Unbiased measurement instruments are needed to reliably estimate health literacy in diverse populations

Elizabeth A. Hahn

Associate Professor

Northwestern University Feinberg School of Medicine

e-hahn@northwestern.edu

HARC X

October 22, 2018

Health Literacy Measurement

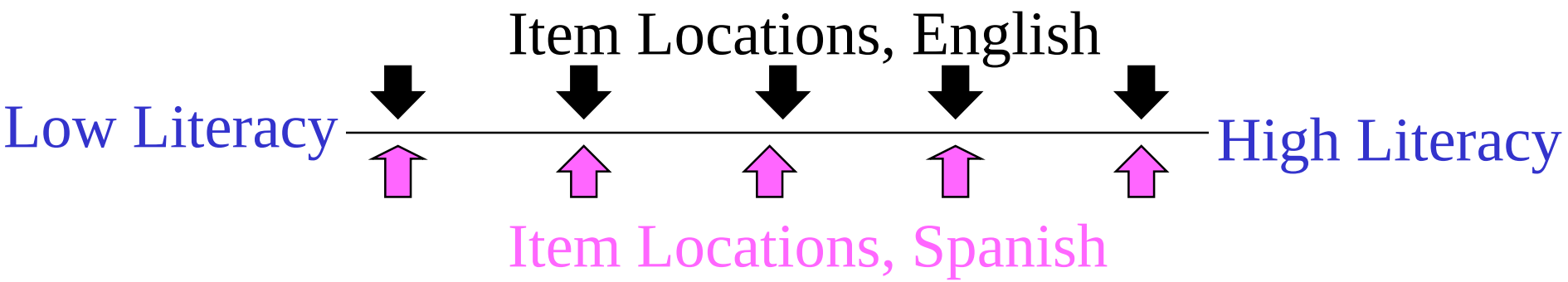
- The link between low health literacy and health-related disparities is well-documented (Berkman et al., 2004; 2011).
- There is a need for better integration of health literacy, health disparities/health equity and patient-centered care initiatives (Hasnain-Wynia et al., 2010; Paasche-Orlow et al., 2010).
- Easy to use and psychometrically sound measurement instruments are needed to reliably estimate health literacy in diverse populations:
  - to facilitate such initiatives
  - to determine the independent effects of limited English proficiency and limited health literacy
- In order to distinguish between Literacy and Language Barriers:
  - English and non-English measures must yield equivalent information

Study Purpose and Sample

- Purpose:
  - to determine whether English and Spanish language versions of Health LiTT (Health Literacy Assessment Using Talking Touchscreen Technology) could use a common set of item calibrations for measure scoring, or if language-specific item calibrations would be required

- Sample:
  - Adults, age 18+, English- or Spanish-speaking, diagnosis of type 2 diabetes and being treated with oral medication or insulin
  - Receiving care in the general medicine clinic of the John H. Stroger, Jr. Hospital of Cook County, Illinois

Item Response Theory (IRT) Item Banks



- the bank of questions defines an underlying trait
- enables test instruments of various lengths and even computerized adaptive tests (CATs)
- the definition of the trait, and the meaning of each item, should be the same across all participant characteristics

  - otherwise, differences due to measurement bias could incorrectly be interpreted as real differences between groups

Health LiTT Prose Item

After a medical test or procedure, be sure to get the results. Ask whether you will get them in person or by phone or by mail. Also ask when you will get the results. Do not assume the results are fine if you do not get them when expected. If you do not get them, call your doctor.

You should _____ all test results.

mail | estimate | obtain | protect

Health LiTT Document Item (English)

Medications for Mr. Beta

Medication	Start Date	End Date	Instructions
Hanebrex: 200 mg tablets	Aug. 27	Sept. 26	1 Tablet daily
Yostatin: 250 mg tablets	Mar. 8	None	1 Tablet twice daily
Nandozol: 90 mcg per puff	Mar. 8	None	1-2 Puffs by mouth every 4-6 hours as needed
Cellacillin: 250 mg tablets	Apr. 22	Apr. 29	2 Tablets on the first day, then 1 Tablet daily after that

Look at the Medications for Mr. Beta. How many tablets of Cellacillin should he take on the third day?

