Received: 9 June 2021 | Revised: 22 October 2021 | Accepted: 14 December 2021

DOI: 10.1111/acem.14428

SYSTEMATIC REVIEW

A systematic review of instruments to measure health literacy of patients in emergency departments

Gijs Hesselink PhD[1,2] | Joey Cheng MSc[1] | Yvonne Schoon MD, PhD[1,3]

[1]Department of Emergency Medicine, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

[2]IQ healthcare, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

[3]Department of Geriatrics, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

Correspondence
Gijs Hesselink, PhD, IQ healthcare, Radboud Institute for Health Sciences, Radboud University Medical Center, P.O. Box 9101, 114 IQ healthcare, 6500 HB Nijmegen, The Netherlands.
Email: gijs.hesselink@radboudumc.nl

Abstract

Objectives: Knowledge of patient's health literacy (HL) in the emergency department (ED) can facilitate care delivery and reduce poor health outcomes. This systematic review investigates HL measurement instruments used in the ED and their psychometric properties, accuracy in detecting limited HL, and feasibility.

Methods: We searched in five biomedical databases for studies published between 1990 and January 2021, evaluating HL measurement instruments tested in the ED on internal consistency, criterion validity, diagnostic accuracy, or feasibility. Reviewers screened studies for relevance and assessed methodologic quality with published criteria. Data were synthesized around study and instrument characteristics and outcomes of interest.

Results: Of the 2,376 references screened, seven met our inclusion criteria. Studied instruments varied in objective (n = 5) and subjective (n = 6) measurement of HL skills, and in HL constructs measured. The Brief Health Literacy Screen (BHLS) and the Subjective Numeracy Scale demonstrate acceptable and good internal consistency across studies. None of the instruments perform consistently well on criterion validity. The Rapid Estimate of Adult Literacy in Medicine–Revised and the Newest Vital Sign, both objective tests with short administration times, demonstrate good accuracy in one study with high risk of bias. The BHLS, a short subjective measure, shows moderate accuracy across studies including one with low risk of bias.

Conclusions: Several short instruments seem valid in measuring HL and accurate in detecting limited HL among ED patients, each with its practical advantages and disadvantages and specific measurement of HL. Additional research is necessary to develop a robust evidence base supporting these instruments.

INTRODUCTION

Health literacy (HL) is the individuals' capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions.[1] Important HL skills include reading and writing ability and numeracy skills. HL is an important determinant for health behavior, including planning and adjusting lifestyle, participation in medical decision making, treatment

Supervising Editor: Kristin L. Rising, MD, MSHP, FACEP.

adherence, and recognizing when and how to access health care services.[2,3]

Limited HL is increasingly perceived as a global public health concern.[4,5] Levels of HL have been surveyed in industrialized countries such as the United States, Canada, Australia, and in the European Union (EU), with the prevalence of limited HL varying from 29% to 62%.[5–7] Limited HL has been associated with a wide range of adverse health effects, including worse self-management skills,[8,9] greater risk of hospitalization, ED visits and lack of preventative care, worse health status, and lower quality of life.[10] The prevalence of limited HL in the emergency department (ED) is wide ranging across studies but generally high with estimates up to 88% depending on the visitor type and on the measurement instruments used.[11,12] Among ED patients, limited HL is associated with worse health status, higher number of health care utilization such as ED recidivism, and higher risk of death.[11–14] Although the explanatory mechanisms underlying the relations between limited HL and adverse outcomes are rather complex, many of the poor outcomes associated with limited HL may be caused or exacerbated by inadequacies in clinician–patient communication.[11,15,16] Not recognizing low literacy among ED visitors by clinicians can lead to suboptimal patient involvement in and receipt of care. Available ED reading materials and standard information by provided clinicians are often too complex for this patient group, thereby increasing the risk of patients being un- or misinformed.[11,16] Moreover, clinicians may approach treatment options differently than patients based on their assumptions of patient's HL and disease knowledge.[16,17] Timely recognition of limited HL in the ED can be a first important step in overcoming these inadequacies and the negative outcomes associated with limited HL.

Over the past decade, there has been a growing effort in developing HL measurement instruments aimed at anticipating on limited HL cases.[18–20] Moreover, the 2004 Institute of Medicine report on HL recommended that HL assessment should be part of health care information systems to facilitate large-scale studies of the effects of HL as well as the evaluation of interventions targeting limited HL.[1] However, there is no actual overview and critical appraisal of HL measurement instruments used in the ED setting. Alqudah et al.[21] reviewed HL measurement instruments in the ED, but their review consisted of publications until 2011 on specific instruments using word recognition procedures with demonstrated concurrent validity and a maximum administration time of 5 min.

An actual and more comprehensive overview of HL measurement instruments studied in the ED informs clinicians on available instruments and may contribute to the identification of an instrument that is favorable for use in their ED. Therefore, our aim was to systematically review scientific literature on instruments used in the ED and their psychometric properties, accuracy in detecting limited HL, and feasibility.

METHODS

We planned and reported this systematic review in accordance with the reporting guidance provided in the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA).[22] The protocol of this review was established a priori and registered on the International Prospective Register of Systematic Reviews (PROSPERO) website with ID CRD42020174997.

Data sources and strategy

We searched for articles published between January 1, 1990, and March 18, 2020, in the following databases: PubMed (including MEDLINE), Cumulative Index to Nursing and Allied Health Literature, Cochrane Library, EMBASE, and PsychInfo. Due to the COVID-19 pandemic, study activities were postponed. An additional search was therefore performed to find relevant articles published between March 19, 2020, and January 11, 2021. Search strategies (Appendix S1) comprised a combination of key search terms related to the concepts of "emergency department," "health literacy," and "measurement tools." Specific HL screenings instruments identified in previous literature studies as the criterion or reference standard were also included in the search strategies, namely, the Test of Functional Health Literacy among Adults (TOFHLA), the Rapid Estimate of Adult Literacy in Medicine (REALM), and the Wide Range Achievement Test (WRAT).[23] Additional relevant articles were searched for by manually checking the reference lists of eligible articles and review articles.

