Journal of Health Communication, 0:1–8, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1081-0730 print/1087-0415 online
DOI: 10.1080/10810730.2014.917744

Routledge
Taylor & Francis Group

Development and Validation of the Comprehensive Health Activities Scale: A New Approach to Health Literacy Measurement

LAURA M. CURTIS[1], WILLIAM REVELLE[2], KATHERINE WAITE[1], ELIZABETH A. H. WILSON[1], DAVID M. CONDON[2], ELIZABETH BOJARSKI[1], DENISE C. PARK[3], DAVID W. BAKER[1], and MICHAEL S. WOLF[1]

[1]*Health Literacy and Learning Program, Division of General Internal Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA*
[2]*Department of Psychology, Northwestern University, Evanston, Illinois, USA*
[3]*Center for Vital Longevity, University of Texas at Dallas, Dallas, Texas, USA*

Current health literacy measures have been criticized for solely measuring reading and numeracy skills when a broader set of skills is necessary for making informed health decisions, especially when information is often conveyed verbally and through multimedia video. The authors devised 9 health tasks and a corresponding 190-item assessment to more comprehensively measure health literacy skills. A sample of 826 participants between the ages of 55 and 74 years who were recruited from an academic general internal medicine practice and three federally qualified health centers in Chicago, Illinois, completed the assessment. Items were reduced using hierarchical factor analysis and item response theory resulting in the 45-item Comprehensive Health Activities Scale. All 45 items loaded on 1 general latent trait, and the resulting scale demonstrated high reliability and strong construct validity using measures of health literacy and global cognitive functioning. The predictive validity of the Comprehensive Health Activities Scale using self-reported general, physical, and mental health status was comparable to or better than widely used measures of health literacy, depending on the outcome. Despite comprehensively measuring health literacy skills, items in the Comprehensive Health Activities Scale supported 1 primary construct. With similar psychometric properties, current measures may be adequate, depending on the purpose of the assessment.

The field of health literacy has developed considerably over the past two decades, with numerous research studies linking crude measures of reading ability and numeracy with a variety of health outcomes. A continued criticism of current health literacy measures is the notion that health literacy refers solely to reading and numeracy skills when, in fact, the conceptual framework of health literacy reflects a far broader set of skills (Baker, 2006; Jordan, Osborne, & Buchbinder, 2011; Nielson-Bohlman, Panzer, & Kindig, 2004). In addition, stimuli that are frequently presented to individuals for existing tests rely heavily on measuring one's ability to process health information using a print or text-based medium. While patients surely must be able to read information in print form and perform basic calculations to effectively engage in the health care system, they must also retain information in support of self-care, listen and process spoken communication, navigate new technologies, and problem-solve in new or changing situations.

Despite these criticisms of the most common tests used in health literacy investigations, their predictive validity is undeniable (DeWalt, Berkman, Sheridan, Lohr, & Pignone, 2004; Institute of Medicine, 2004). Regardless, new health literacy tests continue to emerge; some improve upon the delivery of the assessment (Hahn et al., 2004), shorten the administration time (Arozullah et al., 2007; Lee, Bender, Ruiz, & Cho, 2006), or expand the conceptual understanding of health literacy (McCormack et al., 2010; Osborn, Davis, Bailey, & Wolf, 2010). While these measures hold merit and over time may gather evidence equivalent to commonly used assessments (i.e., Test of Functional Health Literacy in Adults [TOFHLA], Rapid Estimate of Adult Literacy in Medicine [REALM], Newest Vital Sign [NVS]) to support their predictive validity, ultimately they all depend on document-based text, charts, and graphs, either using pencil-and-paper tests or their electronic web-based versions. These new tools, along with current measures, may serve well as a proxy for identifying those who have difficulty with document or web-based tasks but may not be informative for more challenging or perhaps less familiar activities one

Address correspondence to Laura M. Curtis, Health Literacy and Learning Program, Division of General Internal Medicine, Feinberg School of Medicine, Northwestern University, 750 N. Lake Shore Drive, 10th Floor, Chicago, IL 60611, USA. E-mail: l-curtis@northwestern.edu

may encounter in health care. In addition, print materials are often not available and patients, especially those with limited reading ability, often rely on recall of verbal instructions given in an encounter. A measure that more broadly reflects the vast array of current tasks faced by patients and families when accessing, understanding, and applying health information could better guide health systems in matching service demands to patient abilities.