1 | 2 | 3 | 4

Health LiTT Document Item (Spanish)

Medicamentos para el Sr. Beta

Medicamento	Fecha de comienzo	Fecha final	Instrucciones
Hanebrex: pastillas de 200 mg	27 de ago.	26 de sept.	1 pastilla diaria
Yostatin: pastillas de 250 mg	8 de mar.	Ninguna	1 pastilla dos veces al día
Nandozol: 90 mcg por inhalación	8 de mar.	Ninguna	1 a 2 inhalaciones por la boca cada 4 a 6 horas, según sea necesario
Cellacillin: pastillas de 250 mg	22 de abr.	29 de abr.	2 pastillas el primer día y luego 1 pastilla diaria de ahí en adelante

Observe los Medicamentos para el Sr. Beta. ¿Cuántas pastillas de Cellacillin debe tomar el tercer día?

1 | 2 | 3 | 4

Health LiTT Short Form (14 items)

- The 14 items were chosen to represent a mix of item types (prose, document, numeracy) and content (disease/health condition, medical care).
  - 6 Prose items
  - 6 Document items
  - 2 Numeracy items

Psychometric and Statistical Analyses

- Health LiTT uses a multiple-choice format with one correct answer. For item scoring, each response is scored as correct (1) or incorrect (0).
- For scale scoring, IRT-based Bayesian expected a posteriori estimation (EAP) response pattern scoring was conducted, employing either:
  - (a) previously established item parameters derived from the original Health LiTT two-parameter logistic (2PL) model item calibration analyses (Hahn et al., 2011),
  - or (b) language-specific item parameters for DIF-identified items and common-across-groups item parameters for non-DIF items.
- The IRT software package PARSCALE was used for IRT-based scoring (Muraki & Bock, 2003).

Differential Item Functioning (DIF)

- Persons with the same level of overall health literacy are expected to perform similarly on individual Health LiTT items.
- Language (English vs. Spanish) is hypothesized to be a construct-irrelevant grouping factor:
  - High health literate persons, regardless of language, should perform similarly on individual Health LiTT items, and low health literate persons should also perform similarly on individual items.

Part One of the DIF analysis (detection): to identify whether any Health LiTT items displayed DIF by language

- A novel hybrid "logistic ordinal regression (LOR)-plus-IRT" approach to DIF detection was implemented
  - standard DIF detection: liberal McFadden pseudo-R2 change criterion of 0.010
  - sensitivity DIF detection: to increase the ability to detect potential item bias, this criterion was then lowered by half to 0.005
- Evaluated two types of DIF
  - Uniform DIF: bias was constant across varying trait levels
  - Non-uniform DIF: bias varied conditional on trait level

Part Two of the DIF analysis (impact): to evaluate the impact of identified DIF on Health LiTT total scores

- Comparison of unadjusted or "initial" Health LiTT scores to DIF-adjusted or "purified" Health LiTT scores
  - 1) Pearson correlation (initial vs. purified theta scores)
  - 2) a median theta standard error (SE) assessment (the number and percentage of individual difference scores (i.e., initial theta minus purified theta) that exceeded initial theta's median SE
  - 3) an individual theta score standard error (SE) assessment (the number and percentage of individual difference scores that exceeded initial individual theta score SEs)
  - 4) a comparison of Cohen's D language factor effect sizes across competing analyses of variance (ANOVA) (i.e., initial theta scores by language factor vs. purified theta scores by language factor)

Characteristics of Study Participants (Type 2 Diabetes)

	English (*n*=146)	Spanish (*n*=149)	*p*-value
Female	68 (46%)	91 (61%)	0.013
Age in Years, mean (SD)	54.8 (9.8)	54.5 (9.2)	0.674
Ethnicity, Race			
Hispanic, any race	21 (14%)	149 (100%)	--
Non-Hispanic, Black	94 (65%)	--	
Non-Hispanic, Other	31 (21%)	--	
Highest Education			<0.001
Less than HS	35 (24%)	112 (75%)	
HS/GED	46 (32%)	17 (11%)	
More than HS	65 (44%)	20 (14%)	
Prior Computer Use			<0.001
Never	28 (19%)	95 (65%)	
Not in past 12 months	15 (10%)	12 (8%)	
Monthly or weekly	102 (70%)	40 (27%)	
Diabetes Treatment			0.231
Pills only	63 (43%)	70 (47%)	
Insulin only	32 (23%)	23 (15%)	
Pills and insulin	49 (34%)	56 (38%)	
Diabetes Diagnosis ≤2 Years Ago	34 (23%)	27 (18%)	0.318

Characteristics of Study Participants (Type 2 Diabetes)