Eligibility criteria and study selection

A fourth-year medical student (JC) and a senior health scientist (GH) independently assessed inclusion eligibility of the retrieved references. References were included if they: 1) were published full text and with an abstract in English; 2) evaluated one or more instruments aimed at screening patient's HL in the ED; and 3) reported data on one or a combination of the following outcomes: the instruments' internal consistency, criterion validity (assessed by the correlation of the instrument with the short or extended version of the TOFHLA, the REALM, or the WRAT), diagnostic accuracy (i.e., its ability to discriminate between patients with and without limited HL), or feasibility. Conference abstracts and publications without original data were excluded from the analysis. After initial screening of the titles and abstracts, both reviewers read the full texts of included articles and screened these for eligibility. Discrepancies were discussed and taken to a third person (YS) if no agreement could be reached.

Data extraction

Data were extracted using a standardized form that assessed study characteristics (e.g., country, study setting, population, sample size), instrument description, reference methods and results. Data regarding internal consistency included Cronbach's alpha values. Data regarding criterion validity included correlation coefficients (Pearson's

r, Spearman's rho, Kendall's tau depending on type of data). Data regarding diagnostic accuracy included: area under the curve (AUC) scores as the derived summary measure for diagnostic accuracy, sensitivity, and specificity scores and related 95% confidence intervals (CIs). Data regarding feasibility included: the mean total administration time (AT), the mean time on test (TOT), the proportion of administrations with interruptions (PI), and the mean length of interruptions (TOI) per test. Data were extracted by JC and reviewed for completeness and accuracy by GH. Discrepancies were resolved by discussion.

Assessment of study quality

Methodologic quality was assessed independently by JC and GH. Disagreements were resolved by consensus or by involving YS as required. The methodologic quality and applicability of diagnostic accuracy studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2).[24] This tool assesses the risk of bias within four domains: patient selection, index test, reference standard, and flow and timing (Table S1). Risk of bias was assessed for each domain by answering signaling questions with "yes," "no," or "unclear" to assist judgments. Concerns regarding applicability were also determined for the first three domains. Risk of bias and applicability concerns per domain were rated as "high," "low," or "unclear." If, within one domain, all signaling questions were answered "yes" then risk of bias for that domain was judged "low." If one or more signaling questions were answered "no" then risk of bias was judged "high." Studies were overall judged "low risk of bias" or "low concerns regarding applicability" if risk of bias and applicability concerns were scored "low" on all domains relating to either bias or applicability. Studies were judged "at risk of bias" or as having "concerns regarding applicability" if a study was judged "high" or "unclear" on one or more domains.[24] Box H of the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) was used to assess the quality of studies reporting criterion validity. Overall quality was determined by taking the lowest rating of any item in the box (i.e., the "worst score counts" principle).[25] Inter-rater agreement was calculated for the scores on the QUADAS-2 signaling questions combined and for the scores on the COSMIN checklist items by between-group kappa agreement, using the assessments from each reviewer before resolution of disagreements. Publication bias was not assessed because of the small numbers of studies for any given instrument. Moreover, methods to detect publication bias in studies assessing diagnostic accuracy data are considered unreliable.[26,27]

Data synthesis and analysis

Data were organized in tabular form to describe study characteristics and quality, instrument characteristics, comparators, and outcomes of interest. Descriptive statistics were used to summarize psychometric outcome data and compare them against set criteria. For internal consistency, we used Cronbach's alpha cutoffs: >0.9 excellent, >0.8 good, >0.7 acceptable, >0.6 questionable, >0.5 poor, and <0.5 unacceptable.[28] For criterion validity, we used correlation coefficient cutoffs: >0.7 high, 0.5–0.7 moderate, and <0.5 low.[29] For diagnostic accuracy, we used AUC score cutoffs: >0.8 good, 0.6–0.8 moderate, and <0.6 poor.[30] Instruments were also categorized based on their mode of measurement into *objective measurement of HL* derived by one or more direct tests of skills and *subjective measurement of HL* by individuals' self-report of perceived skills.[19] Heterogeneity in clinical instruments and outcome reporting limited our ability to conduct a meta-analysis. Instead, we conducted a descriptive analysis of the psychometric, diagnostic, and feasibility results of each study.

RESULTS

Search results

Our initial search identified 2,145 records. The additional search identified 231 records resulting in a total of 2,376 records. After exclusion of duplicates, 1,578 records were screened by title and abstract. Seventy-one full-text articles were retrieved and reviewed, of which 64 were excluded. Most excluded articles ($n = 46$) did not report data on our outcomes of interest. Other articles were excluded because instruments were not evaluated in the ED, a full-text copy was not available and content turned out to be a conference abstract. No additional relevant articles were found from the reference lists of the articles that were reviewed in full-text and from review articles. Consequently, the final set comprised seven unique published studies that underwent full-text extraction (Figure 1).

Study characteristics

Characteristics of the seven included studies are summarized in Tables 1 and 2. All studies were published between 2011 and 2020 and performed in the United States. The vast majority were single center,[31–34,36] prospective observational cohort studies,[31–37] conducted in urban academic EDs.[31,32,34,36] One study was conducted in four urban academic EDs.[37] Another study was performed in one pediatric ED.[33] One study described a secondary analysis of prospectively collected cohort data from multiple EDs in the United States.[35] Study participants consisted of non–critically ill patients, mostly (older) adults. One study focused on the caregivers of children aged 12 years and younger.[33] All studies consisted of English-speaking participants. Three studies also included Spanish-speaking participants.[33,36,37] Sample sizes varied from 202 to 2,770 participants. The included studies made efforts to minimize the impact of confounding effects on instruments' validity and diagnostic accuracy, mostly by excluding patients with specific mental, cognitive or physical conditions, or impairments that are known to impede an accurate measurement of HL. Two studies described counterbalanced testing of instruments to reduce bias due to test fatigue.[32,33]

