] Michael T Kane's argument-based approach to validation; [14] Samuel J Messick's unified theory of validation; [17, 124] and Francis *et al*'s checklist operationalising measurement characteristics of patient-reported outcome measures. [125] There were 6 studies [36, 83, 93, 96, 102, 107] that indirectly cited Messick, Kane, and/or the 1985, 1999 or 2014 versions of the *Standards* [5, 126, 127] through other citations. A 10th study [88] referenced Buchbinder *et al*. [128], which cites the *Standards*, but there was no clear statement about validity testing to support the citation.

Discussion

This systematic descriptive literature review found that studies in health literacy measurement rarely use or reference a structured theoretical framework for validation planning or testing. Further, this review's use of the *Standards*' framework revealed that validity testing studies for health literacy assessments most frequently, and often only, report evidence based on *relations to other variables*. It is usual and reasonable for a single validity study to not provide comprehensive evidence about a PROM, and this is why an organising framework for evaluating evidence from a range of studies is so important. The findings from this review show that validation practice for health literacy assessments does not use established validity testing criteria and is yet to embrace the structural framework of contemporary validity testing theory. [5, 6]

In this review, evidence based on *relations to other variables* was the most frequent type of validity evidence reported across the 46 studies. It was reported more than twice as frequently as evidence based on *test content*, which was the second most commonly reported source of validity evidence. Evidence based on *internal structure* was reported in almost half the studies. This is not an unexpected result given the propensity for validity testing studies to almost routinely conduct correlation of an assessment with another variable (e.g., a similar or different assessment). [129] In the early 20th Century, the focus of test validation was primarily on predictive validity practices (e.g., prediction of student academic achievement) and so correlation with known criteria was a common validation practice. [48, 130, 131] Development of the theory and practice of validation, and the need to use tests in various contexts with different population groups, has required consideration of the meaning of test scores, and that score interpretations usually lead to decisions or actions that can affect people's lives. [2, 3, 52, 66] As Kane explains, 'ultimately, the need for validation derives

from the scientific and social requirement that public claims and decisions be justified' (p.17). [13] A structured theoretical framework, such as the *Standards*, facilitates validation planning, testing, and integration of evidence for decision making. It can also support new users of a health assessment to judge existing evidence and previous rationales for data interpretation and use, and how these might justify the use of the assessment in a new context.

Reports of evidence based on *response processes* and on *consequences* of testing were negligible in this review. This is the first time this has been observed in the field of health literacy although it has been observed previously in other fields of research. [50, 68, 132] Evidence based on the cognitive (response) processes of respondents (and of assessment users [59, 101]) can be essential to understanding the meanings derived from assessment scores for each new testing purpose. [69] Consequential evidence, although a controversial area of research, [50, 66] can reveal important outcomes for equitable decision making, such as those discussed by Elder *et al*. [83] regarding the use of the REALM, a word recognition assessment, with non-native speakers of English in a world in which health literacy is understood to be about equitable access to, and understanding and use of health information and services. [42, 133-135] Potential risks for unintended consequences of testing can be lessened through the development of the content of health assessments using comprehensive grounded practices that ensure wide and deep coverage of the lived experiences of intended respondents. [36, 136-138]

The findings of this review are important because institutions and governments around the world are increasingly implementing health literacy as a basis for health policy and practice development and evaluation. [43-46, 139] There needs to be certainty that inferences made from health literacy measurement data are leading to accurate and equitable decision making about health care, interventions, and policies, and that these decisions are as fair for the people with the lowest health literacy as for those with the highest. [11, 19, 46, 52, 140-143] Some types of health interventions are known to widen health inequalities. [143-147] Messick emphasises construct underrepresentation and construct-irrelevant variance as causes for negative testing consequences, as related to validity. [124, 148] For example, if a health assessment is biased by a specific perspective about causes of health disparities then construct underrepresentation can be a threat to the validity of inferences and actions taken from the scores. Likewise, if an assessment reflects a particular social perspective (e.g., middle class values and language embedded in the items) then there is the threat that the responses to the assessment are perfused with irrelevant variance derived from that perspective. Evidence from a range of sources is required to justify the use of measurement data in specific contexts (e.g., socioeconomic, demographic, cultural, language), and to assure decision makers of the absence of validity threats. [4, 51, 54]

This is the first time that a comprehensive review of sources of validity evidence for health literacy assessments has been undertaken within the theoretical validity testing framework of the *Standards*. For some methods, coding into the five sources of validity evidence was not straightforward and, in these cases, the *Standards* were consulted closely for guidance. Coding of studies by Elsworth *et al*. and Osborne *et al*. [36, 106] to *relations to other variables* (discriminant evidence) required some deliberation because the evidence in both studies was for discrimination analyses between independent scales *within* a multi-scale health literacy assessment, rather than between different health literacy assessments. The developers of the HLQ view the nine scales as measuring distinct, albeit related, constructs. [36] The *Standards* (p.16) explain that 'external variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs'. [5] It was on the basis of the last part of this statement about tests measuring related or different constructs that these two studies were coded in *relations to other variables* as discriminant evidence.

In a few studies, some assessments seemed to be regarded as proxies for health literacy, which suggested that the researchers were thinking of them as measuring similar constructs to health literacy. In these cases, evidence was coded in *relations to other variables* as convergent evidence (i.e., convergence between measures of the same or similar construct) rather than as criterion-referenced evidence (i.e., prediction of other criteria). For example, Curtis *et al*. [86] explored correlations between the Comprehensive Health Activities Scale (CHAS) with the Mini Mental Status Exam (MMSE) as well as with the TOFHLA, the REALM, and the NVS. [86] Driessnack *et al*. [108] looked at correlations between parents' and children's NVS scores with their self-reports of the number of children's books in the home. Dykhuis *et al*. [87] correlated the Brief Medical Numbers Test (BMNT) with the Montreal Cognitive Assessment (MoCA) as well as with two versions of the REALM.

Further to coding for *relations to other variables* are the distinctions between convergent evidence, criterion-referenced evidence, and evidence for group differences. Coding to convergent evidence was based on analyses of assessments of the same or similar construct (e.g., typically, comparisons of one health literacy assessment with another health literacy assessment). Coding to criterion-referenced evidence was based on analyses of prediction (e.g., a health literacy assessment with a disease knowledge survey). Coding for evidence of group differences was based on analyses of relationships with background characteristics such as demographic information.

Reliability was not coded within the five sources of evidence even though it does contribute to understanding the validity of score interpretations and use, especially for purposes of generalisation. [5] The *Standards* (p.33) classifies reliability into reliability/precision (i.e., consistency of scores

across different instances of testing) and reliability/generalisability coefficients (i.e., in the way that classical test theory refers to reliability as being correlation between scores on two equivalent forms of a test, with the assumption that there is no effect of the first test instance on the second test instance). The predominant focus in the reviewed papers was on the latter conception of reliability, most often calculated using Cronbach's alpha.

Strengths and limitations

An element of bias is potentially present in this review because of the restriction of the search to studies published and health literacy assessments developed and administered in the English language. Future studies may be improved if other languages were included. The health literacy assessments reviewed are those that are predominant in the field and may well provide a foundation for validity studies of more specifically targeted assessments.