We devised a broad range of cognitively challenging yet routine health care scenarios common for the aging population in an attempt to more comprehensively assess individual health literacy skills. Both information presentation format and the abilities necessary to obtain, process, understand, and apply this information were considered in the creation of this tool, herein referred to as the Comprehensive Health Activities Scale (CHAS). Information was presented in form of print documents, common artifacts (i.e., pill bottles), and verbal communication either replicated as spoken from a medical professional or communicated through a streaming health education video. In addition to addressing reading and numeracy abilities using print documents and pill bottles as in the commonly used assessments, the addition of spoken communication and multimedia video allowed us to also measure comprehension and recall of verbal information without written support documents, as is often necessary in medical encounters. As such, the variety of tasks was intended not only to more accurately reflect current patient roles in health care but also to also expand the range of difficulty of items.

Given that causal mechanisms linking health literacy to health outcomes currently remain unclear, the CHAS not only could more precisely identify requisite health skills but also could aid our thinking in how best to devise interventions that can simplify tasks and improve individuals' self-care performance. We review the development and psychometric validation of the CHAS, including construct validity mapped to the three most common health literacy measures and a widely used test of global cognitive function. Its predictive validity was examined by mapping CHAS scores to general health status and functional health status measures including physical health status, depression, and anxiety while comparing its performance to that of current literacy measures in health care.

Method

Sample

We recruited 832 participants between 55 and 74 years of age from one academic general internal medicine clinic and four federally qualified health centers between August 2008 and October 2010 as part of the Literacy and Cognitive Functioning study funded by the National Institute on Aging. Using electronic health records, we initially identified 3,176 patients as eligible to participate. We invited them by mail and by phone to participate in the study. Of the 3,176 patients, our research staff reached 1,904 individuals. Initial screening deemed 244 subjects as ineligible because they had severe cognitive or hearing impairment, they had limited English proficiency, or they were not connected to a clinic physician (defined as fewer than two visits in 2 years). In addition, 794 refused, 14 were deceased, and 20 were eligible but had scheduling conflicts. The final sample included 832 participants, for a determined cooperation rate of 51 percent following American Association for Public Opinion Research guidelines (American Association for Public Opinion Research, 2004).

Procedure

Subjects completed two structured interview sessions with trained research assistants, 7–10 days apart, each lasting approximately 2.5 hours. These analyses include assessments from the first interview including basic demographic information, socioeconomic status, comorbidity, functional health status, three measures of health literacy, and a comprehensive assessment of performance on everyday health tasks. Northwestern University's Institutional Review Board approved the study and patients provided written informed consent before participation.

Instrument Development

Nine different scenarios depicting health-related tasks common for the aging population were developed by members of our research team, including experts in the fields of health literacy, test development, medicine, and cognitive psychology. Most were created before this study and have been piloted and tested (Davis, Wolf, Bass, Middlebrooks, et al., 2006; Davis, Wolf, Bass, Thompson, et al., 2006; Hahn, 2009; Heinemann, Deutsch, Mallinson, & Gershon, 2006–2009; Weiss et al., 2005; Wilson et al., 2010; Wolf, Baker, & Makoul, 2007). The development of the verbal tasks, which were created for this study, is described in more detail elsewhere (McCarthy et al., 2012).

Information was presented to participants using print documents, prescription medication bottles, spoken health communications, and multimedia video, as described in Table 1. After each scenario, participants answered a series of questions that asked them to demonstrate comprehension and use of the material and artifacts. Trained interviewers rated participants' responses as correct or incorrect. Verbatim answers were also recorded, and they were reviewed by the team if the interviewer was in doubt about a particular response. On average, all scenarios in the CHAS take roughly 60 minutes to administer. Tasks in each of these scenarios included simple retrieval of print information, recall of verbal and multimedia information, and demonstration of understanding information from pill bottles; more complicated tasks required calculation, multistep commands, or reasoning. This method was adapted from similar prior studies by the research team, methods from cognitive psychology, and national literacy and health literacy assessments such as the National Adult Literacy Study, the National Assessment of Adult Literacy, and the Health Activities Literacy Scale (Kirsch, 1993; Kutner, Greenberg, Jin, & Paulsen, 2006; Rudd, Kirsch, & Yamamoto, 2004; Willis, Jay, Diehl, & Marsiske, 1992).