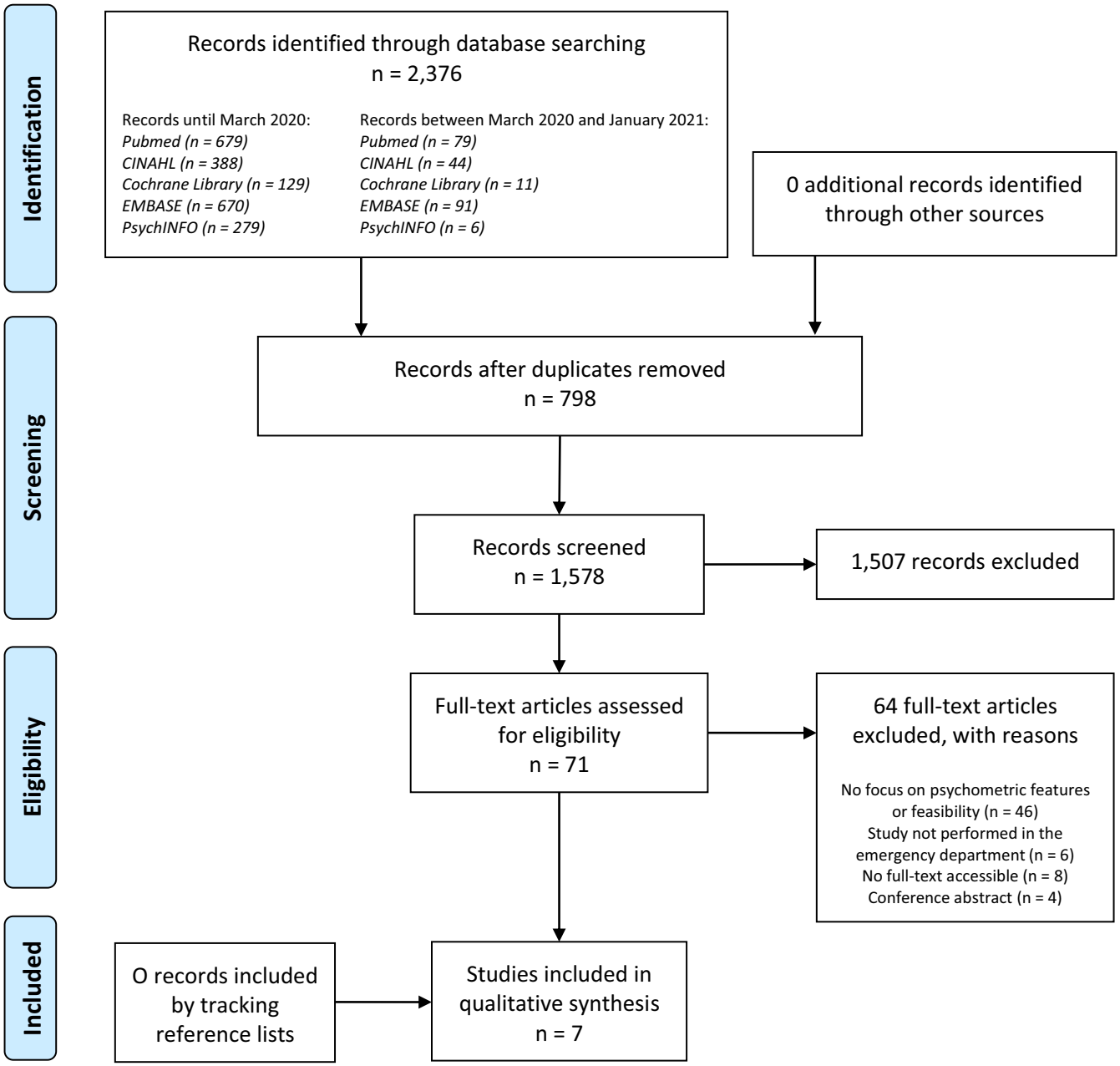

FIGURE 1 Flow chart of the study selection process

One study stratified instrument correlation scores by lower and higher educational level.[33] Three studies evaluated the internal consistency of instruments.[31,32,35] Five reported data on criterion validity.[31–35] Two studies evaluated the diagnostic accuracy[32,37] and the feasibility of instruments.[32,36] Most of the included studies (n = 6; 86%) evaluated multiple instruments[31,32,34–37] or different instrument versions (short and extended).[32,34,35,37]

Study quality

The summarized methodologic quality of each study is presented in Table 2. Details on the study quality are provided in Tables S1 and S2. Inter-rater agreement for the scores on the QUADAS-2 signaling questions was high with a kappa score of 0.89. The percentage agreement between both raters for scores per risk of bias domain varied between 80 and 100. Inter-rater agreement for the scores on the COSMIN checklist items was high with a kappa score of 0.84. None of the three diagnostic studies assessed with the QUADAS-2[31,32,37] scored "low risk of bias" on all four domains. All three showed high risk of bias in the patient selection: i.e., patients were sampled consecutively[31,32] and inappropriate exclusions were not avoided because second-grade reading level patients were excluded, which may have influenced accuracy findings.[37] Only one study showed sufficient information to determine appropriate conduct and interpretation of the index tests.[37] In all three studies the

TABLE 1 Characteristics of the included studies

First author, year	Setting (country)	Design	Population					Sampling	Excluded
				Size	Age (years), mean (±SD)	% Female	% Race		
McNaughton, 2011[31]	Single urban academic ED (United States)	Prospective observational cohort	Adults	207	46[a]	55	68 W; 27 B; 6 O	Convenience	Critically ill; non-English speakers
Carpenter, 2014[32]	Single urban academic ED (United States)	Prospective observational cohort	Adults	435	45 (±16)	55	68 B; 31 W; <1 A; 1 O	Convenience	Critically ill; distressed; altered mental status; aphasia; mentally or visually impaired
Morrison, 2014[33]	Single (sub)urban pediatric ED (United States)	Prospective observational cohort	Caregivers of nonchildren (≤12 years)	501	32[a]	85	47 W; 37 B; 10 H; 5 O	Random	Non-English or Spanish speakers; distressed child; child presented for maltreatment or non-accidental trauma; <5th grade reading level
Kiechle, 2015[34]	Single suburban ED (United States)	Prospective observational cohort	Adults	400	38 (±14)[a]	58	63 W; 30 B; 5 H; 3 O	Sequential block enrollment (based on education levels)	Critically ill; non-English speakers; decisionally or visually impaired; intoxicated
McNaughton, 2015[35]	Multiple EDs (United States)	Secondary analysis of six prospective study cohorts	Adults	207	NR	NR	NR	Convenience	NR
McGuinness, 2020[36]	Single urban academic ED (United States)	Prospective observational cohort	Older adults (55–90 years)	104[b]; 98[c]	68 (NR)[b]; 69 (NR)[c]	51[b]; 52[c]	NR	Convenience	Critically ill; non-English or Spanish speakers; distressed; altered mental status, reading, speech or cognitive disability
Merchant, 2020[37]	Four urban academic EDs (United States)	Prospective observational cohort	Adults (18–64 years)	2,770	44 (±12)	61	56 W; 43 B; 1 O	Random	Critically ill; non-English or Spanish speakers; intoxicated, physically, mentally or cognitively impaired, <2nd grade reading level

Abbreviations: A, Asian; B, Black or Afro American; H, Hispanic or Latino; NR, not reported; O, other; W, White or Caucasian.