	English (*n*=146)	**Spanish (*n*=149)**	***p*-value**
Health LiTT T-score	52.1 (10.6)	47.8 (8.9)	0.001
Health LiTT raw score (0-14)	8.5 (3.1)	7.3 (2.9)	0.006

DIF Detection Results

DIF Detection Criterion	Item Type	Content	Type of DIF[a]
Standard: McFadden pseudo-R^2 change ≥ 0.010	Document	how much cold medicine for a child	non-uniform
	Numeracy	when to take the next pill	non-uniform
	Prose	problems associated with vaccines	uniform
	Prose	benefits of vaccines	non-uniform
Sensitivity: McFadden pseudo-R^2 change ≥ 0.005		[all 4 items above]	[see above]
	Prose	surgery expectations	non-uniform
	Numeracy	BMI interpretation	non-uniform

[a] Type of DIF: non-uniform DIF: effect varied across theta levels
uniform DIF: effect constant across theta levels

DIF Impact Results

Type of Score Impact Evidence	Computational Details	Result
Pearson Correlation	initial vs. purified theta	$r = 0.995$
Difference Score	initial minus purified theta	mean difference=0.0005; SD=0.0888
Difference Score Magnitude (1)	difference score vs. median initial theta SE	0 cases (0%) > initial theta's median SE (0.44)
Difference Score Magnitude (2)	difference score vs. individual initial theta SEs	0 cases (0%) > its own individual initial theta SE
Effect Size Stability	Cohen's D effect sizes: initial vs. purified thetas by language factor ANOVAs	initial thetas by language: Cohen's D=0.49 (medium) purified thetas by language: Cohen's D=0.55 (medium)

initial: Health LiTT scores not corrected for DIF

purified: Health LiTT scores corrected for DIF

SD: standard deviation SE: standard error

ANOVA: analysis of variance

Summary and Conclusions

- Although both standard and sensitivity criterion DIF detection analyses identified DIF items, the impact of DIF on Health LiTT scores appeared to be trivial.
- This means that the original English IRT-based item calibrations can be confidently used to score Health LiTT in Spanish.

Study Strengths

- To our knowledge, this is the first study to implement multimedia assessment for self-administration of questionnaires in English- and Spanish-speaking people with type 2 diabetes receiving care in a safety net system.
- State-of-the-science psychometric analyses were conducted
- Health LiTT measures health literacy in English and Spanish:
  - individual scores are estimated on a linear continuum
  - psychometric measurement equivalence across language

Study Limitations

- For IRT-based DIF analyses, a typical sample size expectation per analyzed DIF group might be *n*=200 (e.g., Zumbo, 1999). However, it is also recognized that this may be restrictive, and that additional empirical research may eventually provide better sample size guidance (Zwick, 2012).
  - In part to address the sample size issue (*n*=146 English, *n*=149 Spanish), we conducted sensitivity analyses for DIF.
- Convenience sampling in one safety net clinic for type 2 diabetes care.
- Most Spanish-speaking participants self-identified their ethnicity as Mexican-American, so may not represent all Spanish-speaking Hispanics.

Implications for Research and Clinical Initiatives

- Distinguishing Between Literacy and Language Barriers
  - Item Response Theory (IRT) is a more sensitive method to determine whether some items are culturally or linguistically biased
  - Health LiTT provides better opportunities to determine the independent effects of limited English proficiency and limited health literacy

- Talking Touchscreen (TT) is easy to use and acceptable for self-administration of a health literacy test
  - self-administration should reduce staff burden and costs, reduce interview bias, and reduce stigma of low literacy
  - TT increases access of underserved populations to new technologies, and can contribute information about the experiences of diverse populations with new technologies
  - Health LiTT is easy to integrate with PRO assessments

Acknowledgements

Grant # R18-HS019335 from the Agency for Healthcare Research and Quality

Hahn EA, Kallen MA, Jacobs EA, Ganschow PS, Garcia SF, Burns JL. English-Spanish Equivalence of the Health Literacy Assessment Using Talking Touchscreen Technology (Health LiTT). Journal of Health Communication 19(sup2):285-301, 2014.

Thanks to: Karon Cook for psychometric consultation; Beatriz Menendez for study coordination; Maria Corona and Veronica Valenzuela for recruiting and interviewing participants; Karuna Sehdev for medical chart abstraction; Helena Correia for coordinating Spanish translations; Michael Bass and Manpreet Lakhan for health information technology support; and Joshua Rutsohn for data management and analysis.

Thanks to all the patients who participated in this study!