DOI: 10.1111/j.1475-6773.2010.01119.x
METHODS ARTICLE

Methods

Short Assessment of Health Literacy—Spanish and English: A Comparable Test of Health Literacy for Spanish and English Speakers

Shoou-Yih Daniel Lee, Brian D. Stucky, Jessica Y. Lee, R. Gary Rozier, and Deborah E. Bender

Objective. The intent of the study was to develop and validate a comparable health literacy test for Spanish-speaking and English-speaking populations.
Study Design. The design of the instrument, named the *Short Assessment of Health Literacy—Spanish and English* (*SAHL-S&E*), combined a word recognition test, as appearing in the *Rapid Estimate of Adult Literacy in Medicine* (*REALM*), and a comprehension test using multiple-choice questions designed by an expert panel. We used the item response theory (IRT) in developing and validating the instrument.
Data Collection. Validation of *SAHL-S&E* involved testing and comparing the instrument with other health literacy instruments in a sample of 201 Spanish-speaking and 202 English-speaking subjects recruited from the Ambulatory Care Center at the University of North Carolina Healthcare System.
Principal Findings. Based on IRT analysis, 18 items were retained in the comparable test. The Spanish version of the test, *SAHL-S*, was highly correlated with other Spanish health literacy instruments, *Short Assessment of Health Literacy for Spanish-Speaking Adults* ($r = 0.88$, $p < .05$) and the Spanish *Test of Functional Health Literacy in Adults* (*TOFHLA*) ($r = 0.62$, $p < .05$). The English version, *SAHL-E*, had high correlations with *REALM* ($r = 0.94$, $p < .05$) and the English *TOFHLA* ($r = 0.68$, $p < .05$). Significant correlations were found between *SAHL-S&E* and years of schooling in both Spanish- and English-speaking samples ($r = 0.15$ and 0.39, respectively). *SAHL-S&E* displayed satisfactory reliability of 0.80 and 0.89 in the Spanish- and English-speaking samples, respectively. IRT analysis indicated that the *SAHL-S&E* score was highly reliable for individuals with a low level of health literacy.
Conclusions. The new instrument, *SAHL-S&E*, has good reliability and validity. It is particularly useful for identifying individuals with low health literacy and could be used to screen for low health literacy among Spanish and English speakers.

Key Words. Health literacy, test instrument, Spanish speakers, English speakers, *SAHL-S&E*

It is hardly news anymore that a significant proportion of adults in the United States have difficulty navigating the health care system and managing personal health issues because of inadequate health literacy or limited "capacity to obtain, process, and understand health information and services needed to make appropriate health decisions" (Seldon et al. 2000). Inadequate health literacy, as a growing body of research has shown, is a risk factor for patients' difficulties in understand health information and following medical instructions (Gazmararian et al. 2003; Parker, Ratzan, and Lurie 2003; Davis et al. 2006; Cho et al. 2008), poor disease/self-management knowledge (Gazmararian et al. 2003), underuse of preventive services and routine physician and dental visits (Lindau et al. 2002; Scott et al. 2002; Baker et al. 2004; Lindau, Basu, and Leitsch 2006; Rogers, Wallace, and Weiss 2006; Jones, Lee, and Rozier 2007), increased hospitalizations and medical costs (Baker et al. 2002; Howard, Gazmararian, and Parker 2005), and high mortality rates (Sudore et al. 2006).

Identifying individuals with inadequate health literacy is difficult because information such as age, educational attainment (i.e., years of schooling), and self-reported literacy skills do not reliably reflect an individual's health literacy level (Bass et al. 2002; Davis, Jackson, George, et al., 1993; Davis, Arnold, Berkel, et al., 1996; Nurss, el-Kebbi, Gallina, et al., 1997). Over the years, several instruments, including the *Test of Functional Health Literacy in Adults* (*TOFHLA*), the *Rapid Estimate of Adult Literacy in Medicine* (*REALM*), and the *Newest Vital Sign*, have been developed to assess health literacy in the United States (Davis et al. 1993; Murphy et al. 1993; Parker et al. 1995; Weiss et al. 2005). Most of the instruments, however, have a strong focus on the English-speaking populations and are inappropriate for assessing the health literacy level of Spanish speakers. In the case of *REALM*, an attempt to develop a Spanish version failed because of the phonetic structure of the Spanish language (Nurss et al. 1995).[1] Where a Spanish version is available, for example,

Address correspondence to Shoou-Yih D. Lee, Ph.D., Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina, 1101 McGavran-Greenberg Hall, CB# 7411, Chapel Hill, NC 27599-7411; e-mail: sylee@email.unc.edu. Brian D. Stucky, M.A., is with the Department of Psychology, University of North Carolina, Chapel Hill, NC. Jessica Y. Lee, D.D.S., Ph.D., is with the Department of Pediatric Dentistry, School of Dentistry, University of North Carolina, Chapel Hill, NC. R. Gary Rozier, D.D.S., is with the Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC. Deborah E. Bender, Ph.D., M.P.H., is with the Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC.