[a]Median score.

[b]Completed the NVS.

[c]Completed the SAHL.

TABLE 2 Tested instruments, outcomes of interest, and methodologic quality per study

First author, year	Tested instruments	Language	Outcomes				QUADAS-2		COSMIN Box H
			IC	CV	DA	F	RB	AC	
McNaughton, 2011[31]	BHLS; SNS-8	English	√	√	√		+	−	Fair
Carpenter, 2014[32]	REALM-R; NVS; SILS questions[a]; BHLS	English	√	√	√	√	+	−	Fair
Morrison, 2014[33]	NVS	English; Spanish		√			NA	NA	Fair
Kiechle, 2015[34]	S-TOFHLA; NVS; SILS questions[a]; BHLS; REALM-R; METER	English		√			NA	NA	Poor
McNaughton, 2015[35]	SNS-3; SNS-8	English	√	√			NA	NA	Fair
McGuinness, 2020[36]	NVS; SAHL	English; Spanish				√	NA	NA	NA
Merchant, 2020[37]	SILS questions[a]; BHLS	English; Spanish			√		−	−	NA

Abbreviations: BHLS, Brief Health Literacy Screen; CV, Criterion validity; COSMIN, COnsensus-based Standards for the selection of health status Measurement Instruments; DA, diagnostic accuracy; F, feasibility; IC, internal consistency; METER, Medical Term Recognition Test; NA, not applicable; NVS, Newest Vital Sign; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies; REALM-R, Rapid Estimate of Adult Literacy in Medicine-Revised; SNS, Short Numeracy Scale; SILS, Single Item Literacy Screener; S-TOHFLA, Short Test of Functional Health Literacy among Adults. +, At risk of bias, concerns regarding applicability; −, low risk of bias, low concerns regarding applicability.

[a]All three SILS questions: i.e., 1) "How often do you have someone (like a family member, friend, hospital or clinic worker, a caregiver, or anyone else) help you read materials given to you by the hospital, clinic, or your health care provider?" 2) "How confident are you in filling out medical forms by yourself?" 3) "How often do you have problems learning about your medical condition or health because of difficulty reading and understanding written information given to you by the hospital, clinic, or your health care provider?"

reference standard used may have introduced bias as the standards used are not the undisputable gold standard (i.e., estimates of test accuracy are not based on the assumption that the standard is 100% sensitive and specific disagreements between the reference standard and index test result from incorrect classification by the index test). Furthermore, in one study screeners collected scores on both the index tests and the reference standard, thereby increasing the risk of incorporation bias, which can also falsely increase sensitivity and specificity.[32] All three showed appropriate flow and timing of the index test and reference standard. Applicability concerns were considered low for all three studies. Of the five studies assessed with the COSMIN Box H,[31–35] four scored "fair,"[31–33,35] and one scored "poor."[34]

Instruments

Characteristics

Table 2 provides an overview of the studied HL instruments. In total, 11 unique instruments were evaluated on either internal consistency, criterion validity, diagnostic accuracy, and/or feasibility. Five instruments used direct tests to assess individuals' HL skills (objective measurement) by letting them solve tasks dealing with print literacy, numeracy, or oral literacy. The Short Test of Functional Health Literacy in Adults (S-TOFHLA) includes a condensed 36-item version of the TOFHLA testing reading comprehension.[34] The Rapid Estimate of Adult Literacy in Medicine-Revised (REALM-R) is a shortened version of the REALM, which tests individuals' pronunciation of eight medical words (e.g., anemia and osteoporosis).[32,34] The Newest Vital Sign (NVS) consists of a fictitious ice cream nutrition label that is handed to the patient, as the interviewer asks six accompanying questions to assess literacy and numeracy skills.[32,34,36] The Medical Term Recognition Test (METER) contains a list of 40 medical words mixed in with nonwords. The patient is asked to identify the real words.[34] The Short Assessment of Health Literacy (SAHL) includes 18 interviewer-administered items designed to assess patients' ability to read and understand common medical terms. Each item contains a medical term printed in boldface and two association words (i.e., the key and the distracter). Correct answers are determined by both correct pronunciation and accurate association.[36,38]

Six instruments used the elicitation of self-reported perceived skills in print literacy and numeracy. The three Single Item Literacy Screener (SILS) consist of one question (5-point Likert scale) to assess individuals' self-perceived 1) need for help in reading hospital materials, 2) confidence in filling out medical forms, and 3) difficulty understanding written information in trying to learn more about a medical condition.[32,34,37] The Brief Health Literacy Screen (BHLS) are the three SILS-questions combined in one instrument.[31,32,34,37] The two Subjective Numeracy Scale (SNS) versions are measures of individuals' perceived ability to perform various mathematical tasks and preference for the use of numerical versus prose information. The SNS-8 consists of eight questions (6-point Likert scale) asking individuals to assess their numeracy skills in different contexts and their preferences for the presentation of numerical and probabilistic information.[31,35] The SNS-3 is a condensed version of the SNS-8 with two questions on numeracy skills and one on subject preference.[35] Thresholds for detecting limited HL were provided

for each of the studied instruments based on previously published scoring rules for the instrument. All instruments were administered in English. In three studies, the NVS, the SAHL, the BHLS, and the three SILS questions were also administered in Spanish.[33,36,37]

Internal consistency

Three instruments were tested on internal consistency (Table 3). The BHLS demonstrated acceptable internal consistency with Cronbach's alpha scores of 0.74 and 0.78.[31,32] The same applied to the SNS-3 with a reported alpha of 0.78.[35] The SNS-8 showed good internal consistency with alpha scores of 0.82 and 0.83 reported in two studies by McNaughton et al.[31,35]