Table 1. Description of health scenarios

Information presentation: Task	Description
Print documents	
Consent to a procedure	Read a consent form given before an angiography and exhibit understanding of the procedure, potential complications, and physician responsibilities
Monitor blood sugar	Calculate and interpret numeric information from a chart listing 7 days of recorded blood sugar levels before and after meals for a diabetic patient
Prepare for a procedure	Review instructions for colonoscopy preparation and demonstrate comprehension of patient duties before the procedure
Choose a facility	Examine written text about pressure sore prevention, a chart comparing prevention at two nursing homes, and a map in order to select the best facility
Monitor condition	Review and demonstrate understanding of a sheet about heart failure symptoms, monitoring, and self-care activities, as well as a calendar indicating weight and swelling status
Medication bottles	
Manage prescription medications	Review prescription bottles from two hypothetical prescription medication regimens; pronounce the names of the medications, make inferences on usage, and dose both regimens over a 24-hr period using a medication box
Spoken instructions	
Understand new diagnosis	Receive oral instructions from a physician regarding a diagnosis and course of treatment for gastroesophageal reflux; answer questions to assess immediate and delayed recall about self-management
Recall home care instructions	Listen to wound care instructions for a laceration upon discharge from the emergency department; recall information about follow-up appointments, frequency of dressing change, and signs of infection
Multimedia video	
Recall symptom prevention information	Watch a video clip on identifying, monitoring, and controlling asthma triggers; recall information immediately following the video and at the end of interview

Measurement

Health Literacy

We used the TOFHLA, REALM, and NVS to assess health literacy (Davis et al., 1993; Parker, Baker, Williams, & Nurss, 1995; Weiss et al., 2005). The TOFHLA uses actual materials that patients might encounter in health care to test their reading fluency. It takes approximately 20 minutes to complete, and it consists of two parts: a 50-item reading comprehension section using the Cloze procedure and a 17-item numeracy assessment. Total scores are weighted for possible scores ranging from 0 to 100 and can be interpreted as inadequate (0–59), marginal (60–74), or adequate (75–100) literacy. The REALM is a word-recognition test in which patients are asked to read aloud as many words as they can from a list of 66 health-related terms arranged in order of increasing difficulty. Scores are based on the total number of words pronounced correctly, with dictionary pronunciation being the scoring standard and interpreted as low (0–44), marginal (45–60), or adequate (61–66) literacy. Last, the NVS is a screening tool used to determine risk for limited health literacy. Patients are given a copy of a nutrition label and asked six questions about how they would interpret and act on the information. Scores are classified as high likelihood of limited literacy (0–1), possibility of limited literacy (2–3), and adequate literacy (4–6).

Global Cognitive Functioning

We used the Mini Mental Status Exam (MMSE) to measure global cognitive functioning (Cockrell & Folstein, 1988). Total scores range from 0 to 30 and can be classified into severe cognitive impairment (0–17), mild cognitive impairment (18–23), and no cognitive impairment (24–30).

Functional Health Status

Self-reported health status was determined by asking participants to rate their overall health as *excellent, very good, good, fair*, or *poor*. Self-reported health was then dichotomized as *excellent, very good*, or *good* versus *fair* or *poor*. Physical health status was assessed using the SF-36 physical function short-form from the Patient-Reported Outcomes Measurement Information System (PROMIS; Hays, Bjorner, Revicki, Spritzer, & Cella, 2009; Teresi et al., 2009), with scores ranging from 0 to 100 (Ware, Kosinski, & Keller, 1994). Depression and anxiety were measured using the PROMIS emotional health short-form measures. Scores on the depression scale range from 8 to 40, and scores on the anxiety scale range from 7 to 35.

Analyses

Item Response Variability

Across the nine health scenarios, 190 items were initially scored as correct or incorrect. Frequencies of correct responses were calculated and tetrachoric correlations were computed to examine the pattern of correlations between the dichotomous items. If items are too highly correlated, the estimate of reliability is inflated and assumptions of item response theory are violated (Cattell, 1978; Reeve & Fayers, 2005). Items with fewer than 5% or greater than 95% correct, or that were highly correlated with other items (>0.80) were examined and those found to be too easy, difficult, or redundant with other items were either eliminated or rescored by combining them with similar items. In addition, when participants were asked to recall a list of items, we examined different ways of scoring (either by giving a point for each item listed or based on the total number recalled from the list, such as 3/5 or 4/5).

Factor Structure

We hypothesized a general nonspecific factor would account for substantial variance in all items with the possibility of lower order factors. We examined the factor structure of the items by determining the number of components that minimized the average squared partial correlation using Velicer's Minimum Average Partial criteria and by using the Very Simple Structure algorithm, which compares the fit of the simplified factor matrix to one that uses only the largest factor loadings (Revelle & Rocklin, 1979; Velicer, 1976). On the basis of these results, we then ran omega analyses, a method of hierarchical factor analysis that uses the Schmid-Leiman transformation to first estimate a general factor followed by evaluation of the lower level factor loadings after accounting for variation by the general factor (McDonald, 1999; Revelle & Zinbarg, 2009; Schmid & Leiman, 1957). We compared the fit of these models using the root mean squared error of approximation (RMSEA) and the root mean square residual (RMSR). A model with low RMSEA (<0.06) and RMSR (<0.08) is considered to adequately fit the data (Hu & Bentler, 1999).