TOFHLA-Spanish, the Spanish instrument is usually developed using a rudimentary translation-and-back-translation technique and is not validated psychometrically. A recent study comparing the psychometric properties of the English and Spanish versions of shortened *TOFHLA* raised a significant concern about their comparability (Aguirre, Ebrahim, and Shea 2005).

Our research team developed an easy-to-use health literacy test, the *Short Assessment of Health Literacy for Spanish-speaking Adults* (*SAHLSA*), for Spanish speakers (Lee et al. 2006). The *SAHLSA* is focused on testing an individual's reading ability in the health context. The test contains 50 test items and has good psychometric qualities. It has been adopted in research and clinical practice in the United States (Keselman et al. 2007) and is being validated for use in Latin American countries (Huamán-Calderón, Quiliano-Terreros, and Vílchez-Román 2009). Since the publication of *SAHLSA*, many users have expressed the need for an English version to allow comparisons of health literacy level between Spanish and English speakers for research and clinical purposes. In this paper, we report our subsequent effort to develop a comparable test for Spanish and English speakers, named *Short Assessment of Health Literacy—Spanish and English* (*SAHL-S&E*), based on the same methods used in developing *SAHLSA*. The test contains 18 items and is easy to administer. In taking the test, examinees are asked to read aloud each of the 18 medical terms and then associate each term to another word similar in meaning to demonstrate comprehension. The following sections describe the development of the *SAHL-S&E*, the methods used to validate the instrument, results of the validation, and recommendations for the use of the instrument.

METHODS

Instrument Development

The test items in *SAHL-S&E* were selected from the Spanish and English versions of an instrument that contained the 66 medical terms in the *REALM*, a test of reading ability based on word recognition (Davis et al. 1993). As a departure from *REALM*, we incorporated in the instrument simple multiple-choice questions to assess the examinee's comprehension. Specifically, two common, simple words were chosen to match each of the *REALM* medical terms ("don't know" was also included as an option). One of the words was meaningfully associated with the *REALM* medical term and the other was not. The test is akin to one form of educational achievement testing: "defining,"

which measures understanding or comprehension based on correct identification of a paraphrased version of an original concept, fact, principle, or procedure as presented during instruction (Haladyna 1999). Because the purpose of the multiple-choice questions was to verify the comprehension of the given medical terms, examinees were instructed not to guess. The difficulty of the two added words was kept minimal so that any examinee with a low level of education could understand them.

As reported in Lee et al. (2006), the instrument was developed by an expert panel through a Delphi process. The panel consisted of five experts who were fluent in both English and Spanish and had extensive experience working with Spanish speakers in educational, medical, and public health settings. The panel first translated the 66 *REALM* medical terms into Spanish. The translation took into account both the dictionary definition and the commonality of usage in daily conversations. The panel then selected the key and appropriate distractor for each *REALM* medical term. The process produced both the English and Spanish drafts of the instrument. A pretest with 10 English-speaking and 10 Spanish-speaking subjects found the drafts were appropriate, requiring no further change.

Field Test and Verification of the Association Questions

The field test was conducted with 202 English-speaking and 201 Spanish-speaking respondents, recruited at the Ambulatory Care Center of the University of North Carolina Healthcare System. To be eligible for participation in the study, the subjects had to meet the following criteria: (1) be fluent in either English or Spanish; (2) aged 18 or older but <80 years old; (3) without obvious signs of cognitive impairment; (4) without vision or hearing problems; and (5) showing no sign of drug or alcohol intoxication. The research protocol was approved by the Institutional Review Board at the School of Public Health, the University of North Carolina at Chapel Hill.

The two groups of respondents had similar gender composition; female respondents representing approximately 56 percent of the total sample. On average, Spanish-speaking respondents tended to be younger (34.2 versus 43.7 years) and have fewer years of schooling (10.1 versus 13.0 years) than English-speaking respondents. Around 65 percent of the Spanish-speaking respondents were Mexican. The interview was conducted by six trained bilingual interviewers using a questionnaire that included the 66 test items and questions regarding the respondents' demographic attributes (i.e., years of school-

ing, gender, age, and marital status). Also included in the interview was the *TOFHLA*, used as a comparison in instrument validation.

Using data collected from English-speaking respondents, we were able to verify the design and selection of words for the association (comprehension) test in the instrument. The verification was based on the correlation between the *REALM* score and the association test score. A high correlation ($r = 0.76$) was found, suggesting the design of the association test was adequate.