Just as there were two papers known to the authors of an instrument that is frequently used to measure health literacy, and two further papers were identified from published literature reviews, it may be that more papers that would be relevant to this review were not identified. However, since the 1991 publication of the REALM, which was not designed as a health literacy assessment but has since been used as such, we predict that most assessments for the measurement of health literacy will be identified for this purpose, and would thus have been captured by the present search strategy. Validation practice is complex and there are many groups publishing validity testing studies that may have limited training and experience in the area. [1-4] There was a lack of clarity in some papers and theses about the methods used and results obtained, which caused difficulties with classifying the evidence within the *Standards* framework, so some misclassification is possible for some papers. Future work in this area would be improved if researchers used clearly defined and structured validity testing frameworks (i.e., the five validity evidence sources of the *Standards*) in which to classify evidence.

The main strength of this study was that validity is clearly defined as the extent to which theory and evidence (quantitative and qualitative) support score interpretation and use. This definition is in accordance with leading authorities in the validity testing literature. [2, 5, 13, 51] A second strength of this study was the use of an established and well-researched theoretical validity testing framework, the *Standards*, to examine sources of evidence for health literacy assessments. Different health literacy assessments have different measurement purposes. Validation planning with a structured framework would help to determine the sources of evidence needed to justify the inferences from data, and to guide potential users. Application of theory to validation practice will provide a scientific basis for the development and testing of health assessments, enable systematic evaluations of validity evidence, and help detect possible threats to the validity of the interpretation and use of data in different contexts. [2, 3, 15],

Conclusions

Arguments for the validity of decisions based on health assessment data must be based on evidence that the data are valid for the decision purpose to ensure the integrity of the consequences of the measurement, yet this is frequently overlooked. This literature review demonstrated the use of the *Standards'* validity testing framework to collate and assess existing evidence and identify gaps in the evidence for health literacy assessments. Potentially, the framework could be used to assess the validity of data interpretation and use of other health assessments in different contexts. Developers of health assessments can use the *Standards*' framework to clearly outline their measurement purpose, and to define the relevant and appropriate validity evidence needed to ensure evidence-based, valid and equitable decision making for health. This view of validity being about score interpretation and use challenges the long-held view that validity is about the properties of the assessment instrument itself. It is also the basis for establishing a sound argument for the authority of decisions based on health assessment data, which is critical to health services research and to the health and health equity of the populations affected by those decisions.

References

1. McClimans, L., *A theoretical framework for patient-reported outcome measures.* Theoretical medicine and bioethics, 2010. **31**(3): p. 225-240.
2. Zumbo, B.D. and E.K. Chan, eds. *Validity and validation in social, behavioral, and health sciences*. Social Indicators Research Series. 2014, Springer International Publishing: Switzerland.
3. Sawatzky, R., et al., *Montreal Accord on patient-reported outcomes (PROs) use series–Paper 7: modern perspectives of measurement validation emphasize justification of inferences based on patient reported outcome scores.* Journal of Clinical Epidemiology, 2017. **89**: p. 154-159.
4. Kwon, J.Y., S. Thorne, and R. Sawatzky, *Interpretation and use of patient-reported outcome measures through a philosophical lens.* Quality of Life Research, 2019: p. 1-8.
5. AERA, APA, and NCME, *Standards for educational and psychological testing*. 2014, Washington, DC: American Educational Research Association.
6. Hawkins, M., G.R. Elsworth, and R.H. Osborne, *Application of validity theory and methodology to patient-reported outcome measures (PROMs): building an argument for validity.* Quality of Life Research, 2018. **27**(7): p. 1695-1710.
7. O'Leary, T.M., J.A. Hattie, and P. Griffin, *Actual interpretations and use of scores as aspects of validity.* Educational Measurement: Issues and Practice, 2017. **36**(2): p. 16-23.
8. Bakker, M.M., et al., *Acting together–WHO National Health Literacy Demonstration Projects (NHLDPs) address health literacy needs in the European Region.* Public Health Panorama, 2019. **5**(2-3): p. 233-243.
9. Klinker, C.D., et al., *Health Literacy is Associated with Health Behaviors in Students from Vocational Education and Training Schools: A Danish Population-Based Survey.* Journal of Environmental Research and Public Health, 2020. **17**(2): p. 671.

10. Chapelle, C.A., *The TOEFL validity argument.* Building a validity argument for the Test of English as a Foreign Language, 2008: p. 319-352.
11. Elsworth, G.R., S. Nolte, and R.H. Osborne, *Factor structure and measurement invariance of the Health Education Impact Questionnaire: Does the subjectivity of the response perspective threaten the contextual validity of inferences?* SAGE Open Medicine, 2015. **3**: p. 2050312115585041.
12. Shepard, L.A., *The centrality of test use and consequences for test validity.* Educational Measurement: Issues and Practice, 1997. **16**(2): p. 5-24.
13. Kane, M.T., *Validation*, in *Educational Measurement*, R.L. Brennan, Editor. 2006, Rowman & Littlefield Publishers / Amer Council Ac1 (Pre Acq). p. 17-64.
14. Kane, M.T., *An argument-based approach to validity.* Psychological Bulletin, 1992. **112**(3): p. 527-535.
15. Kane, M.T., *Explicating validity.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 198-211.
16. Messick, S., *Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning.* American Psychologist, 1995. **50**(9): p. 741.
17. Messick, S., *Validity of test interpretation and use.* ETS Research Report Series, 1990. **1990**(1): p. 1487-1495.
18. Moss, P.A., B.J. Girard, and L.C. Haniford, *Validity in educational assessment.* Review of research in education, 2006: p. 109-162.
19. Batterham, R., et al., *Health literacy: applying current concepts to improve health services and reduce health inequalities.* Public health, 2016. **132**: p. 3-12.
20. Shepard, L.A., *Evaluating test validity: Reprise and progress.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 268-280.
21. Hubley, A.M. and B.D. Zumbo, *A dialectic on validity: Where we have been and where we are going.* The Journal of General Psychology, 1996. **123**(3): p. 207-215.
22. Nutbeam, D., *The evolving concept of health literacy.* Soc Sci Med, 2008. **67**(12): p. 2072-8.
23. Sørensen, K., et al., *Health literacy and public health: a systematic review and integration of definitions and models.* BMC Public Health, 2012. **12**: p. 80.
24. Sykes, S., et al., *Understanding critical health literacy: a concept analysis.* BMC Public Health, 2013. **13**(1): p. 150.
25. Pleasant, A., J. McKinney, and R. Rikard, *Health literacy measurement: a proposed research agenda.* Journal of Health Communication, 2011. **16**(sup3): p. 11-21.
26. Jordan, J.E., R.H. Osborne, and R. Buchbinder, *Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses.* J Clin Epidemiol, 2010.
27. Altin, S.V., et al., *The evolution of health literacy assessment tools: a systematic review.* BMC public health, 2014. **14**(1): p. 1207.
28. McCormack, L., et al., *Recommendations for advancing health literacy measurement.* Journal of Health Communication, 2013. **18**(sup1): p. 9-14.
29. Mancuso, J.M., *Assessment and measurement of health literacy: an integrative review of the literature.* Nursing & health sciences, 2009. **11**(1): p. 77-89.
30. Haun, J.N., et al., *Health literacy measurement: an inventory and descriptive summary of 51 instruments.* J Health Commun, 2014. **19 Suppl 2**: p. 302-33.
31. Guzys, D., et al., *A critical review of population health literacy assessment.* BMC Public Health, 2015. **15**(1): p. 215.
32. Barry, A.E., et al., *Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals.* Health Education & Behavior, 2014. **41**(1): p. 12-18.
33. Davis, T.C., et al., *Rapid assessment of literacy levels of adult primary care patients.* Family medicine, 1991. **23**(6): p. 433-435.
34. Weiss, B.D., et al., *Quick assessment of literacy in primary care: the newest vital sign.* The Annals of Family Medicine, 2005. **3**(6): p. 514-522.