Criterion validity

Ten instruments were tested on validity against one or more reference standards (Table 3). Only the METER showed a high correlation with the REALM-R ($r = 0.73$).[34] This may be explained by the fact that both instruments use the same medical words in testing literacy.[34] The METER showed moderate validity ($r = 0.53$) with the S-TOFHLA as the reference standard.[34] The S-TOFHLA showed moderate validity against the REALM-R as reference standard and vice versa.[32,34] The validity of the NVS was poor to moderate with correlation coefficients varying from 0.45 to 0.62 against the S-TOFHLA and the REALM-R as reference standards. The SILS questions showed poor validity against three reference standards (i.e., the S-TOFHLA, the REALM-R, and the WRAT-4 mathematical subtest), both individually (r = 0.38–0.43)[34] and as questions combined in the BHLS (r = 0.24–0.49).[31,32,34] For the SNS-8 and SNS-3, the correlations with S-TOFHLA and REALM-R were poor (r = 0.36–0.40).[31,35] In contrast, both instruments showed moderate correlation with the WRAT-4, which included a substantial items on numeracy like the SNS ($r = 0.57$ and $r = 0.59$).[31,35]

Diagnostic accuracy

Accuracy in detecting limited HL was tested for seven instruments (Table 3). The REALM-R and NVS demonstrated good accuracy using the S-TOFHLA as reference standard. AUC values for the REALM-R and the NVS were 0.80 (95% CI = 0.73–0.86) and 0.83 (95% CI = 0.78–0.87), respectively.[32] The REALM-R provides reasonable sensitivity and specificity for the ED setting with detecting 81% of patients with limited HL and correctly reporting 62% of patients with adequate HL. The NVS appears to be highly sensitive (98%), but less specific (46%).[32] The BHLS showed moderate diagnostic accuracy against various reference methods (i.e., the S-TOFHLA, the REALM-R, the WRAT-4, and the SAHL) with AUC values ranging between 0.62 (95% CI 0.59–0.64) and 0.77 (95% CI = 0.70–0.83).[31,32,37] Diagnostic accuracy of the three separate SILS questions, both in the English and in Spanish version, were poor when using the SAHL as the reference standard. AUC values ranged between 0.58 (95% CI = 0.56–0.61) and 0.63 (95% CI = 0.60–0.66).[37] Finally, the SNS-8 demonstrated moderate accuracy in detecting limited HL against the WRAT-4 with an AUC of 0.77 (95% CI = 0.70–0.82).[31]

Feasibility

Four instruments were evaluated on feasibility (Table 4). The REALM-R demonstrated the shortest mean administration time (1.06 min).[32] The NVS and the SAHL show similar ease of use with regard to time of administration (means range between 3.31 and 3.57 min).[32,36] Interruptions during administration were minimal for all three instruments (<6.1%), particularly for the REALM-R (0.5%). Compared to the other tested instruments, the S-TOFHLA had the longest administration time (mean = 6.55 min) and the highest percentage (13.1) of interruptions during test performance.[32]

DISCUSSION

To our knowledge, this is the first comprehensive review of HL measurement instruments tested in the ED. Although a substantial number of instruments have been recently developed and tested in various health care settings and across different populations,[18–20,24] the evaluation of such instruments in the ED setting remains scarce. Only seven studies, all performed in the United States, fulfilled our inclusion criteria. Nevertheless, our findings provide a valuable overview that could help ED professionals in selecting the most appropriate HL measurement instrument for clinical and scientific purposes.[1]

With regard to psychometrics, the following conclusions can be drawn. First, the BHLS and the SNS (short and extended) have good evidence for reliability in measuring HL in the ED as they demonstrated acceptable and good internal consistency across studies. The internal reliability of a HL measure may be high in other nonacute health care settings, but this outcome could be different when measuring patient's HL level under different circumstances like in the ED. Second, none of the studied HL instruments performed consistently well on criterion validity. Most instruments demonstrated poor or moderate validity against different reference standards. The question may arise whether the true HL status of ED patients can be captured well enough by a relatively simple screening tool often measuring individuals' self-perceived HL abilities that are sensitive for bias.[19,39] Another explanation may be found in inappropriate validation criteria used by the studies, because selected reference methods may have measured a different domain of the multidimensional concept of HL than the studied HL instrument itself (e.g., reading ability and word comprehension versus numeracy skills).[19,40] Third, a limited number of instruments were tested on accuracy in detecting limited HL among ED visitors with mixed results. The REALM-R and the NVS showed good diagnostic accuracy (especially high sensitivity) against one reference standard. However, these performance outcomes

TABLE 3 Characteristics and psychometric properties of studied HL measurement instruments, categorized by mode of measurement

Instrument	Constructs measured	Items	Cronbach's alpha	Correlation coefficient			Diagnostic accuracy for detecting LHL		
				With S-TOFHLA	With REALM-R	With WRAT-4	AUC (95% CI)	SE (95% CI)	SP (95% CI)
Objective measurement approach (n = 5)									
S-TOFHLA	Close-type comprehension	36	NR	NR	0.56 [34]	NR	NR	NR	NR
REALM-R	Recognition and pronunciation of medical words	8	NR	0.54 [32]; 0.56 [34]	NR	NR	0.80 (0.73–0.86)[g] [32]	80.8 (73.2–88.3)[g] [32]	61.7 (56.5–67.0)[g] [32]
NVS	Reading and comprehension of nutrition label	6	NR	0.60 [32]; 0.62 [34]; 0.45 [33]; 0.32[a] [33]; 0.47[b] [33]	0.57 [34]	NR	0.83 (0.78–0.87)[g] [32]	98.0 (93.1–99.8)[g] [32]	45.7 (40.3–51.3)[g] [32]
METER	Recognition of medical words	80	NR	0.53 [34]	0.73 [34]	NR	NR	NR	NR
SAHL	Reading and comprehension of medical words	18	NR	NR	NR	NR	NR	NR	NR
Subjective measurement approach (n = 6)									
SILS-help with reading	Screening question[c]	1	NA	0.42 [34]	0.38 [34]	NR	0.59 (0.56–0.62)[h] [37]; 0.58 (0.56–0.61)[i] [37]	42.7 (33.2–52.3)[g] [32]	85.5 (81.7–89.3)[g] [32]
SILS-confident with forms	Screening question[d]	1	NA	0.39 [34]	0.40 [34]	NR	0.62 (0.59–0.65)[h] [37]; 0.60 (0.57–0.63)[i] [37]	54.8 (45.2–64.4)[g] [32]	80.4 (76.1–84.6)[g] [32]
SILS-understanding information	Screening question[e]	1	NA	0.40 [34]	0.43 [34]	NR	0.63 (0.60–0.66)[h] [37]; 0.59 (0.56–0.62)[i] [37]	40.4 (31.0–49.8)[g] [32]	95.5 (92.2–97.7)[g] [32]
BHLS	Reading and comprehension of medical information	3	0.74 [31]; 0.78 [32]	0.33 [31]; 0.24–0.49 [32]; 0.41 [34]	0.26 [31]; 0.44 [34]	0.26[d] [31]	0.74 (0.62–0.87)[g] [31]; 0.72 (0.62–0.81)[j] [31]; 0.77 (0.70–0.83)[k] [31]; 0.77 (0.70–0.83)[g] [32]; 0.66 (0.63–0.70)[h] [37]; 0.62 (0.59–0.64)[i] [37]	68.0 (59.0–77.1)[g] [32]	75.5 (70.8–80.1)[g] [32]