Psychometric Assessment and Selection of Comparable Items for Spanish- and English Speakers

For the purpose of developing a comparable test for Spanish and English speakers, we used item response theory (IRT). IRT is a modern, model-based, and item-oriented psychometric approach to scale development. In addition to testing the psychometric qualities of test items, it has the capability of examining the equivalence of test items between groups, thereby allowing the development of comparable tests (Embretson and Reise 2000; Ellis and Mead 2002).

IRT assumes that responses to items are related to a single underlying latent variable. We examined this assumption using both exploratory and confirmatory factor analyses of the interitem tetrachoric correlation matrix via the WLMSV algorithm in the software *Mplus* (Muthén and Muthén 2008). Initially, exploratory factor analysis, including the scree plot, was conducted to determine the necessary number of factors to achieve adequate model fit (using evaluation of common fit indices and comparisons of eigenvalues) (Hambleton and Rovinelli 1986). Confirmatory factor analysis was then performed to confirm unidimensionality.

We then performed IRT to calibrate the test items in the Spanish and English versions of the original 66-item instrument. IRT assumes that an examinee's response to an item on a test is related to a latent trait (θ), which the test is presumed to measure. It also assumes that the relationship can be represented by a mathematical function (usually an s-shaped, logistic function) known as an item characteristic curve (ICC). The ICCs of dichotomously scored items are commonly evaluated using three-, two-, and one-parameter logistic models (3PLM, 2PLM, and 1PLM). The 3PLM is written as

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{[1 + \exp\{-Da_i(\theta - b_i)\}]}$$

where $P_i(\theta)$ is the probability that an examinee with ability θ (in this case, reading ability) answers item i correctly; a_i is the discrimination parameter indicating the degree to which small differences in ability are associated with

different probabilities of correctly answering item i; b_i is the difficulty parameter corresponding to the ability level associated with a .50 probability of answering item i correctly; and c_i is the guessing parameter or the probability that an examinee who is infinitely low on the ability answers item i correctly; and D is a scaling constant of 1.7 used to transform the metric from logistic to normal. The 2PLM assumes no guessing and estimates item difficulty and discrimination. The 1PLM estimates item difficulty only and assumes that the discrimination parameter is equal across items. The 2PLM and 3PLM usually provide a better fit for dichotomous items (Embretson and Reise 2000). We examined the relative fit of the two models and estimated the parameters using the *MULTILOG* program (Thissen 1991).

In order to create a comparable test, the psychometric properties of the items must be shown to be equal in both the Spanish- and English-speaking samples. In IRT, the test of differential item functioning (DIF) is used to assess whether item discrepancy exists between separate groups (Embretson and Reise 2000). In the case of the 2PLM, for example, DIF may occur for either the discrimination or difficulty parameter. DIF in a discrimination parameter indicates that an item is more representative of the underlying construct in one group than the other. DIF in a difficulty parameter suggests that an item is more or less difficult in one group than the other, after accounting for overall group differences. In the context of this study, DIF could be interpreted as a Spanish-to-English, or vice versa, translation effect or a potential cultural difference (Orlando and Marshall 2002). Ignoring DIF, therefore, could lead to incorrect conclusions about group differences or similarities.

DIF could also be viewed as an approach to ensuring "construct consistency" between samples. DIF on an item indicates that the construct the item is intended to measure is different between groups. When items with DIF are eliminated, we are left with a set of items that are measuring the same construct in practice. Thus, our goal was to identify items that were DIF-limited so that they could be administered to Spanish and English speakers. DIF analysis was performed using the IRT-LR DIF procedure in the software *IRTLRDIF* (Thissen 2001).

Validity and Reliability Tests

Construct validity and reliability of the comparable test were also examined. In testing construct validity, we performed the following analyses: (1) correlating the Spanish version of the test to *SAHLSA* and Spanish *TOFHLA*,[2] (2) correlating the English version of the test to *REALM* and English *TOFHLA*,[3]

and (3) correlating the examinee's test score to his/her educational attainment (i.e., years of schooling).

Reliability was examined using two approaches. First, we calculated Cronbach α for each version of the test. Cronbach α, a measure of internal reliability, indicates the extent to which the reliability of the test scores was similar across samples. Second, using an IRT-based approach, test information was computed. Differing from the traditional reliability coefficients (e.g., Cronbach α), test information reflects how reliably (or precisely) the *SAHL-S&E* items measure health-related reading ability across the range of literacy (Embretson and Reise 2000; Ellis and Mead 2002).