35. Jessup, R.L., et al., *Using co-design to develop interventions to address health literacy needs in a hospitalised population.* BMC Health Services Research, 2018. **18**(1): p. 989.
36. Osborne, R.H., et al., *The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ).* BMC Public Health, 2013. **13**: p. 658.
37. Jessup, R.L. and R. Buchbinder, *What if I cannot choose wisely? Addressing suboptimal health literacy in our patients to reduce over-diagnosis and overtreatment.* Internal Medicine Journal, 2018. **48**(9): p. 1154-1157.
38. Roberts, J., *Local action on health inequalities: Improving health literacy to reduce health inequalities*. 2015, UCL Institute of Health Equity: London.
39. Batterham, R.W., et al., *The OPtimising HEalth LIterAcy (Ophelia) process: study protocol for using health literacy profiling and community engagement to create and implement health reform.* BMC Public Health, 2014. **14**(1): p. 694-703.
40. Beauchamp, A., et al., *Systematic development and implementation of interventions to Optimise Health Literacy and Access (Ophelia).* BMC Public Health, 2017. **17**(1): p. 230.
41. Barry, M.M., M. D'Eath, and J. Sixsmith, *Interventions for Improving Population Health Literacy: Insights From a Rapid Review of the Evidence.* Journal of Health Communication, 2013. **18**(12): p. 1507-1522.
42. Australian Bureau of Statistics. *National Health Survey: Health Literacy, 2018*. 2019 14 October 2019]; Available from: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4364.0.55.014Main+Features12018?OpenDocument.
43. Trezona, A., G. Rowlands, and D. Nutbeam, *Progress in implementing national policies and strategies for health literacy—What have we learned so far?* International Journal of Environmental Research and Public Health, 2018. **15**(7): p. 1554.
44. WHO Regional Office for Europe, *WHO Health Evidence Network synthesis report 65. What is the evidence on the methods, frameworks and indicators used to evaluate health literacy policies, programmes and interventions at the regional, national and organizational levels?* 2019: Copenhagen.
45. Putoni, S., *Health Literacy in Wales - A Scoping Document for Wales*. 2010, Welsh Assembly Government: Wales.
46. Scottish Government NHS Scotland, *Making it Easier: A Health Literacy Action Plan for Scotland 2017-2025*. 2017: Edinburgh.
47. Kelley, T.L., *Interpretation of educational measurements*. Measurement and adjustment series. 1927, Yonkers-on-Hudson, N.Y.: World Book. xiii, 363 p.
48. Sireci, S.G., *On the validity of useless tests.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 226-235.
49. Hawkins, M., G.R. Elsworth, and R.H. Osborne, *Questionnaire validation practice: a protocol for a systematic descriptive literature review of health literacy assessments.* BMJ Open, 2019. **9:e030753**(10).
50. Cizek, G.J., S.L. Rosenberg, and H.H. Koons, *Sources of validity evidence for educational and psychological tests.* Educational and psychological measurement, 2008. **68**(3): p. 397-412.
51. Messick, S., *Validity*, in *Educational Measurement*, R. Linn, Editor. 1989, American Council on Education/Macmillan Publishing Company: New York.
52. Messick, S., *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment.* ETS Research Report Series, 1998: p. 3-20.
53. Cronbach, L.J., *Five perspectives on validity argument*, in *Test validity*, H. Wainer and H.I. Braun, Editors. 1988, Lawrence Erlbaum Associates Inc: New Jersey. p. 3-17.
54. House, E., *Evaluating with validity*. 1980, Beverly Hills, California: Sage Publications.
55. Shepard, L.A., *Evaluating test validity*, in *Review of Research in Education*, L. Darling-Hammond, Editor. 1993, American Educational Research Asociation. p. 405-450.
56. Kane, M.T., *Validating the interpretations and uses of test scores.* Journal of Educational Measurement, 2013. **50**(1): p. 1-73.
57. Kane, M.T., *Validity as the evaluation of the claims based on test scores.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 309-311.