TABLE 3 (Continued)

Instrument	Constructs measured	Items	Cronbach's alpha	Correlation coefficient			Diagnostic accuracy for detecting LHL		
				With S-TOFHLA	With REALM-R	With WRAT-4	AUC (95% CI)	SE (95% CI)	SP (95% CI)
SNS-8	Self-reported numeracy abilities and preferences	8	0.82 [31]; 0.83 [35]	0.36 [31]; 0.40 [35]	0.36 [31]; 0.37 [35]	0.57[f] [31]; 0.59 [35]	0.77 (0.70–0.82)[k] [31]	NR	NR
SNS-3	Self-reported numeracy abilities and preferences	3	0.78 [35]	0.38 [35]	0.38 [35]	0.59 [35]	NR	NR	NR

Abbreviations: AUC, area under the curve; BHLS, Brief Health Literacy Screen; HL, health literacy; METER, Medical Term Recognition Test; NA, not applicable; NVS, Newest Vital Sign; NR, not reported; SE, sensitivity; SAHL, Short Assessment of Health Literacy; SILS, Single Item Literacy Screener; SNS, Short Numeracy Scale; SP, specificity; S-TOHFLA, Short Test of Functional Health Literacy among Adults; REALM-R, Rapid Estimate of Adult Literacy in Medicine-Revised; WRAT, Wide Range Achievement Test.

[a]Group with higher educational attainment.

[b]Group with lower educational attainment.

[c]"How often do you have someone (like a family member, friend, hospital or clinic worker, a caregiver, or anyone else) help you read materials given to you by the hospital, clinic, or your health care provider?"

[d]"How confident are you in filling out medical forms by yourself?"

[e]"How often do you have problems learning about your medical condition or health because of difficulty reading and understanding written information given to you by the hospital, clinic, or your health care provider?"

[f]40-item subtest of the WRAT-4 to objectively measure mathematic skills.

Reference methods:

[g]S-TOFHLA.

[h]SAHL-English version.

[i]SAHL-Spanish version.

[j]REALM-R.

[k]WRAT-4 mathematical subtest.

TABLE 4 HL measurement instruments tested on feasibility

Instrument	First author (year)	Sample (*n*)	Mean time of administration (min)[a]	Time on test (min/s)[b]	% Interrupted	Mean time of interruptions (min/s)
S-TOFHLA	Carpenter (2014)[32]	434	6.55 min	6.07 min	13.1	3.77 min
NVS	Carpenter (2014)[32]	428	3.31 min	3.13 min	6.0	2.85 min
	McGuinness (2020)[36]	104	3.57 min[c]	NR	4.8	5.54 s
SAHL	McGuinness (2020)[36]	98	3.45 min[c]	NR	6.1	4.96 s
REALM-R	Carpenter (2014)[32]	433	1.06 min	1.06 min	0.5	1.50 min

Abbreviations: HL, health literacy; NVS, Newest Vital Sign; REALM-R, Rapid Estimate of Adult Literacy in Medicine-Revised; S-TOFHLA, Short Test of Functional Health Literacy in Adults; SAHL, Short Assessment of Health Literacy.

[a]Total time for the test.

[b]Total time – interrupted time.

[c]Originally reported in seconds and converted to minutes.

originate from a single study performed in one ED with a high risk of bias and, therefore, should be interpreted with caution. This also applies to the SNS-8 showing moderate accuracy. The BHLS showed moderate accuracy in detecting limited HL and better performance compared to each of the poorly performing SILS questions. Evidence for the diagnostic performance of the BHLS is strengthened by the fact that accuracy was tested against different reference standards across multiple studies including > 3,000 adult participants. One of these studies was a multicenter study with low risk of bias.

Apart from psychometric properties, previous reviews dealing with HL measurement emphasized that the choice to use a particular HL instrument also depends on practical considerations related to administering the instrument and the advantages and disadvantages of using an objective versus subjective measurement mode.[18,19] Although the BHLS and the SNS-8 showed moderate accuracy in detecting limited HL and were not tested on feasibility in the ED, both measures seem to have several practical benefits. First, they appear to incur minimal administration time, which is important for ED professionals with limited available time for screening HL levels. Second, both measurements are based on self-reported answers that do not require in-person testing by trained staff using prepared materials.[23] Third, the self-perceived assessment of HL by both instruments involves less cognitive effort and a reduced risk for shame or stigma compared to objective tests like the REALM-R and the NVS.[19,41,42] In contrast, the major benefit of the REALM-R and the NVS is their direct and objective measurement of the individuals' skill based on empirically grounded data.[19] Unlike the BHLS and the SNS-8, these objective tests are not prone to bias associated with self-reports (e.g., patients overestimating their abilities due to perceived social desirability).[31] Moreover, both the REALM-R and the NVS demonstrated short administration times in the ED that correspond with previous evaluations of the instruments in other health care settings.[23] As previous reviews on HL measurement instruments already argued for the general population,[19] the choice for a particular instrument in the ED finally depends on the specific HL domain(s) that one wants to measure and respond to in the ED.[19] The BHLS, the REALM-R, and the NVS share a common focus on measuring individuals' reading ability and comprehension of health information while the SNS-8 specifically assesses numeracy skills and preferences.