RESULTS

Examination of Unidimensionality

Before conducting factor analysis, 3 of the 66 items—"flu," "cancer," and "eye"—were removed in both the Spanish- and English-speaking samples because more than 98 percent of the respondents provided correct responses, indicating that those items provided little useful information. For the remaining 63 items in each sample, comparisons of fit indices and interpretability of communalities indicated that a one-factor model fit better than did models with more or fewer factors. Additionally, scree plots show a clear dominance of the first factor. In the Spanish-speaking sample, the eigenvalue for the first factor of the 63 items was over four times larger than that of the second largest, and the second largest eigenvalue was similar to the smaller ones, suggesting the items were indicators of a common, latent factor. Similarly, the eigenvalue of the first factor in the English-speaking sample was over eight times larger than that of the second largest factor (Appendix SA2).

Results of confirmatory factor analysis also indicated generally good fit of the single-factor model (i.e., unidimensionality) in both the Spanish- and English-speaking samples. For the Spanish-speaking sample, the single factor model had a χ^2 value of 76 ($df = 55$, $p = .03$), TLI = 0.935, and RMSEA = 0.044. The corresponding fit indices for the English-speaking samples were $\chi^2 = 61$ ($df = 45$, $p = .058$), TLI = 0.989, and RMSEA = 0.042.

Item Calibration

IRT was conducted separately for the remaining 63 items in each sample. Results from likelihood ratio tests indicated that the 2PLM provided the best fit, suggesting that the effect of guessing was minimal.

Following Lee et al. (2006), we considered items with a discrimination parameter >1.0 but <3.0 (to ensure all items reasonably discriminated between individuals) and a difficulty parameter between -3.0 and $+3.0$ to be satisfactory. Using these criteria, 17 additional items were removed from the English version of the instrument. Notably, most of the removed items had discrimination parameters >3.0. Sixteen items (not necessarily the same) were also removed from the Spanish version. The majority of these items had discrimination parameters <1.0 or threshold parameters <-3.0. Of the remaining items, 32 appeared in both versions of the instrument.

DIF Test

To determine the final set of items for inclusion in the comparable test, DIF analysis was conducted on the 32 common items. Because of the number of statistical tests involved in determining DIF (in this case 32), the Benjamini and Hochberg (1995) correction was used to control for multiple comparisons. Results indicated that 14 of the 32 items had significant DIF (Table 1). The remaining 18 items comprised the comparable test, which we named the *SAHL-S&E*.

Validity and Reliability Tests

SAHL-S was highly correlated with *SAHLSA* ($r=0.88$, $p<.05$) and Spanish *TOFHLA* ($r=0.62$, $p<.05$) in the Spanish-speaking sample. *SAHL-E* also had high correlations with *REALM* ($r=0.94$, $p<.05$) and English *TOFHLA* ($r=0.68$, $p<.05$) in the English-speaking sample. Significant correlations were also found between *SAHL-S&E* and years of schooling in both the Spanish- and English-speaking samples ($r=0.15$, $p<.05$ and $r=0.39$, $p<.05$, respectively).

SAHL-S&E displayed satisfactory reliability of 0.80 and 0.89 in the Spanish- and English-speaking samples, respectively. The test information function indicates that scores on the *SAHL-S&E* are highly reliable (i.e., $>\alpha=0.90$) for individuals with a low level of reading ability (i.e., between approximately -3 and -1 SD below the mean) (Appendix SA3).

Finally, we examined the plot of *SAHL-S&E* scores vis-à-vis *SAHLA-50*, English *TOFHLA*, and *REALM* scores and determined that subjects with a *SAHL-S&E* score between 0 and 14 had a significant chance (76–85 percent) of being classified as having low health literacy based on these other instruments. Additional analyses of association confirmed that *SAHL-S&E* ≤ 14 represented a proper cutoff point for low health literacy (or, more specifically, low health-related reading ability). Based on this criterion, 54 (27.0 percent) of the