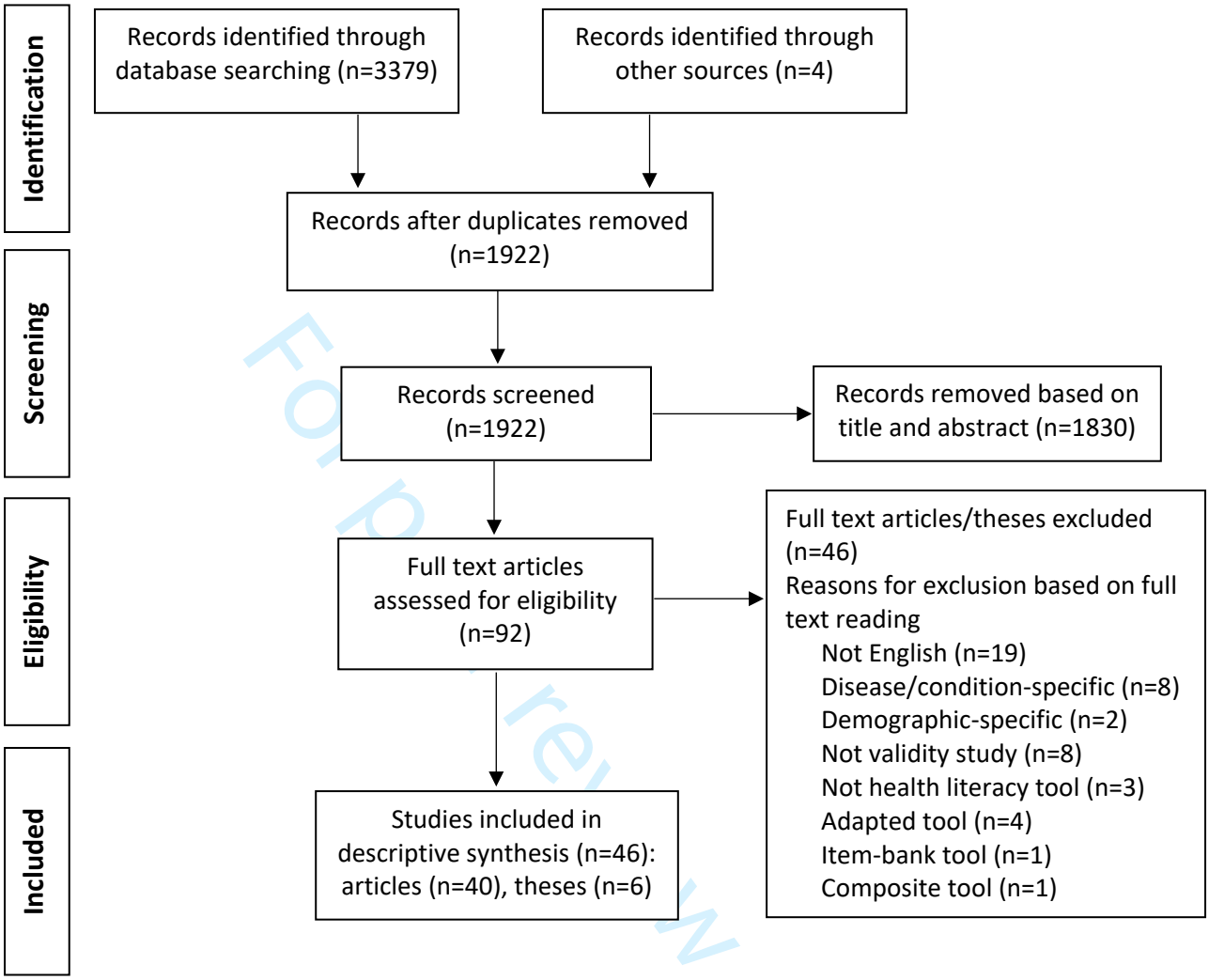
58. Cox, D.W. and J.J. Owen, *Validity evidence for a perceived social support measure in a population health context*, in *Validity and validation in social, behavioral, and health sciences*, B.D. Zumbo and E.K. Chan, Editors. 2014, Springer International Publishing: Switzerland.
59. Zumbo, B.D. and A.M. Hubley, eds. *Understanding and investigating response processes in validation research*. Social Indicators Research Series, ed. A.C. Michalos. Vol. 69. 2017, Springer International Publishing: Switzerland.
60. Hubley, A.M. and B.D. Zumbo, *Response processes in the context of validity: Setting the stage*, in *Understanding and investigating response processes in validation research*, B.D. Zumbo and A.M. Hubley, Editors. 2017, Springer International Publishing: Switzerland. p. 1-12.
61. Onwuegbuzie, A.J. and N.L. Leech, *Validity and qualitative research: An oxymoron?* Quality and Quantity, 2007. **41**(2): p. 233-249.
62. Onwuegbuzie, A.J. and N.L. Leech, *On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies.* International journal of social research methodology, 2005. **8**(5): p. 375-387.
63. Castillo-Díaz, M. and J.-L. Padilla, *How cognitive interviewing can provide validity evidence of the response processes to scale items.* Social indicators research, 2013. **114**(3): p. 963-975.
64. Padilla, J.-L. and I. Benítez, *Validity evidence based on response processes.* Psicothema, 2014. **26**(1): p. 136-144.
65. Padilla, J.-L., I. Benítez, and M. Castillo, *Obtaining validity evidence by cognitive interviewing to interpret psychometric results.* Methodology, 2013. **9**(3): p. 113-122.
66. Moss, P.A., *The role of consequences in validity theory.* Educational Measurement: Issues and Practice, 1998. **17**(2): p. 6-12.
67. Hubley, A.M. and B.D. Zumbo, *Validity and the consequences of test interpretation and use.* Social Indicators Research, 2011. **103**(2): p. 219.
68. Zumbo, B.D. and A.M. Hubley, *Bringing consequences and side effects of testing and assessment to the foreground.* Assessment in Education: Principles, Policy & Practice, 2016. **23**(2): p. 299-303.
69. Kane, M. and R. Mislevy, *Validating score interpretations based on response processes*, in *Validation of score meaning for the next generation of assessments*, K. Ercikan and J.W. Pellegrino, Editors. 2017, Routledge: New York. p. 11-24.
70. Terwee, C.B., et al., *Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist.* Quality of Life Research, 2012. **21**(4): p. 651-657.
71. Devellis, R.F., *A consumer's guide to finding, evaluating, and reporting on measurement instruments.* Arthritis and Rheumatism, 1996. **9**(3): p. 239-245.
72. Lohr, K.N., *Assessing health status and quality-of-life instruments: attributes and review criteria.* Quality of Life Research, 2002. **11**(3): p. 193-205.
73. King, W.R. and J. He, *Understanding the role and methods of meta-analysis in IS research.* Communications of the Association for Information Systems, 2005. **16**(1): p. 32.
74. Yang, H. and M. Tate, *A descriptive literature review and classification of cloud computing research.* Communications of the Association for Information Systems, 2012. **31**: p. 2.
75. Schlagenhaufer, C. and M. Amberg. *A Descriptive Literature Review and Classification Framework for Gamification in Information Systems*. in *European Conference on Information Systems*. 2015. Germany: Gartner.
76. Roter, D.L., J.A. Hall, and N.R. Katz, *Patient-physician communication: a descriptive summary of the literature.* Patient Education and Counseling, 1988. **12**(2): p. 99-119.
77. Guzzo, R.A., S.E. Jackson, and R.A. Katzell, *Meta-analysis analysis*, in *Research in Organizational Behavior*, L.L. Cummings and B.M. Staw, Editors. 1987, JAI Press: Greenwich, CT. p. 407-442.
78. Paré, G., et al., *Synthesizing information systems knowledge: A typology of literature reviews.* Information and Management, 2015. **52**(2): p. 183-199.

79. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS medicine, 2009. **6**(7): p. e1000097.
80. Barber, M.N., et al., *Up to a quarter of the Australian population may have suboptimal health literacy depending upon the measurement tool: results from a population-based survey.* Health Promot Int, 2009. **24**(3): p. 252-61.
81. Wallace, L.S., et al., *Brief report: screening items to identify patients with limited health literacy skills.* Journal of General Internal Medicine, 2006. **21**(8): p. 874-877.
82. Davis, T.C., et al., *Rapid estimate of adult literacy in medicine: a shortened screening instrument.* 1993. **25**(6): p. 391-395.
83. Elder, C., et al., *Assessing health literacy: A new domain for collaboration between language testers and health professionals.* Language Assessment Quarterly, 2012. **9**(3): p. 205-224.
84. Dumenci, L., et al., *On the Validity of the Shortened Rapid Estimate of Adult Literacy in Medicine (REALM) Scale as a Measure of Health Literacy.* Communication Methods and Measures, 2013. **7**(2): p. 134-143.
85. Parker, R.M., et al., *The test of functional health literacy in adults.* Journal of General Internal Medicine, 1995. **10**(10): p. 537-541.
86. Curtis, L.M., et al., *Development and validation of the comprehensive health activities scale: a new approach to health literacy measurement.* Journal of Health Communication, 2015. **20**(2): p. 157-164.
87. Dykhuis, K.E., et al., *A New Measure of Health Numeracy: Brief Medical Numbers Test (BMNT).* Psychosomatics, 2019. **60**(3): p. 271-277.
88. Jordan, J.E., et al., *The Health Literacy Management Scale (HeLMS): A measure of an individual's capacity to seek, understand and use health information within the healthcare setting.* Patient education and counseling, 2013. **91**(2): p. 228-235.
89. Zhang, X.-H., et al., *Development and validation of a functional health literacy test.* The Patient: Patient-Centered Outcomes Research, 2009. **2**(3): p. 169-178.
90. Harper, R., *Comprehensive health literacy assessment for college students*, in *Department of Journalism and Technical Communication*. 2013, Colorado State University: Fort Collins, Colorado.
91. Bann, C.M., et al., *The health literacy skills instrument: a 10-item short form.* Journal of Health Communication, 2012. **17**(sup3): p. 191-202.
92. McCormack, L., et al., *Measuring health literacy: a pilot study of a new skills-based instrument.* Journal of Health Communication, 2010. **15**(S2): p. 51-71.
93. DeBello, M.C., *The development and psychometric testing of the health literacy knowledge, application, and confidence scale (HLKACS)*, in *College of Education and College of Nursing*. 2016, Eastern Michigan University: Michigan.
94. Baker, D.W., et al., *Development of a brief test to measure functional health literacy.* Patient Education and Counseling, 1999. **38**(1): p. 33-42.
95. Shaw, T.C., *Uncovering health literacy: Developing a remotely administered questionnaire for determining health literacy levels in health disparate populations.* Journal of Hospital Administration, 2014. **3**(4): p. 140.
96. Begoray, D.L. and B. Kwan, *A Canadian exploratory study to define a measure of health literacy.* Health Promotion International, 2011. **27**(1): p. 23-32.
97. Chew, L.D., K.A. Bradley, and E.J. Boyko, *Brief questions to identify patients with inadequate health literacy.* Family Medicine, 2004. **11**: p. 12.
98. Chesser, A.K., et al., *Health literacy assessment of the STOFHLA: paper versus electronic administration continuation study.* Health Education & Behavior, 2014. **41**(1): p. 19-24.
99. Dageforde, L.A., et al., *Validation of the written administration of the short literacy survey.* Journal of Health Communication, 2015. **20**(7): p. 835-842.
100. Sørensen, K., et al., *Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q).* BMC Public Health, 2013. **13**(1): p. 1-22.