Items with factor loadings greater than 0.3 on the common latent variable were retained (Nunnally & Bernstein, 1984; Tabachnick & Fidell, 2001). A second omega analysis was run to determine whether the remaining items measure a latent variable in common and the extent to which this latent variable accounts for the variance in the scale scores (Cronbach, 1951; McDonald, 1999; Revelle & Rocklin, 1979; Revelle & Zinbarg, 2009). Omega hierarchical (ω_H) estimates the degree to which the test measures a single primary latent variable, omega total (ω_T) estimates the reliable variance in a test, and Cronbach's α measures the internal reliability of the test (Revelle & Zinbarg, 2009; Zinbarg, Revelle, Yovel, & Li, 2005). We conducted diagnostic statistics to examine the quality of the omega solution including comparison of the relative size of the g factor eigenvalue to the other eigenvalues.

Item Response Theory

We used item response theory to further assess the psychometric properties of the remaining items. Item response theory allows one to evaluate the ability of a test to discriminate between respondents with different levels of an underlying trait. The probability of a person with a given level of ability answering an item correctly was modeled using a two-parameter logistic model. The discrimination parameter (α) represents the strength of the relation between a particular item and an underlying trait and indicates how effectively an item can differentiate between participants at different levels of the latent trait. It typically ranges from 0.5 to 2.5, with more effective items having larger values (Reeve & Fayers, 2005). The information for an item is maximized at the point at which 50% of the respondents answer the item correctly and is the best estimate for an item's difficulty (β). Items tend to range between -3 (*extremely easy*) and 3 (*extremely difficult*). Although calculated using factor analysis of the tetrachoric correlations, the item response theory and factor analysis parameters are just transforms of each other (Kamata & Bauer, 2008; McDonald, 1999).

When forming a test and evaluating the items within a test, the most useful items are the ones that give the most information about a person's score. Because the test information is just the sum of the item information, one can design a measure that provides maximum information at a particular point by including items of a specific difficulty, or to have relatively uniform information across a range of abilities by including items across all difficulties. We used the test information curve—a graphical representation of the information function for all the items in the test across the range of difficulty—to display the information of the test as a whole. Information for each item at each level of difficulty was provided in the appendix to allow one to determine the amount of information and from that the reliability (1–1/information) for a subset of items. Psychometric analyses were done using omega and irt.fa functions from the psych package version 1.1-12 (Revelle, 2011) in the open source data analysis and statistical system R version 2.14.1 (R Development Core Team, 2011).

Scoring

An overall score was created based on the items that loaded on the general factor. Scores were calculated as the percent correct out of the total possible multiplied by 100 and were not calculated for participants missing more than half of the items on a particular scale.

Construct and Predictive Validity Testing

Construct validity was assessed by examining correlations between the CHAS score and patients' scores on the TOFHLA, REALM, NVS, and the MMSE. For predictive validity, the ability of our measure to predict physical health and blood pressure control was tested using linear and logistic regression, respectively, controlling for age, sex, race, marital status, and total number of chronic conditions. Similar models were run for the literacy measures and the MMSE in order to compare the predictive validity of the CHAS to that of these four measures. For the continuous physical health, anxiety, and depression outcomes, standardized betas were calculated in order to compare the strength of association between the CHAS, health literacy, and MMSE estimates despite differing units of measurement. In addition, the amount of variance explained by the CHAS score was compared to the amount explained by the literacy measures and the MMSE using the Vuong Test, a likelihood ratio–based approach for nonnested models (Vuong, 1989). For the dichotomous outcome of self-reported health status, models were similarly assessed by comparing the equality of the area under the receiver operating characteristic curves for each model. Scoring and validity testing of the CHAS was done in STATA version 12.1.

Results

Sample

The sample of 826 participants included in these analyses is described in Table 2. Two thirds of participants (68%) were women, and the majority either Black (43%) or non-Hispanic White (50%). Participants were socially and economically diverse in years of schooling, household income, employment, marital status, and living situation. Individuals on average had two comorbid chronic conditions ($M = 1.9$, $SD = 1.4$) and were taking a mean of 3.6 prescription medications ($SD = 3.1$).

Item Response Variability

The percentage correct for the initial 190 items was generally high with an average of 65 ($SD = 23$). Many items were highly correlated with tetrachoric correlations ranging from 0.04 to 0.99. We examined the 22 items correctly answered by less than 5% or more than 95% of respondents or had correlations >0.80 with other items. These items were either eliminated or rescored by combining with similar items, and listed items were scored based on the total number recalled, leaving a total of 70 remaining items.