Table 1: Results of the DIF Analysis

English Item	*Discrimination a*	*Difficulty b*	*Spanish Item*	*Discrimination a*	*Difficulty b*	*DIF*
Dose	1.08	– 1.67	Dosis	1.37	– 1.95	
Nerves	2.17	– 1.35	Nervios	1.75	– 1.69	
Kidney	1.72	– 3.10	Riñón	2.01	– 2.05	
Hormones	1.48	– 1.65	Hormonas	1.41	– 1.45	
Herpes	1.38	– 2.53	Herpes	1.28	– 0.90	*
Caffeine	1.03	– 2.49	Cafeína	0.92	– 1.17	*
Incest	1.90	– 1.27	Incesto	1.46	0.15	*
Asthma	2.10	– 1.57	Asma	3.21	– 2.26	*
Seizure	1.49	– 1.76	Convulsiones	1.25	– 2.39	
Depression	1.63	– 2.39	Depresión	2.13	– 1.49	*
Infection	2.18	– 2.27	Infección	1.57	– 2.36	
Pregnancy	1.88	– 1.92	Embarazo	2.00	– 1.80	
Syphilis	0.93	– 0.14	Sífilis	1.50	0.13	
Abnormal	1.70	– 1.54	Anormal	1.36	– 1.36	
Nutrition	1.21	– 2.28	Nutrición	2.38	– 1.59	
Miscarriage	1.56	– 2.28	Aborto espontáneo	1.68	– 1.93	
Hemorrhoids	1.17	– 0.84	Hemorroides	1.47	– 1.13	
Directed	1.91	– 1.43	Indicado	1.09	– 1.47	
Irritation	1.41	– 1.46	Irritación	1.03	– 1.04	*
Alcoholism	1.59	– 2.00	Alcoholismo	1.94	– 2.32	
Sexually	0.63	0.19	Sexualmente	1.06	– 1.87	*
Colitis	1.55	0.80	Colitis	1.20	– 0.56	*
Testicle	1.51	– 1.31	Testículo	1.27	– 0.72	*
Occupation	1.63	– 2.37	Empleo	2.30	– 2.42	
Constipation	1.51	– 1.25	Estreñimiento	1.25	– 1.90	
Medication	1.51	– 2.31	Medicamento	1.88	– 2.36	
Diagnosis	1.36	– 1.23	Diagnóstico	1.85	– 1.28	
Osteoporosis	1.77	– 0.16	Osteoporosis	0.95	– 1.48	*
Prostate	1.08	– 1.53	Próstata	1.23	– 0.58	*
Hepatitis	0.81	– 1.55	Hepatitis	1.60	– 0.67	*
Anemia	1.93	– 0.66	Anemia	1.46	– 2.18	*
Obesity	1.59	– 1.28	Obesidad	2.29	– 1.53	*

*Significant difference ($p<.05$) in item parameters between the Spanish- and English-speaking samples using the Benjamini–Hochberg correction.

DIF, differential item functioning.

Spanish speakers and 48 (23.8 percent) of the English speakers in our sample had a low level of health literacy.

DISCUSSION

This paper reports the development of *SAHL-S&E*, designed to provide a comparable test of health literacy for Spanish-speaking and English-speaking

populations. Results show that the instrument has good validity and reliability. Guessing does not appear to be a concern if clear instruction is given before the test. The instrument contains only 18 items and is easy to administer. We estimate that the administration would take only 2–3 minutes and require minimal training. (The Spanish and English version of *SAHL-S&E* and the user guides are included in the Appendix.) A rather high cutoff point is found for low health literacy (≤ 14), suggesting that the *SAHL-S&E* is particularly useful for identifying individuals with low health literacy. The test information function confirms that the instrument is highly reliable at the lower range of scores.

In validating the instrument, we found that *SAHL-S* had a higher correlation with *SAHLSA* than with Spanish *TOFHLA*. Similarly, the correlation between *SAHL-E* and *REALM* was higher than that between *SAHL-E* and English *TOFHLA*. The findings may reflect the fact that the design of *SAHL-S&E*, essentially a word recognition test of reading ability, is the same as *SAHLSA* and similar to *REALM*. We also found that the resulting instrument had a lower correlation with years of schooling in the Spanish-speaking sample. There are two plausible explanations. First, in comparison with education experience of English speakers, Spanish speakers, whose education was obtained in multiple countries and systems, may have more heterogeneous education experience. Second, although consistent with the standard testing in the U.S. education system, the format of the test (a pronunciation test and a multiple-choice test for comprehension) may be unusual for Spanish speakers. In other words, Spanish-speaking respondents in our sample, compared with English-speaking respondents with the same level of formal schooling, may be less familiar with the multiple-choice format of the test and thus have a poorer performance on the test.

Several limitations are worth noting. The instrument was developed based on standard, "dictionary" Spanish and English. Further testing of the instrument may be needed in different Spanish- and English-speaking subpopulations who are accustomed to using different idiomatic expressions. As with other health literacy instruments such as *TOFHLA* and *REALM*, *SAHL-S&E* is a reading test. It assesses specifically an individual's reading skill in the health care context. The design is based on the assumption that reading ability is a basic literacy skill, without which patients would have difficulty functioning in and negotiating the health care system. However reasonable the assumption is, it should be noted that the instrument does not capture other skills such as numeracy and interpersonal communication that may also be important in health care. Furthermore, similar to prior instrument development studies, our study did not include a random, representative sample of Spanish

speakers and English speakers in the community. The clinic-based participants recruited for the study may be more receptive to a health literacy test. What kind of difficulties may arise in applying the *SAHL-S&E* to a community-based sample remains to be evaluated. Finally, as we have noted, the instrument is particularly suitable for identifying individuals with low health literacy. For individuals with a >14 score, the instrument may not be sensitive enough to distinguish different health literacy levels.