101. Hawkins, M., et al., *The Health Literacy Questionnaire (HLQ) at the patient-clinician interface: a qualitative study of what patients and clinicians mean by their HLQ scores.* BMC Health Services Research, 2017. **17**(1): p. 309.
102. Morris, R.L., et al., *Measurement properties of the Health Literacy Questionnaire (HLQ) among older adults who present to the emergency department after a fall: a Rasch analysis.* BMC health services research, 2017. **17**(1): p. 605.
103. Sand-Jecklin, K. and S. Coyle, *Efficiently assessing patient health literacy: the BHLS instrument.* Clinical Nursing Research, 2014. **23**(6): p. 581-600.
104. Haun, J., et al., *Measurement variation across health literacy assessments: implications for assessment selection in research and practice.* Journal of Health Communication, 2012. **17**(sup3): p. 141-159.
105. Miller, B., *Investigating the Construct of Health Literacy Assessment: A Cross-Validation Approach*, in *Graduate School, the College of Education and Psychology and the Department of Educational Research & Administration*. 2018, The University of Southern Mississippi: Ann Arbor, MI.
106. Elsworth, G.R., A. Beauchamp, and R.H. Osborne, *Measuring health literacy in community agencies: a Bayesian study of the factor structure and measurement invariance of the health literacy questionnaire (HLQ).* BMC Health Serv Res, 2016. **16**(1): p. 508.
107. Goodwin, B.C., et al., *Health literacy and the health status of men with prostate cancer.* Psycho-Oncology, 2018. **27**(10): p. 2374-2381.
108. Driessnack, M., et al., *Using the "Newest Vital Sign" to assess health literacy in children.* Journal of Pediatric Health Care, 2014. **28**(2): p. 165-171.
109. Goodman, M.S., et al., *Do subjective measures improve the ability to identify limited health literacy in a clinical setting?* Journal of the American Board of Family Medicine, 2015. **28**(5): p. 584-594.
110. Cavanaugh, K.L., et al., *Performance of a brief survey to assess health literacy in patients receiving hemodialysis.* Clinical Kidney Journal, 2015. **8**(4): p. 462-468.
111. Chew, L.D., et al., *Validation of screening questions for limited health literacy in a large VA outpatient population.* Journal of General Internal Medicine, 2008. **23**(5): p. 561-566.
112. Housten, A.J., et al., *Limitations of the S-TOFHLA in measuring poor numeracy: a cross-sectional study.* BMC Public Health, 2018. **18**(1): p. 405.
113. Kirk, J.K., et al., *Performance of health literacy tests among older adults with diabetes.* Journal of General Internal Medicine, 2012. **27**(5): p. 534-540.
114. Ko, Y., et al., *Development and validation of a general health literacy test in Singapore.* Health Promotion International, 2011. **27**(1): p. 45-51.
115. Kordovski, V.M., et al., *Is the newest vital sign a useful measure of health literacy in HIV disease?* Journal of the International Association of Providers of AIDS Care, 2017. **16**(6): p. 595-602.
116. McNaughton, C., et al., *Short, subjective measures of numeracy and general health literacy in an adult emergency department.* Academic Emergency Medicine, 2011. **18**(11): p. 1148-1155.
117. Morris, N.S., et al., *The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability.* BMC Family Practice, 2006. **7**(1): p. 21.
118. Quinzanos, I., et al., *Cross-sectional correlation of single-item health literacy screening questions with established measures of health literacy in patients with rheumatoid arthritis.* Rheumatology International, 2015. **35**(9): p. 1497-1502.
119. Rawson, K.A., et al., *The METER: a brief, self-administered measure of health literacy.* Journal of General Internal Medicine, 2010. **25**(1): p. 67-71.
120. Wallston, K.A., et al., *Psychometric properties of the brief health literacy screen in clinical practice.* J Gen Intern Med, 2014. **29**(1): p. 119-26.
121. Hadden, K.B., *Health Literacy and Pregnancy: Validation of a New Measure and Relationships of Health Literacy to Pregnancy Risk Factors*. 2012, University of Arkansas for Medical Sciences: Arkansas.

122. Soelberg, J., *Determining the reliability and validity of the newest vitalsign in the inpatient setting*. 2015, Rush University: Chicago, Illinois.
123. Haun, J.N., *Health Literacy: The Validation of a Short Form Health Literacy Screening Assessment in an Ambulatory Care Setting*. 2007, University of Florida: Florida.
124. Messick, S., *Foundations of validity: Meaning and consequences in psychological assessment.* ETS Research Report Series, 1993. **1993**(2): p. i-18.
125. Francis, D.O., et al., *Checklist to operationalize measurement characteristics of patient-reported outcome measures.* Systematic Reviews, 2016. **5**(1): p. 129.
126. American Educational Research Association, et al., *Standards for educational and psychological testing*. 1999, Washington, DC: American Educational Research Association.
127. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing*. 1985: American Educational Research Association.
128. Buchbinder, R., et al., *A validity-driven approach to the understanding of the personal and societal burden of low back pain: development of a conceptual and measurement model.* Arthritis Research & Therapy, 2011. **13**(5): p. R152.
129. McClimans, L., *Interpretability, validity, and the minimum important difference.* Theoretical Medicine and Bioethics 2011. **32**(6): p. 389-401.
130. Kane, M. and B. Bridgeman, *Research on Validity Theory and Practice at ETS*, in *Advancing Human Assessment: The Methodological, Psychological and Policy Contribution of ETS*, R.E. Bennett and M. von Davier, Editors. 2017, Springer Nature: Cham, Switzerland. p. 489-552.
131. Landy, F.J., *Stamp collecting versus science: Validation as hypothesis testing.* American Psychologist, 1986. **41**(11): p. 1183.
132. Spurgeon, S.L., *Evaluating the unintended consequences of assessment practices: Construct irrelevance and construct underrepresentation.* Measurement and Evaluation in Counseling and Development, 2017. **50**(4): p. 275-281.
133. Nutbeam, D., *Health promotion glossary.* Health Promotion International, 1998. **13**(4): p. 349-364.
134. New Zealand Ministry of Health, *Content Guide 2017/18: New Zealand Health Survey*. 2018, NZ Ministry of Health: Wellington, New Zealand.
135. Bo, A., et al., *National indicators of health literacy: ability to understand health information and to engage actively with healthcare providers - a population-based survey among Danish adults.* BMC Public Health, 2014. **14**(1): p. 1095.
136. Busija, L., R. Buchbinder, and R.H. Osborne, *A grounded patient-centered approach generated the personal and societal burden of osteoarthritis model.* Journal of Clinical Epidemiology, 2013. **66**(9): p. 994-1005.
137. Rosas, S.R. and J.W. Ridings, *The use of concept mapping in measurement development and evaluation: application and future directions.* Evaluation and Program Planning, 2017. **60**: p. 265-276.
138. Soellner, R., N. Lenartz, and G. Rudinger, *Concept mapping as an approach for expert-guided model building: The example of health literacy.* Evaluation and Program Planning, 2017. **60**: p. 245-253.
139. WHO Regional Office for Europe. *Health literacy in action*. 2019 [cited 2019 18 October]; Available from: http://www.euro.who.int/en/health-topics/disease-prevention/health-literacy/health-literacy-in-action.
140. Nguyen, T.H., et al., *State of the science of health literacy measures: Validity implications for minority populations.* Patient Educ Couns, 2015.
141. Marmot, M., *Fair society, healthy lives: the Marmot Review: strategic review of health inequalities in England Post-2010.* 2010.
142. Kane, M., *Validity and fairness.* Language testing, 2010. **27**(2): p. 177-182.
143. Carey, G., B. Crammond, and E. De Leeuw, *Towards health equity: a framework for the application of proportionate universalism.* International Journal for Equity in Health, 2015. **14**(1): p. 81.