This review may guide clinicians and policy-makers in their efforts to identify patients with limited HL as a primary step to improve patient–provider communication in the ED and ultimately health outcomes for this vulnerable population. Once limited HL is identified, tailored strategies can be used to improve information comprehension and to facilitate shared decision making in the ED as highlighted in the proceedings of the 2016 Academic Emergency Medicine Consensus Conference on Shared Decision Making in the Emergency Department.[43] Although strong evidence for one or more accurate and feasible HL instruments remains limited and findings only apply for English- and Spanish-speaking populations, our systematic review informs ED clinicians and policy-makers by presenting currently available instruments, along with their advantages and disadvantages, that can be used for the assessment of HL in daily practice. They should take these above-mentioned considerations into account when selecting a measurement instrument as a mean to overcome the barriers to health care delivery and the negative outcomes associated with limited HL. Various interventions, albeit with limited evidence base in the ED setting, are available with the potential to overcome such barriers throughout a patient's ED visit. Interventions at the clinical–patient level include clear communication (e.g., slow down, use of plain language, lowering the level of detail, presenting essential information by itself or first), confirmation of understanding (e.g., teach-back method), and reinforcement (e.g., combining verbal information with illustrations).[15,44–47] Interventions at the system–patient level include clear educational materials at the appropriate literacy level, visual aids, clear medication labeling, and shame-free clinical environments.[15,44,46]

LIMITATIONS

Our systematic review has several limitations. First, the heterogeneity of studied instruments measuring different domains of the multidimensional concept of HL and the variety of reference standards used make it difficult to compare instruments' validity and

diagnostic performance in the ED. Second, the evidence provided on criterion validity and diagnostic performance is based on a limited number of studies with poor methodologic quality and should therefore be interpreted with caution. Third, all included studies were performed in the United States and involved instruments that were mainly tested in a single urban ED among English-speaking patients. These aspects may limit the external validity of instruments and the extrapolation of estimates of validity and diagnostic accuracy to ED settings elsewhere serving populations with dissimilar sociodemographic characteristics (e.g., non–English speakers). Moreover, in most of the included studies patients were not formally screened on potentially confounding characteristics (e.g., dementia or undue distress). Therefore, mild cognitive dysfunctions may have been undetected and influenced the findings.[48] Finally, as with any systematic review, selection bias is possible. Although we conducted an extensive search of electronic literature, the search was limited by peer-reviewed full-text publications with an abstract in English language only.

CONCLUSION

This review highlights the existence of several short and simple instruments that appear valid in measuring health literacy levels and accurate in detecting limited health literacy among ED patients. These instruments differ in the mode of measurement, each with its practical advantages and disadvantages, and in the measurement of health literacy domains. Unfortunately, the low number of included studies and their methodologic limitations hinder the demonstration of robust evidence supporting one or more instruments. In the context of the widespread problem of limited health literacy and the ED as a first point of contact for many patients where they receive important health information, our findings call for more research to develop a robust evidence base for rapid and easy-to-administer health literacy measurement instruments that are psychometrically and diagnostically sound regardless of language spoken.[37] Future research may benefit from the following considerations. First, a better alignment of instruments with definitions of health literacy and tested against a corresponding criterion standard would facilitate a more meaningful comparison of instruments. Second, future testing of instruments following the Standards for Reporting Diagnostic Accuracy (STARD)[49] and in accordance with the QUADAS-2 criteria would improve determining instrument's diagnostic accuracy with limited risk of bias. Third, the use of larger study samples including non-English speakers across multiple and different types of EDs, also outside the United States, could improve the instruments' external validity. Findings of this systematic review should encourage and guide health care professionals and scientists further in their efforts to detect ED patients with limited health literacy and to facilitate their involvement in and receipt of optimal health care.

CONFLICT OF INTEREST

The authors have no potential conflicts to disclose.

AUTHOR CONTRIBUTIONS

Gijs Hesselink and Yvonne Schoon conceived and designed the study. Gijs Hesselink and Joey Cheng were responsible for data collection. Gijs Hesselink and Joey Cheng analyzed and interpreted the data. Gijs Hesselink and Joey Cheng drafted the paper, which was critically revised for important intellectual content by Yvonne Schoon. Gijs Hesselink takes responsibility for the paper as a whole.

ORCID

Gijs Hesselink https://orcid.org/0000-0003-2532-0724

REFERENCES

1. Institute of Medicine (US) Committee on Health Literacy; Nielsen-Bohlman L, Panzer AM, Kindig DA, editors. *Health Literacy: A Prescription to End Confusion*. National Academies Press; 2004.
2. Berkman ND, Davis TC, McCormack L. Health literacy: what is it? *J Health Commun*. 2010;15(Suppl 2):9-19.
3. Peerson A, Saunders M. Health literacy revisited: what do we mean and why does it matter? *Health Promot Int* 2009;24:285-296.
4. Madeeha M, Rubab ZZ, Azhar H. Health literacy as a global public health concern: a systematic review. *J Pharmacol Clin Res*. 2017;4:555632.
5. Kutner M, Greenberg E, Jin Y, Boyle B, Hsu YC, Dunleavy E. *Literacy in everyday life: Results from the 2003 National Assessment of Adult Literacy. U.S. Department of Education*. National Center for Education Statistics; 2007.
6. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *Eur J Public Health*. 2015;25:1053-1058.
7. Australian Bureau of Statistics. *Health Literacy, Australia*. Australian Bureau of Statistics; 2006.
8. Smith SG, Curtis LM, O'Conor R, Federman AD, Wolf MS. ABCs or 123s? The independent contributions of literacy and numeracy skills on health task performance among older adults. *Patient Educ Couns*. 2015;98:991-997.
9. Chen AM, Yehle KS, Albert NM, et al. Relationships between health literacy and heart failure knowledge, self-efficacy, and self-care adherence. *Res Social Adm Pharm*. 2014;10:378-386.
10. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. *Ann Intern Med*. 2011;155:97-107.
11. Herndon JB, Chaney M, Carden D. Health literacy and emergency department outcomes: a systematic review. *Ann Emerg Med*. 2011;57:334-345.
12. Griffey RT, Kennedy SK, McGownan L, et al. Is low health literacy associated with increased emergency department utilization and recidivism? *Acad Emerg Med*. 2014;21:1109-1115.
13. McNaughton CD, Kripalani S, Cawthon C, Mion LC, Wallston KA, Roumie CL. Association of health literacy with elevated blood pressure: a cohort study of hospitalized patients. *Med Care*. 2014;52:346-353.
14. McNaughton CD, Cawthon C, Kripalani S, et al. Health literacy and mortality: a cohort study of patients hospitalized for acute heart failure. *J Am Heart Assoc*. 2015;4:e001799.
15. Sudore RL, Schillinger D. Interventions to improve care for patients with limited health literacy. *J Clin Outcomes Manag*. 2009;16:20-29.
16. Shaw A, Ibrahim S, Reid F, Ussher M, Rowlands G. Patients' perspectives of the doctor-patient relationship and information giving across a range of literacy levels. *Patient Educ Couns*. 2009;75:114-120.
17. Smith SK, Dixon A, Trevena L, Nutbeam D, McCaffery KJ. Exploring patient involvement in healthcare decision making across different education and functional health literacy groups. *Soc Sci Med*. 2009;69:1805-1812.