Table 2. Characteristics of sample ($N = 826$)

Variable	*M* (*SD*) or %
Age, *M* (*SD*)	63.1 (5.5)
Gender (%)	
Female	67.9
Race (%)	
Black	42.6
White	50.2
Other	7.2
Education (%)	
High school or less	27.1
Some college or technical school	22.0
College graduate	20.5
Graduate degree	30.4
Income (%)	
Less than $10,000	12.2
$10,000–$24,999	19.6
$25,000–$49,999	15.3
More than $50,000	52.9
Employment status (%)	
Full time	20.3
Part time	15.0
Not working	64.7
Chronic conditions (%)	
Hypertension	59.9
Diabetes	15.8
Coronary artery disease	6.9
Heart failure	5.0
Bronchitis or emphysema	13.2
Asthma	18.7
Arthritis	47.4
Cancer	7.4
Depression	20.5
Total number, *M* (*SD*)	1.9 (1.4)
Number of prescription medications, *M* (*SD*)	3.6 (3.1)

Factor Structure

Very Simple Structure and Velicer's Minimum Average Partial criterion indicated a strong first factor with the possibility of four additional factors in the remaining 70 items. Therefore, we ran an omega analysis extracting one general factor and four lower level factors. Items with factor loadings ≥ 0.3 on the common latent variable were retained, resulting in 92% reliable variance (ω_T) and 92% internal consistency (α) in the remaining 45 items. This common latent trait accounted for 79% of the variance in the items (ω_H). With this general factor partialed out, less than one third of the items ($n = 12$) also loaded on one of the four lower level factors (≥ 0.3). The eigenvalues for these factors were much smaller when compared with that for the common latent variable (1.0, 0.8, 0.9, 0.9 for factors 1 through 4, respectively, compared with 8.0) and although the fit is slightly better for the four-factor omega model, the difference is negligible (four-factor: RMSEA = 0.02, RMSR = 0.03; one-factor: RMSEA = 0.03, RMSR = 0.05). These results suggest a strong primary factor underlying responses in this sample with little influence from additional subfactors.

Item Response Theory

With support for a strong primary dimension and unidimensionality of the items, assumptions for item response theory analyses were satisfied (Reeve & Fayers, 2005). Overall test information is displayed in Figure 1, indicating the test is able to discriminate well across all ability levels but more so at middle to lower ability levels. Information for each of the 45 items is displayed in the Appendix. Item information in each column was summed in order to calculate test reliability (1–1/Information) for subsets of items, in this case

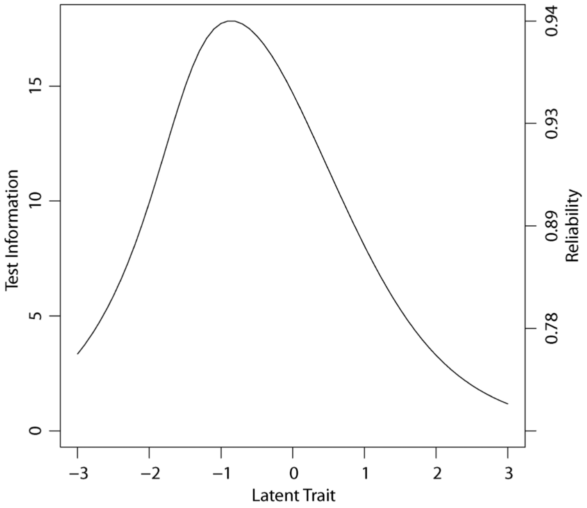

Fig. 1. Test information curve for the Comprehensive Health Activity Scale.

Table 3. Validity of the Comprehensive Health Activity Scale and current measures of health literacy and global cognitive function

	Construct validity	Predictive validity						
	CHAS	Self-reported health	Physical health		Depression		Anxiety	
	ρ	OR (95% CI)	β (95% CI)	Stand. β	β (95% CI)	Stand. β	β (95% CI)	Stand. β
CHAS	—	1.03 (1.02, 1.04)***	0.10 (0.04, 0.16)**	0.12	−0.06 (−0.09, −0.04)***	−0.22	−0.03 (−0.05, −0.01)**	−0.11
Health literacy								
TOFHLA	0.81***	1.03 (1.02, 1.04)***	0.09 (0.02, 0.17)**	0.08	−0.05 (−0.08, −0.03)***	−0.14	−0.02 (−0.05, 0.004)	−0.07
REALM	0.68***	1.03 (1.01, 1.01)***	0.10 (−0.001, 0.19)	0.06	−0.03 (−0.06, 0.01)	−0.05	0.01 (−0.02, −0.05)	0.02
NVS	0.75***	1.19 (1.06, 1.33)**	0.96 (0.36, 1.56)**	0.11	−0.51 (−0.73, −0.28)***	−0.17	−0.26 (−0.48, −0.04)*	−0.10
Global cognitive function								
MMSE	0.68***	1.14 (1.05, 1.25)**	0.63 (0.11, 1.16)*	0.08	−0.42 (−0.61, −0.22)***	−0.15	−0.19 (−0.38, 0.01)	−0.07