144. Beauchamp, A., et al., *The effect of obesity prevention interventions according to socioeconomic position: a systematic review.* Obes Rev, 2014. **15**(7): p. 541-54.
145. Beeston, C., et al., *Health inequalities policy review for the Scottish Ministerial Task Force on health inequalities*. 2014: Edinburgh.
146. Addison, M., et al., *Equal North: how can we reduce health inequalities in the North of England? A prioritization exercise with researchers, policymakers and practitioners.* Journal of Public Health, 2018: p. 1-13.
147. Capewell, S. and H. Graham, *Will cardiovascular disease prevention widen health inequalities?* PLoS Medicine, 2010. **7**(8): p. e1000320.
148. Messick, S., *Test validity: A matter of consequence.* Social Indicators Research, 1998. **45**(1-3): p. 35-44.



MY

EBSCOhost

Monday, March 11, 2019 7:57:11 PM

#	Query	Limiters/Expanders	Last Run Via	Results
S28	S24 AND S25 AND S26 AND S27	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	1,036
S27	S12 OR S13 OR S14 OR S15 OR S16 OR S22 OR S23	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	3,396,491
S26	S11 OR S21	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	68,560
S25	S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S18 OR S19 OR S20	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	5,965,966
S24	S1 OR S2 OR S17	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	813,727
S23	(MH "Focus Groups") OR (MH "Interviews as Topic") OR (MH "Data Accuracy")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	79,811
S22	(MH "Psychometrics")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced	69,650

			Search Database - MEDLINE Complete	
S21	(MH "Health Literacy") OR (MH "Health Education") OR (MH "Consumer Health Information")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	65,418
S20	(MH "Self-Assessment")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	11,920
S19	(MH "Health Surveys") OR (MH "Surveys and Questionnaires") OR (MH "Health Care Surveys")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	486,718
S18	(MH "Patient Outcome Assessment") OR (MH "Self Report") OR (MH "Patient Reported Outcome Measures")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	31,421
S17	(MH "Validation Studies as Topic")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE Complete	1,977
S16	TI “focus group*” OR AB “focus group*”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	36,147
S15	TI “think aloud” OR AB “think aloud”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	1,091

S14	TI interview* OR AB interview*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	433,189
S13	TI Psychometric* OR AB Psychometric*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	32,322
S12	TI measur* OR AB measur*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	2,602,779
S11	TI “health literacy” OR AB “health literacy”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	4,293
S10	TI tool* OR AB tool*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	682,713
S9	TI test* OR AB test*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	2,261,528
S8	TI assess* OR AB assess*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	1,782,492
S7	TI “self rated" OR AB “self rated"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced	7,786

			Search Database - Academic Search Complete	
S6	TI “self report*” OR AB “self report*”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	92,555
S5	TI survey* OR AB survey*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	590,730
S4	TI questionnaire* OR AB questionnaire*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	287,547
S3	TI “patient reported outcome*” OR AB “patient reported outcome*”	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	7,191
S2	TI Verif* OR AB Verif*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	276,503
S1	TI valid* OR AB valid*	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - Academic Search Complete	639,560

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1, 4
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2-3 This systematic review is not registered
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Protocol paper: https://bmjopen.bmj.com/content/9/10/e030753
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6-7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6-7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Supplementary file 1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6-8 including Figure 1.
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	7-8
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7-8
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	3, 16
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	5 (Table 1. The five sources of validity evidence)

PRISMA 2009 Checklist

Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	7-8

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	3, 16
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	7-8
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	8-9
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	8-9
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	3, 16
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	8-13
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	8-13
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	3, 16
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8-13
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	13-16
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	17
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	2

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

PRISMA 2009 Checklist

Supplementary file 3: Data extraction framework

Note: The table has been sorted to highlight the studies that directly and indirectly referenced a validity-testing framework - see Column 2

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Dykhuis et al (2019)	Direct to Francis et al (2016) checklist [1]	USA	Test BMNT	REALM-R, REALM-SF	Cronbach's alpha	1	x	x	2	x
Elsworth et al (2016)	Direct to Messick 1992 In Alkin MC [2]	Australia	Test HLQ	x	Cronbach's alpha, Composite reliability	x	x	2	2	x
Hawkins et al (2017)	Direct ref to *Standards* 2014, Kane 1992, Messick 1993 [3-5]	Australia	Test HLQ	x	Inter-rater	1	2	x	x	x
Begoray (2012)	Indirect to *Standards* [6]	Canada	Develop and test health literacy assessment (no name) (9 s-r items, 2 Cloze)	REALM	Cronbach's alpha	1	x	x	2	x

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
DeBello (2016) thesis	Indirect to *Standards* [7]	USA	Test HKACS	x	Cronbach's alpha; Test-retest	2	x	1	4	x
Elder et al (2012)	Indirect to *Standards* [8, 9]	Australia	Test REALM (13 items)	TOFHLA, NVS, Definition scores / AQOL	Cronbach's alpha; Inter-rater	3	x	x	4	1
Goodwin et al (2018)	Indirect to *Standards* [10, 11]	Australia	HLQ	x	Cronbach's alpha	x	x	1	1	x
Morris et al (2017)	Indirectly to *Standards* [12]	Australia	Test HLQ	x	Person separation index (PSI) in IRT	x	x	3	x	x
Osborne et al (2013)	Indirectly to *Standards* [10]	Australia	Develop and test HLQ	x	Composite reliability	7	1	3	1	x

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Baker et al (1999)	x	USA	Develop and test S-TOFHLA	REALM	Cronbach's alpha	1	x	x	1	x
Bann et al (2012)	x	USA	Develop and test HLSI-SF	S-TOFHLA, self-report questions	Cronbach's alpha	3	x	2	1	x
Barber et al (2009)	x	Australia	Test REALM, TOFHLA, NVS	AQOL	x	x	x	x	3	x
Cavanaugh et al (2015)	x	USA	Test BHLS (3 items)	REALM, S-TOFHLA, MMSE / CHeKS, PiKS	Cronbach's alpha	x	x	x	6	x
Chesser et al (2014)	x	USA	Test S-TOFHLA	x	x	1	1	x	x	x
Chew et al 2004	x	USA	Develop and test 3 screening questions	S-TOFHLA	x	2	x	x	1	x
Chew et al 2008	x	USA	Test 3 screening questions	S-TOFHLA, REALM	x	x	x	x	1	x
Curtis et al (2015)	x	USA	Develop and test CHAS	S-TOFHLA, REALM, NVS, MMSE / self-reported health status, SF-36, PROMIS short form	Cronbach's alpha; IRT TIF; Omega analysis	2	x	3	6	x