Despite these limitations, the instrument is robust and has several practical applications. First, unlike other instruments, the comparability between the Spanish and English versions of the instrument is established through rigorous psychometric evaluation. It offers a reliable way to assess and compare the level of low health literacy between Spanish and English speakers.

Second, the instrument may be used to screen for individual health literacy level in public health and clinical settings that serve a high concentration of English-speaking or Spanish-speaking patients or a mixed patient population. Being able to identify patients with low health literacy can alert health care providers to the possibility that these patients may have difficulty with printed educational materials, communicating their symptoms to physicians, or following medical instructions (Bass et al. 2002; Chew, Bradley, and Boyko 2004; Institute of Medicine 2004). Increased awareness among health care practitioners of the special health and personal needs of low health literacy patients may help reduce the level of linguistic complexity used in provider–patient communications, thus preventing serious medical errors due to misunderstanding. This, in turn, has the potential to improve quality of care and reduce health care cost. These potential advantages asides, the value of health literacy screening may still be debatable because of concerns about patient stigmatization and embarrassment (Parikh et al. 1996; Wolf et al. 2007). Two recent studies suggest that patients are not adverse to health literacy screening if protection of personal information is exercised (Ryan et al. 2008; VanGeest, Welch, and Weinber 2010). However, more research is needed to assess the conditions under which health literacy screening may be appropriate in clinical settings.

Third, the instrument could be used to assess the level of health literacy in local communities. The information could be used to guide the design of appropriate health educational materials (written and/or multimedia) or for devising community intervention programs that are comparable with the health literacy level of the local population (Brandes 1996; Davis et al. 1998).

Finally, a comparable health literacy instrument for Spanish and English speakers would facilitate comparisons in research. Instead of stratifying subjects

on language in health literacy research, researchers could combine samples and use *SAHL-S&E* to identify those with low health literacy in their analysis.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Data for this study were collected with support from the Agency for Healthcare Research and Quality (R03-HS13233) and the analysis was supported by grants from the National Institute of Dental and Craniofacial Research (RO1 DE018236 and RO1DE0180451).

Disclosures: None.

Disclaimers: The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality, the National Institute of Dental and Craniofacial Research, or the University of North Carolina at Chapel Hill.

NOTES

1. In comparison to English, Spanish has regular phoneme–grapheme correspondence, meaning that one sound is usually represented by one letter and vice versa. Therefore, it is relatively easy to pronounce words in Spanish so long as one can recognize letters, and a low-level reader can usually score high on a word recognition test. This feature of the Spanish language violates the design basis of the *REALM* that there exists a high correspondence between reading ability and comprehension.
2. In a previous study, the *SAHLSA* score was found to be significantly and positively associated with the physical health status of Spanish-speaking subjects ($p < .05$), holding constant age and years of education (Lee et al. 2006). The instrument also displayed high internal reliability (Cronbach $\alpha = 0.92$) and test–retest reliability (Pearson $r = 0.86$).
3. *REALM* has good correlation scores, ranging from 0.88 to 0.97, with three other general reading tests. Its test–retest reliability is 0.99 (Davis et al. 1993). English *TOFHLA* has a high correlation with *REALM* ($r = 0.84$). Its test–retest reliability is 0.98 (Parker et al. 1995).

REFERENCES

Aguirre, A. C., N. Ebrahim, and J. A. Shea. 2005. "Performance of the English and Spanish S-TOFHLA among Publicly Insured Medicaid and Medicare Patients." *Patient Education and Counseling* 56 (3): 332–39.

Baker, D. W., J. A. Gazmararian, M. V. Williams, T. Scott, R. M. Parker, D. Green, J. Ren, and J. Peel. 2002. "Functional Health Literacy and the Risk of Hospital Admission among Medicare Managed Care Enrollees." *American Journal of Public Health* 9 (8): 1278–83.

———. 2004. "Health Literacy and Use of Outpatient Physician Services by Medicare Managed Care Enrollees." *Journal of General Internal Medicine* 19 (3): 215–20.

Bass, P. F. I., J. F. Wilson, C. H. Griffith, and D. R. Barnett. 2002. "Residents' Ability to Identify Patients with Poor Literacy Skills." *Academic Medicine* 77 (10): 1039–41.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of Royal Statistical Society: Series B* 57: 289–300.

Brandes, W. L. 1996. *Literacy, Health and the Law: An Exploration of the Law and the Plight of Marginal Readers within the Health Care System: Advocating for Patients and Providers.* Philadelphia, PA: Health Promotion Council of Southeastern Pennsylvania Inc.

Chew, L. D., K. A. Bradley, and E. J. Boyko. 2004. "Brief Questions to Identify Patients with Inadequate Health Literacy." *Family Medicine* 36: 588–94.