Note. CHAS = Comprehensive Health Activity Scale; OR = odds ratio; CI = confidence interval; β = unstandardized beta; Stand. β = standardized beta; TOFHLA = Test of Functional Health Literacy in Adults; REALM = Rapid Estimate of Adult Literacy in Medicine; NVS = Newest Vital Sign; MMSE = Mini Mental Status Examination.
***$p < .001$. **$p < .01$. *$p < .05$.

partitioned by information presentation modality. Using this table, one can similarly calculate information and reliability for any subset of items.

Construct and Predictive Validity

The CHAS score correlated highly with current measures of health literacy and the MMSE, being more closely related to the TOFHLA and the NVS ($r = 0.81$ and $r = 0.75$, respectively), compared with the REALM and the MMSE ($r = 0.68$ for both; Table 3). Regarding predictive validity, the CHAS assessment was significantly associated with self-reported health (OR = 1.03; 95% CI [1.02, 1.04]; $p < .001$), physical health ($\beta = 0.10$; 95% CI [0.04, 0.16]; $p = .002$), depression ($\beta = -0.06$; 95% CI [−0.09, −0.04]; $p < .001$), and anxiety ($\beta = -0.03$; 95% CI [−0.05, −0.01]; $p = .01$). The amount of variance in physical health and in anxiety explained by the CHAS was not significantly better or worse than that of current health literacy and global cognitive functioning measures (all ps > .20). However, the CHAS did explain significantly more variance in the depression outcome than the REALM (Vuong Z-statistic 2.67; $p < .01$). In addition, the CHAS similarly explained variance in self-reported health to the TOFHLA and the REALM but significantly more variance than the NVS (areas under the receiver operating characteristic curve [AUC] = 0.84 vs. 0.83; $p = .04$) and the MMSE (AUC = 0.84 vs. 0.82; $p = .02$).

Discussion

The meaning and measure of health literacy has been and continues to be a frequently discussed issue in the field as well as the subject of many ongoing studies. In the development of the CHAS, we intentionally designed the measure to tap into a far broader set of health skills than simply navigating text in prose selections or documents including tables and graphs and instead focused on realistic, complex health-based scenarios that patients often experience. In our study, the CHAS demonstrated high reliability, strong construct validity with the three most commonly used health literacy measures, and strong predictive validity with all self-reported general, physical, and mental health status measures. Although some items provided more information than others, all seem to get at the same latent trait of health literacy skills. Yet, we ask whether the CHAS, as a more dynamic assessment tool, should replace existing health literacy tests despite requiring more time and resources to administer. There are several important factors to consider in seeking out answers as to how to best assess the construct of health literacy.

We measured a variety of tasks including retrieval and comprehension of print information, recalling spoken and multimedia information, medication organization, and tasks requiring calculation and multistep commands, but all items mapped to a single latent factor, suggesting measurement of the same construct. In other words, one's ability to comprehend print information may be similar to his or her ability to retain information from spoken and multimedia sources.

While the REALM, TOFHLA, and NVS may be highly criticized as very crude measures of health literacy that focus on reading comprehension and/or numeracy rather than the full range of skills necessary to function in the health care environment, they too are likely capturing this latent trait. Despite these criticisms, the ability of the REALM, TOFHLA, and NVS to predict outcomes is unquestionable. While we expected our more comprehensive assessment to have considerably greater predictive validity, this was not the case; the CHAS performed comparably, if not only slightly better than current measures. Considering brevity when measuring health literacy, which is essential in research and clinical settings, these more crude assessments may be adequate as measures of a health literacy construct.

Existing health literacy tests offer validated thresholds for identifying at-risk individuals, which can be valuable in both a research and clinical setting. However, if the intent is to learn precisely how to revise an existing health care task or process, the CHAS items and/or its approach might be able to offer more assistance. By determining which items are more difficult at varying levels of ability, tasks or processes can be simplified in order to make them clearer for the targeted population. For example, the most difficult items in the CHAS were related to recalling information from a multimedia video. If clinics use videos to educate patients about their condition or how to perform tasks at home, providing tangible supports for participants to take with them may support recall of this information (Wilson et al., 2010). In addition, if researchers want to test an intervention designed to improve a specific task or skill set (e.g., accurately taking medication), a measure testing that particular skill would be beneficial.