Authors: Hawkins M; Elsworth GR; Hoban E; Osborne RH. (2019)

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
				emotional health						
Dageforde et al (2015)	x	USA	Test SLS (3 items)	REALM, S-TOFHLA	Cronbach's alpha; Wilcoxen signed rank test	1	x	x	2	x
Davis et al (1991)	x	USA	Test REALM	PIAT-R, SORT	Test-retest; Inter-rater	1	x	x	1	x
Davis et al (1993)	x	USA	Develop and test REALM-SF	PIAT-R, SORT-R, WRAT-R	x	1	x	x	1	x
Driessnack et al (2014)	x	USA	Test NVS	N of children's books	Cronbach's alpha	x	x	x	3	x
Goodman et al (2015)	x	USA	Test BHLS (3 items)	REALM-R, NVS	x	x	x	x	1	x
Hadden (2012) thesis	x	USA	Test HLSI-SF	S-TOFHLA, Perceptions of Difficulty with Health Literacy Skills	x	x	x	x	5	x
Harper (2013) thesis	x	USA	Develop and test a new health literacy assessment [no name]	S-TOFHLA	Cronbach's alpha	4	x	2	3	x

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Haun (2012)	x	USA	Test S-TOFHLA, REALM, BRIEF (4 items)	x	Cronbach's alpha	x	x	1	4	x
Haun (2007) thesis	x	USA	Test BRIEF (4 items)	S-TOFHLA and REALM	Cronbach's alpha	x	x	x	3	x
Housten et al (2018)	x	USA	Test S-TOFHLA	SNS, GL (GL1, GL2, GL3)	x	x	x	x	1	x
Jordan et al (2013)	x	Australia	Develop and test HeLMS	x	Cronbach's alpha; Test-retest	3	1	3	2	x
Kirk et al (2012)	x	USA	Test REALM-SF, NVS	S-TOFHLA	x	x	x	x	2	x
Ko et al (2012)	x	Singapore	Develop and test HLTS	NVS	Cronbach's alpha	x	x	x	4	x
Kordovski et al (2017)	x	USA	Test NVS	REALM, SILS	Cronbach's alpha	x	x	x	7	x
McCormack et al (2010)	x	USA	Develop and test HLSI	S-TOFHLA, self-report questions	Cronbach's alpha	3	x	2	3	x
McNaughton et al (2011)	x	USA	Test SLS (3 items) and SNS (8 items)	S-TOFHLA, REALM, WRAT4	Cronbach's alpha	x	x	x	3	x
Miller (2018) thesis	x	USA	Test HLSI (Cloze only), NVS, S-TOFHLA	x	Cronbach's alpha	x	x	1	1	x

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Morris et al (2006)	x	USA	Test SILS 1	S-TOFHLA	x	x	x	x	1	x
Parker et al (1995)	x	USA	Develop and test TOFHLA	WRAT-R, REALM	Cronbach's alpha; Split halves coefficient	1	x	x	1	x
Quinzanos et al (2015)	x	USA	test SILS 1 and SILS 2	REALM and S-TOFHLA	x	x	x	x	1	x
Rawson et al (2010)	x	USA	Develop and test METER	REALM	Cronbach's alpha	x	x	x	2	x
Sand-Jecklin et al (2014)	x	USA	Develop and test BHLS (5 items)	S-TOFHLA	Cronbach's alpha	x	x	2	4	x
Shaw et al (2014)	x	USA	Develop and test remote admin health literacy assessment	S-TOFHLA	x	1	x	x	1	x
Soelberg (2015)	x	USA	Test NVS	S-TOFHLA	Cronbach's alpha	x	x	x	2	x
Sørensen et al (2013)	x	Netherlands	Develop and test HLS-EU-Q	x	Cronbach's alpha	7	1	1	x	x
Wallace et al (2006)	x	USA	Test 3 screening questions	REALM	x	x	x	x	2	x
Wallston et al (2014)	x	USA	Test BHLS (3 items)	S-TOFHLA	Cronbach's alpha; Inter-rater	x	x	x	4	x

Author	Reference to validity testing framework	Country	HL assessment/s under investigation	Comparator HL assessment/s	Reliability	Test content	Response processes	Internal structure	Relations to other variables	Validity and the consequences of testing
Weiss et al (2005)	x	USA	Develop and test NVS	TOFHLA	Cronbach's alpha	3	x	x	3	x
Zhang et al (2009)	x	Singapore	Develop and test FHLT (21 items)	REALM	Cronbach's alpha; Test-retest	3	1	1	5	x
Totals						**52**	**7**	**28**	**107**	**1**

References

1. Francis, D.O., et al., *Checklist to operationalize measurement characteristics of patient-reported outcome measures.* Systematic Reviews, 2016. **5**(1): p. 129.
2. Alkin, M.C., *Encyclopedia of educational research*. Vol. 3. 1992: Macmillan.
3. AERA, APA, and NCME, *Standards for educational and psychological testing*. 2014, Washington, DC: American Educational Research Association.
4. Kane, M.T., *An argument-based approach to validity.* Psychological Bulletin, 1992. **112**(3): p. 527-535.
5. Messick, S., *Foundations of validity: Meaning and consequences in psychological assessment.* ETS Research Report Series, 1993. **1993**(2): p. i-18.
6. Hubley, A.M. and B.D. Zumbo, *A dialectic on validity: Where we have been and where we are going.* The Journal of General Psychology, 1996. **123**(3): p. 207-215.
7. Waltz, C.F., O.L. Strickland, and E.R. Lenz, *Measurement in nursing and health research*. 2010: Springer publishing company.
8. Abedi, J., C.H. Hofstetter, and C. Lord, *Assessment accommodations for English language learners: Implications for policy-based empirical research.* Review of Educational Research, 2004. **74**(1): p. 1-28.
9. LaCelle-Peterson, M.W. and C. Rivera, *Is it real for all kids? A framework for equitable assessment policies for English language learners.* Harvard Educational Review, 1994. **64**(1): p. 55.
10. Buchbinder, R., et al., *A validity-driven approach to the understanding of the personal and societal burden of low back pain: development of a conceptual and measurement model.* Arthritis Research & Therapy, 2011. **13**(5): p. R152.
11. Hawkins, M., G.R. Elsworth, and R.H. Osborne, *Application of validity theory and methodology to patient-reported outcome measures (PROMs): building an argument for validity.* Quality of Life Research, 2018. **27**(7): p. 1695-1710.
12. Hawkins, M., et al., *The Health Literacy Questionnaire (HLQ) at the patient-clinician interface: a qualitative study of what patients and clinicians mean by their HLQ scores.* BMC Health Services Research, 2017. **17**(1): p. 309.

Supplementary file 4: Detail of data extraction framework

Data were extracted in Excel. These are the data extraction category headings from the Excel spreadsheet.