Cho, Y. I., S. Y. Lee, A. M. Arozullah, and K. S. Crittenden. 2008. "Effects of Health Literacy on Health Status and Health service Utilization amongst the Elderly." *Society Science Medicine* 66 (8): 1809–16.

Davis, T. C., C. Arnold, H. J. Berkel, I. Nandy, R. H. Jackson, and J. Glass. 1996. "Knowledge and Attitude on Screening Mammography among Low-Literate, Low-Income Women." *Cancer* 78 (9): 1912–20.

Davis, T. C., R. H. Jackson, R. B. George, S. W. Long, D. Talley, P. W. Murphy, E. J. Mayeaux, and T. Truong. 1993. "Reading Ability in Patients in Substance Misuse Treatment Centers." *International Journal of Addiction* 28 (6): 571–82.

Davis, T. C., S. W. Long, R. H. Jackson, E. J. Mayeaux, R. B. George, P. W. Murphy, and M. A. Crouch. 1993. "Rapid Estimate of Adult Literacy in Medicine: A Shortened Screening Instrument." *Family Medicine* 25 (6): 391–95.

Davis, T. C., R. Michielutte, E. N. Askov, M. V. Williams, and B. D. Weiss. 1998. "Practical Assessment of Adult Literacy in Health Care." *Health Education and Behavior* 25 (5): 613–24.

Davis, T. C., M. S. Wolf, P. F. Bass, M. Middlebrooks, E. Kennen, D. W. Baker, C. L. Bennett, R. Durazo-Arvizu, A. Bocchini, and S. Savory. 2006. "Low Literacy Impairs Comprehension of Prescription Drug Warning Labels." *Journal of General Internal Medicine* 21 (8): 847–51.

Ellis, B. B., and A. D. Mead. 2002. "Item Analysis: Theory and Practice Using Classical and Modern Test Theory." In *Handbook of Research Methods in Industrial and Organizational Psychology*, edited by S. G. Rogelberg, pp. 324–43. Malden, MA: Blackwell.

Embretson, S. E., and S. P. Reise. 2000. *Item Response Theory for Psychologists.* Hillsdale, NJ: Erlbaum.

Gazmararian, J. A., M. V. Williams, J. Peel, and D. W. Baker. 2003. "Health Literacy and Knowledge of Chronic Disease." *Patient Education Counseling* 51 (3): 267–75.

Haladyna, T. M. 1999. *Developing and Validating Multiple-Choice Test Items.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., and R. J. Rovinelli. 1986. "Assessing the Dimensionality of a Set of Test Items." *Applied Psychological Measurement* 10: 287–302.

Howard, D. H., J. Gazmararian, and R. M. Parker. 2005. "The Impact of Low Health Literacy on the Medical Costs of Medicare Managed Care Enrollees." *American Journal of Medicine* 118 (4): 371–7.

Huamán-Calderón, D., R. Quiliano-Terreros, and C. Vílchez-Román. 2009. "Embarazo no deseado y fuentes de información impresas y audiovisuales, en mujeres peruanas (2004–2005) [Unwanted pregnancy and access to printed media in Peruvian women]." *Revista Médica de Chile* 137: 46–52.

Institute of Medicine. 2004. *Health Literacy: A Prescription to End Confusion.* Washington, DC: The National Academy of Sciences.

Jones, M., J. Y. Lee, and R. G. Rozier. 2007. "Oral Health Literacy among Adult Patients Seeking Dental Care." *Journal of American Dental Association* 038: 1199–208.

Keselman, A., T. Tse, J. Crowell, A. Browne, L. Ngo, and Q. Zeng. 2007. "Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study." *Journal Medical Internet Research* 9 (1): e5.

Lee, S. Y., D. E. Bender, R. E. Ruiz, and Y. I. Cho. 2006. "Development of an Easy-To-Use Spanish Health Literacy Test." *Health Service Research* 41 (4, part 1): 1392–412.

Lindau, S. T., A. Basu, and S. A. Leitsch. 2006. "Health Literacy as a Predictor of Follow-Up after an Abnormal Pap Smear: A Prospective Study." *Journal of General Internal Medicine* 21: 829–34.

Lindau, S. T., C. Tomori, T. Lyons, L. Langseth, C. L. Bennett, and P. Garcia. 2002. "The Association of Health Literacy with Cervical Cancer Prevention Knowledge and Health Behavior in a Multiethnic Cohort of Women." *American Journal of Obstetrics and Gynecology* 186: 938–43.

Murphy, P. W., T. C. Davis, S. W. Long, R. H. Jackson, and B. C. Decker. 1993. "Rapid Estimate of Adult Literacy in Medicine (REALM): A Quick Reading Test for Patients." *Journal of Reading* 37: 121–30.