In both research and clinical settings, there are requisite attributes that are important to consider for any standard assessment. As such, practical limitations of the CHAS include technology and stimuli requirements (i.e., pill bottles and dosing trays) as well as the time it takes to complete all nine tasks. However, this item-level analysis identifies the most informative items and tasks in the CHAS. These results, in turn, could be used to inform the development of adaptive procedures for CHAS administration. Because all items measure the same latent trait, it does not matter which tasks are chosen as long as the items provide information for the targeted individuals, allowing for practical flexibility in assessment administration. Using portions of the CHAS specific to what a patient might be expected to perform in a clinical context could appear more relevant and less as a test of ability, and thereby reduce stigma. For example, presenting patients with diabetes with the hypothetical blood sugar monitoring task would be a more intuitive choice rather than to give a patient a list of words to pronounce or to read a Medicaid application.

In addition, the tasks and scoring criteria of the CHAS will be publicly available for others to use and could potentially be administered online. The verbal and multimedia tasks are already presented to patients on a laptop computer; while paper documents were used for the majority of tasks, information could be presented in the same way on the screen as other new measures have demonstrated (Hahn et al., 2004; McCormack et al., 2010). Further research is needed to determine whether this manner of administration would affect the psychometric properties of this assessment.

While we believe the CHAS has applicability for individuals across all age groups, our sample was limited to older patients by its development being linked to the larger Literacy and Cognitive Functioning study. However, the stimuli provided reflects tasks that would be common for those with some level of comorbidity or familiarity with the health care system. Future investigations should evaluate the CHAS among a more diverse sample in terms of age and health care experience. In addition, more objective health outcomes including blood pressure control, hospitalizations, and emergency department visits from patients' medical records will be used to determine further predictive validity once data become available.

Conclusion

While the intent of this study was to more comprehensively measure aspects of health literacy that current measures might potentially be missing, such as understanding of oral instructions and multimedia learning, our evidence suggests that a more comprehensive measure may not be necessary depending on the purpose of the assessment. Yet, these findings are important in that we offer a framework of the degree of complexity for many common health activities, offering insight into the most pressing targets for simplifying health care. Whether or not any new tool will replace the current gold standards will largely be determined by the availability of evidence supporting their use, access, and feasibility. Even the most impressive psychometric properties cannot ensure widespread adoption if a health literacy assessment is not publicly available or cannot be easily administered in diverse clinic settings.

Funding

This project was supported by the National Institute on Aging grant R01 AG030611 (principal investigator: Wolf).

Supplemental Material

Supplemental data for this article (Appendix: Item response theory information, discrimination, and difficulty results by task) can be accessed on the publisher's website at http://dx.doi.org/10.1080/10810730.2014.917744.

References

American Association for Public Opinion Research. (2004). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Ann Arbor, MI: American Association for Public Opinion Research.

Arozullah, A. M., Yarnold, P. R., Bennett, C. L., Soltysik, R. C., Wolf, M. S., Ferreira, R. M., ... Davis, T. (2007). Development and validation of a short-form, rapid estimate of adult literacy in medicine. *Medical Care*, *45*, 1026–1033.

Baker, D. W. (2006). The meaning and the measure of health literacy. *Journal of General Internal Medicine*, *21*, 878–883.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York, NY: Plenum Press.

Cockrell, J. R., & Folstein, M. F. (1988). Mini-Mental State Examination (MMSE). *Psychopharmacology Bulletin*, *24*, 689–692.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Davis, T. C., Long, S. W., Jackson, R. H., Mayeaux, E. J., George, R. B., Murphy, P. W., . . . Crouch, M. A. (1993). Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Family Medicine*, *25*, 391.

Davis, T. C., Wolf, M. S., Bass, P. F., 3rd, Middlebrooks, M., Kennen, E., Baker, D. W., . . . Parker, R. M. (2006). Low literacy impairs comprehension of prescription drug warning labels. *Journal of General Internal Medicine*, *21*, 847–851.

Davis, T. C., Wolf, M. S., Bass, P. F., 3rd, Thompson, J. A., Tilson, H. H., Neuberger, M., . . . Parker, R. M. (2006). Literacy and misunderstanding prescription drug labels. *Annals of Internal Medicine*, *145*, 887–894.