1. Evidence based on test content

1. Test content evaluated: yes/no/unclear

1. Test content: literature review

1. Test content: prior existing measures of the construct

1. Test content: expert review

1. Test content: participant involvement in construct / item development - structured workshops, concept mapping

1. Test content: participant involvement in construct / item development - interviews

1. Test content: participant feedback processes about items

1. Test content: construct description (incl. high/low descriptors)

1. Test content: item intent descriptions

1. Test content: examination of administration methods

1. Test content: other method (e.g., Item difficulty)

2. Evidence based on response processes

2. Response processes evaluated: yes/no/unclear

2. Response processes - respondents: cognitive interviews

2. Response processes - respondents: think aloud protocols

2. Response processes - respondents: recording and timing responses to items

2. Response processes - users: cognitive interviews

2. Response processes - users: think aloud protocols

2. Response processes - users: recording and timing responses to items

2. Response processes: other method (e.g., determining construct irrelevant factors and construct underrepresentation)

3. Evidence based on internal structure

3. Internal structure evaluated: yes/no/unclear

3. Internal structure: exploratory factor analysis (EFA)

3. Internal structure: confirmatory factor analysis (CFA)

3. Internal structure: multi-group factor analysis (MGFA) (SEM, measurement invariance)

3. Internal structure: correlation patterns and multi-trait scaling analysis (inter-item, item-total and item-remainder correlations)

3. Internal structure: differential item functioning (DIF)

3. Internal structure: other method

4. Evidence based on relations to other variables

4. Relations to other variables evaluated: yes/no/unclear

4. Relations to other variables: convergent validity (between measures of the same or similar construct)

4. Relations to other variables: discriminant validity

4. Relations to other variables: test-criterion relationships (how accurately test scores predict criterion performance)

4. Relations to other variables: group differences (relationships with background characteristics such as demographics information)

4. Relations to other variables: validity generalisation (e.g., meta-analyses / statistical summaries of past studies; cumulative databases)

4. Relations to other variables: nomological networks

4. Relations to other variables: other method

5. Evidence based on validity and the consequences of testing

5. Consequences of testing evaluated: yes/no/unclear

5. Consequences of testing: methods to test for consequential validity (intended consequences e.g., benefits)

5. Consequences of testing: methods to test for consequential validity (unintended consequences e.g., negative effects)

5. Consequences of testing: methods to test for construct underrepresentation

5. Consequences of testing: methods to test for construct-irrelevant components

5. Consequences of testing: methods to test for claims made beyond the intended score interpretation

5. Consequences of testing: methods to test for consequences for clinical implications

5. Consequences of testing: other methods to test consequential validity

5. Consequences of testing: other method (e.g., fairness - low/high-stakes consequences)

Supplementary file 5 – Supplementary Tables 1 to 4

Supplementary Table 1. Evidence based on test content

Number of instances of evidence based on test content across all studies		
Method to generate evidence		
Literature review	4	8%
Existing measures of the construct	8	15%
Expert review	14	27%
Participant involvement:		
Concept mapping	3	6%
Interviews	2	4%
Participant feedback processes about items	4	8%
Construct descriptions (e.g., high/low)	4	8%
Item intent descriptions	1	2%
Examination of administration methods	3	6%
Other method (e.g., item difficulty):		
Item difficulty	5	10%
Items tested against item intents	1	2%
IRT analysis for item selection within domains	1	2%
Item selection based on hospital medical texts	1	2%
Item selection based on HL conceptual model	1	2%
Total instances of evidence based on test content	*52*	*100%*

Supplementary Table 2. Evidence based on response processes

Number of instances of evidence based on response processes across all studies		
Method to generate evidence		
With respondents:		
Cognitive interviews	3	43%
Recording and timing responses to items	3	43%
With users:		
Cognitive interviews	1	14%
Total instances of evidence based on response processes	*7*	*100%*

Supplementary Table 3. Evidence based on internal structure

Number of instances of evidence based on internal structure across all studies		
Method to generate evidence		
Exploratory factor analysis (incl. PCA*)	7	25%
Confirmatory factory analysis (incl. IRT** item discriminations)	7	25%
Multi-group factor analysis	1	4%
Correlation patterns / multi-trait scaling analysis:		
Tetrachoric correlations	1	4%

Inter-item correlations	1	4%
Item-total correlations	1	4%
Item-remainder correlations	2	7%
Differential item functioning	3	11%
Other method:		
Very Simple Structure	1	4%
Velicer's Minimum Average partial criterion	1	4%
Rasch analysis (overall fit, individual person/item fit)	1	4%
Intra-factor correlations	1	4%
IRT for item discriminations	1	4%
Total instances of evidence based on response processes	*28*	*100%*

*PCA = principal component analysis; **IRT = item response theory

Supplementary Table 4. Evidence based on relations to other variables

Summary of number of instances of evidence based on relations to other variables across all studies		
Type of evidence		
Convergent evidence	57	53%
Discriminant evidence	3	3%
Criterion-referenced evidence	17	16%
Evidence for group differences	30	28%
Evidence for generalisation	0	0%
Total instances of evidence based on relations to other variables	107	100%
Number of instances of evidence based on relations to other variables across all studies		
Convergent evidence (relationships between items and scales of the same or similar structure) (n=38 studies):		
Spearman's correlation coefficient	11	19%
Pearson correlation coefficient	11	19%
Linear regression models	5	9%
Logistic regression models	2	4%
Receiver operating characteristic / Area under the ROC (AUROC)	11	19%
Wilcoxen signed rank test	2	4%
Cross tabulations / calculated agreement and disagreement	2	4%
Goodman-Kruskal gamma correlation	1	2%
Bland-Altman plots	1	2%
Cohen's Kappa	1	2%
Sensitivity and specificity	1	2%
Stratum-specific likelihood ratios	1	2%
Unnamed / unclear correlation calculations with similar measures	8	14%
Total instances of convergent evidence	*57*	*100%*
Discriminant evidence (measures of different constructs are sufficiently uncorrelated) (n=2 studies)		
Comparison of AVE and shared variance between HLQ scales	1	33%
Pearson correlation coefficient	1	33%%
Multiscale factor analysis	1	33%
Total instances of discriminant evidence	*3*	*100%*

Criterion-referenced evidence (how accurately test scores predict criterion performance) (n=9 studies):		
Spearman's correlation coefficient	2	12%
Pearson correlation coefficient	1	6%
Linear regression models	6	35%
Logistic regression models	2	12%
ROC/AUROC	1	6%
Chi-squared test of independence	3	18%
ANOVA	1	6%
Cohen's d	1	6%
Total instances of criterion-referenced evidence	*17*	*100%*
Evidence for group differences (relationships of test scores with background characteristics such as demographic information) (n=19 studies):	**N**	**%**
Linear regression models	4	13%
Logistic regression models	3	10%
Univariate associations	1	3%
Spearman's correlation coefficient	1	3%
Chi-squared test	3	10%
Analysis of variance (ANOVA)	5	17%
Analysis of covariance (ANCOVA)	1	3%
Cross tabulations	1	3%
Area under the ROC (AUROC)	1	3%
Kruskal-Wallis test	1	3%
Mann-Whitney U test	2	7%
Goodman-Kruskal gamma correlation	1	3%
Independent sample t-test	3	10%
Exploratory partial correlation analysis	1	3%
Bayesian fit statistics	1	3%
Descriptive statistics (sub-group differences)	1	3%
Total instances of evidence of group differences	*30*	*100%*
Evidence for generalisation (degree to which evidence can be generalised to a new situation) (n=0 studies):	**N**	**%**
Only research synthesis-type studies - see validity generalisation in the *Standards*.	0	0%