18. Altin SV, Finke I, Kautz-Freimuth S, Stock S. The evolution of health literacy assessment tools: a systematic review. *BMC Public Health*. 2014;24(14):1207.
19. Nguyen TH, Paasche-Orlow MK, McCormack LA. The state of the science of health literacy measurement. *Stud Health Technol Inform*. 2017;240:17-33.
20. Liu H, Zeng H, Shen Y, et al. Assessment tools for health literacy among the general population: a systematic review. *Int J Environ Res Public Health*. 2018;15:1711.
21. Alqudah M, Johnson M, Cowin L, George A. Measuring health literacy in emergency departments. *J Nurs Educ*. 2014;4:1-10.
22. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol*. 2009;62:e1-34.
23. Collins SA, Currie LM, Bakken S, Vawdrey DK, Stone PW. Health literacy screening instruments for eHealth applications: a systematic review. *J Biomed Inform*. 2012;45:598-607.
24. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529-536.
25. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21:651-657.
26. Leeflang MM. Systematic reviews and meta-analyses of diagnostic test accuracy. *Clin Microbiol Infect*. 2014;20:105-113.
27. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58:882-893.
28. George D, Mallery P. *SPSS for Windows Step by Step: A Simple Guide and Reference*. 4th ed, 11.0 update. Allyn & Bacon; 2003.
29. Hinkle DE, Wiersma W, Jurs SG. *Applied Statistics for the Behavioral Sciences*. 5th ed. Houghton Mifflin; 2003.
30. van Bokhorst–de van der Schueren MA, Guaitoli PR, Jansma EP, de Vet HC. A systematic review of malnutrition screening tools for the nursing home setting. *J Am Med Dir Assoc*. 2014;15:171-184.
31. McNaughton C, Wallston KA, Rothman RL, Marcovitz DE, Storrow AB. Short, subjective measures of numeracy and general health literacy in an adult emergency department. *Acad Emerg Med*. 2011;18:1148-1155.
32. Carpenter CR, Kaphingst KA, Goodman MS, Lin MJ, Melson AT, Griffey RT. Feasibility and diagnostic accuracy of brief health literacy and numeracy screening instruments in an urban emergency department. *Acad Emerg Med*. 2014;21:137-146.
33. Morrison AK, Schapira MM, Hoffmann RG, Brousseau DC. Measuring health literacy in caregivers of children: a comparison of the Newest Vital Sign and S-TOFHLA. *Clin Pediatr (Phila)*. 2014;53:1264-1270.
34. Kiechle ES, Hnat AT, Norman KE, Viera AJ, DeWalt DA, Brice JH. Comparison of brief health literacy screens in the emergency department. *J Health Commun*. 2015;20:539-545.
35. McNaughton CD, Cavanaugh KL, Kripalani S, Rothman RL, Wallston KA. Validation of a short, 3-item version of the subjective numeracy scale. *Med Decis Making*. 2015;35:932-936.
36. McGuinness M, Bucher J, Karz J, et al. Feasibility of health literacy tools for older patients in the emergency department. *West J Emerg Med*. 2020;21:1270-1274.
37. Merchant RC, Marks SJ, Clark MA, Carey MP, Lui T. Limited ability of three health literacy screening items to identify adult English- and Spanish-speaking emergency department patients with lower health literacy. *Ann Emerg Med*. 2020;75:691-703.
38. Lee SY, Stucky BD, Lee JY, Rozier RG, Bender DE. Short Assessment of Health Literacy-Spanish and English: a comparable test of health literacy for Spanish and English speakers. *Health Serv Res*. 2010;45:1105-1120.
39. Lee SY, Tsai TI, Tsai YW. Accuracy in self-reported health literacy screening: a difference between men and women in Taiwan. *BMJ Open*. 2013;3:e002928.
40. Rademakers J, Waverijn G, Rijken M, Osborne R, Heijmans M. Towards a comprehensive, person-centred assessment of health literacy: translation, cultural adaptation and psychometric test of the Dutch Health Literacy Questionnaire. *BMC Public Health*. 2020;20:1850.
41. Easton P, Entwistle VA, Williams B. How the stigma of low literacy can impair patient-professional spoken interactions and affect health: insights from a qualitative investigation. *BMC Health Serv Res*. 2013;13:319.
42. VanGeest JB, Welch VL, Weiner SJ. Patients' perceptions of screening for health literacy: reactions to the Newest Vital Sign. *J Health Commun*. 2010;15:402-412.
43. Griffey RT, McNaughton CD, McCarthy DM, et al. Shared decision making in the emergency department among patients with limited health literacy: beyond slower and louder. *Acad Emerg Med*. 2016;23:1403-1409.
44. Berkman ND, Sheridan SL, Donahue KE, et al. Health literacy interventions and outcomes: an updated systematic review. *Evid Rep Technol Assess (Full Rep)*. 2011;1-941.
45. Griffey RT, Shin N, Jones S, et al. The impact of teach-back on comprehension of discharge instructions and satisfaction among emergency patients with limited health literacy: a randomized, controlled study. *J Commun Healthc*. 2015;8:10-21.
46. Visscher BB, Steunenberg B, Heijmans M, et al. Evidence on the effectiveness of health literacy interventions in the EU: a systematic review. *BMC Public Health*. 2018;18:1414.
47. Sheridan SL, Halpern DJ, Viera AJ, Berkman ND, Donahue KE, Crotty K. Interventions for individuals with low health literacy: a systematic review. *J Health Commun*. 2011;16(Suppl 3):30-54.
48. Kaphingst KA, Goodman MS, MacMillan WD, Carpenter CR, Griffey RT. Effect of cognitive dysfunction on the relationship between age and health literacy. *Patient Educ Couns*. 2014;95:218-225.
49. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med*. 2003;138:W1-12.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Hesselink G, Cheng J, Schoon Y. A systematic review of instruments to measure health literacy of patients in emergency departments. *Acad Emerg Med*. 2022;29:890–901. doi:10.1111/acem.14428