DeWalt, D. A., Berkman, N. D., Sheridan, S., Lohr, K. N., & Pignone, M. P. (2004). Literacy and health outcomes: A systematic review of the literature. *Journal of General Internal Medicine*, *19*, 1228–1239.

Hahn, E. A. (2009, February). *Refining and standardizing Hhalth literacy assessment: English and Spanish item banks*. Presented at the Instiute of Medicine workshop on measures of health literacy, Washington, DC.

Hahn, E. A., Cella, D., Dobrez, D., Shiomoto, G., Marcus, E., Taylor, S. G., . . . Webster, K. (2004). The talking touchscreen: A new approach to outcomes assessment in low literacy. *Psycho-Oncology*, *13*, 86–95.

Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. *Quality of Life Research*, *18*, 873–880.

Heinemann, A., Deutsch, A., Mallinson, T., & Gershon, R. (2006–2009). Rehabilitation research and training center on measuring rehabilitation outcomes and effectiveness (H133B040032). Chicago, IL: National Institute for Disability and Rehabilitation Research.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Institute of Medicine. (2004). *Health literacy: A prescription to end confusion*. Washington, DC: National Academies Press.

Jordan, J. E., Osborne, R. H., & Buchbinder, R. (2011). Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses. *Journal of Clinical Epidemiology*, *64*, 366–379.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory. *Structural Equation Modeling-a Multidisciplinary Journal*, *15*, 136–153.

Kirsch, I. S. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Washington, DC: U.S. Government Printing Office.

Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy (NCES 2006-483)*. Washington, DC: National Center for Education Statistics.

Lee, S. Y., Bender, D. E., Ruiz, R. E., & Cho, Y. I. (2006). Development of an easy-to-use Spanish Health Literacy test. *Health Services Research*, *41*(4 Pt. 1), 1392–1412.

McCarthy, D. M., Waite, K. R., Curtis, L. M., Engel, K. G., Baker, D. W., & Wolf, M. S. (2012). What did the doctor say? Health literacy and recall of medical instructions. *Medical Care*, *50*, 277–282.

McCormack, L., Bann, C., Squiers, L., Berkman, N. D., Squire, C., Schillinger, D., . . . Hibbard, J. (2010). Measuring health literacy: A pilot study of a new skills-based instrument. *Journal of Health Communication*, *15*(Suppl 2), 51–71.

McDonald, R. P. (1999). *Test theory: A unified treatmen*. Mahwah, NJ: Erlbaum.

Nielson-Bohlman, L., Panzer, A., & Kindig, D. (Eds.). (2004). *Health literacy: A prescription to end confusion*. Washington, DC: Institute of Medicine, National Academies Press.

Nunnally, J. C., & Bernstein, I. H. (1984). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Osborn, C. Y., Davis, T. C., Bailey, S. C., & Wolf, M. S. (2010). Health literacy in the context of HIV treatment: Introducing the Brief Estimate of Health Knowledge and Action (BEHKA)-HIV version. *AIDS Behavior*, *14*, 181–188.

Parker, R. M., Baker, D. W., Williams, M. V., & Nurss, J. R. (1995). The test of functional health literacy in adults: A new instrument for measuring patients' literacy skills. *Journal of General Internal Medicine*, *10*, 537–541.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org

Reeve, B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (2nd ed., pp. 55–74). Oxford, England: Oxford University Press.

Revelle, W. (2011). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*, 403–414.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.

Rudd, R. E., Kirsch, I., & Yamamoto, K. (2004). *Literacy and health in America*. Princeton, NJ: Educational Testing Service.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., . . . Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly*, *51*, 148–180.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333.

Ware, Jr. J. E., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A user's manual* (3rd ed.). Boston, MA: The Health Institute.

Weiss, B. D., Mays, M. Z., Martz, W., Castro, K. M., DeWalt, D. A., Pignone, M. P., . . . Hale, F. A. (2005). Quick assessment of literacy in primary care: The Newest Vital Sign. *Annals of Family Medicine*, *3*, 514–522.

Willis, S. L., Jay, G. M., Diehl, M., & Marsiske, M. (1992). Longitudinal change and prediction of everyday task competence in the elderly. *Research on Aging*, *14*, 68–91.

Wilson, E. A. H., Park, D. C., Curtis, L. M., Cameron, K. A., Clayman, M. L., Makoul, G., . . . Wolf, M. S. (2010). Media and memory: The efficacy of video and print materials for promoting patient education about asthma. *Patient Education and Counseling*, *80*, 393–398.

Wolf, M. S., Baker, D. W., & Makoul, G. (2007). Physician–patient communication about colorectal cancer screening. *Journal of General Internal Medicine*, *22*, 1493–1499.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's (omega H): Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.