Muthén, L. K., and B. O. Muthén. 2008. *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Nurss, J. R., D. W. Baker, T. C. David, R. M. Parker, and M. V. Williams. 1995. "Difficulties in Functional Health Literacy Screening in Spanish-Speaking Adults." *Journal of Reading* 38: 632–7.

Nurss, J. R., I. M. el-Kebbi, D. L. Gallina, D. C. Ziemer, V. C. Musey, S. Lewis, Q. Liao, and L. S. Phillips. 1997. "Diabetes in Urban African Americans: Functional Health Literacy of Municipal Hospital Outpatients with Diabetes." *Diabetes Education* 23 (5): 563–8.

Orlando, M., and G. N. Marshall. 2002. "Differential Item Functioning in a Spanish Translation of the PTSD Checklist: Detection and Evaluation of Impact." *Psychological Assessment* 14: 50–9.

Parikh, N. S., R. M. Parker, J. R. Nurss, D. W. Baker, and M. V. Williams. 1996. "Shame and Health Literacy: The Unspoken Connection." *Patient Education and Counseling* 27 (1): 33–9.

Parker, R. M., D. W. Baker, M. V. Williams, and J. R. Nurss. 1995. "The Test of Functional HEALTH Literacy in Adults: A New Instrument for Measuring Patients' Literacy Skills." *Journal of General Internal Medicine* 10 (10): 537–41.

Parker, R. M., S. C. Ratzan, and N. Lurie. 2003. "Health Literacy: A Policy Challenge for Advancing High-Quality Health Care." *Health Affairs* 22 (4): 147–153.

Rogers, E. S., L. S. Wallace, and B. D. Weiss. 2006. "Misperceptions of Medical Understanding in Low-Literacy Patients: Implications for Cancer Prevention." *Cancer Control* 13 (3): 225–29.

Ryan, J. G., F. Leguen, B. D. Weiss, S. Albury, T. Jennings, F. Velez, and N. Salibi. 2008. "Will Patients Agree to Have Their Literacy Skills Assessed in Clinical Practice?" *Health Education Research* 23: 603–11.

Scott, T. L., J. A. Gazmararian, M. V. Williams, and D. W. Baker. 2002. "Health Literacy and Preventive Health Care Use among Medicare Enrollees in a Managed Care Organization." *Medical Care* 40 (5): 395–404.

Seldon, C. R., M. Zorn, S. C. Ratzan, and R. M. Parker. 2000. *National Library of Medicine Current Bibliographies in Medicine: Health Literacy.* Bethesda, MD: National Institutes of Health.

Sudore, R. L., K. Yaffe, S. Satterfield, T. B. Harris, K. M. Mehta, E. M. Simonsick, A. B. Newman, C. Rosano, R. Rooks, S. M. Rubin, H. N. Ayonayon, and D. Schillinger. 2006. "Limited Literacy and Mortality in the Elderly: The Health, Aging, and Body Composition Study." *Journal of General Internal Medicine* 21 (8): 806–12.

Thissen, D. 1991. *MULTILOG User's Guide: Multiple Categorical Item Analysis and Test Scoring Using Item Response Theory.* Chicago: Scientific Software International Inc.

———. 2001. *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning.* Chapel Hill, NC: University of North Carolina.

VanGeest, J. B., V. L. Welch, and S. J. Weinber. 2010. "Patients' Perceptions of Screening for Health Literacy: Reactions to the Newest Vital Sign." *Journal of Health Communication*, in press.

Weiss, B. D., M. Z. Mays, W. Martz, K. M. Castro, D. A. DeWalt, M. P. Pignone, J. Mockbee, and F. A. Hale. 2005. "Quick Assessment of Literacy in Primary Care: The Newest Vital Sign." *Annals of Family Medicine* 3 (6): 514–22.

Wolf, M. S., M. V. Williams, R. M. Parker, N. S. Parikh, A. W. Nowlan, and D. W. Baker. 2007. "Patients' Shame and Attitudes toward Discussing the Results of Literacy Screening." *Journal of Health Communication* 12 (8): 721–32.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Eigenvalue Plot for the 63-Item Instruments in Spanish (top) and English (bottom).

Appendix SA3: SAHL-S&E Test Information Functions.

Appendix SA4: The 18 Items of *SAHL-S*, Rank-Ordered According to the Parameter *b* of Item Difficulty (Keys and Distracters Are Listed in the Same Random Order as in the Field Interview).

Appendix SA5: Instruction for Administering *SAHL-S*.

Appendix SA6: The 18 Items of *SAHL-E*, Rank-Ordered According to the Parameter *b* of Item Difficulty (Keys and Distracters Are Listed in the Same Random Order as in the Field Interview).

Appendix SA7: Instruction for Administering *SAHL-E*.