Differential Item Functioning by Survey Language Among Older Hispanics Enrolled in Medicare Managed Care

A New Method for Anchor Item Selection

Claude Messan Setodji, PhD, Steven P. Reise, PhD,† Leo S. Morales, MD, PhD,‡ Marie N. Fongwa, RN, PhD,§ and Ron D. Hays, PhD*‖*

Objective: To propose a permutation-based approach of anchor item detection and evaluate differential item functioning (DIF) related to language of administration (English vs. Spanish) for 9 questions assessing patients' perceptions of their providers from the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Medicare 2.0 survey.

Method and Study Design: CAHPS 2.0 health plan survey data collected from 703 Hispanics who completed the survey in Spanish were matched on personal characteristics to 703 Hispanics that completed the survey in English. Steps to be followed for the detection of anchor items using the permutation tests are proposed and these tests in conjunction with item response theory were used for the identification of anchor items and DIF detection.

Results: Of the questions studied, 4 were selected as anchor items and 3 of the remaining questions were found to have DIF ($P < 0.05$). The 3 questions with DIF asked about seeing the doctor within 15 minutes of the appointment time, respect for what patients had to say, and provider spending enough time with patients.

Conclusions: Failure to account for language differences in CAHPS survey items may result in misleading conclusions about disparities in health care experiences between Spanish and English speakers. Statistical adjustments are needed when using the items with DIF.

From the *RAND Corporation, Pittsburgh, PA; †Department of Psychology, UCLA, Los Angeles, CA; §UCLA School of Nursing, Los Angeles, CA; ‖UCLA School of Medicine, Los Angeles, CA; and ‡Group Health Research Institute, Seattle, WA.

Supported by the University of California, Los Angeles, Resource Centers for Minority Aging Research Center for Health Improvement of Minority Elderly (RCMAR/CHIME) under NIH/NIA Grant P30-AG021684 (to C.M.S.). Also supported by UCLA/ DREW Project EXPORT, NCMHD, 2P20MD000182, and the UCLA Older Americans Independence Center, NIH/NIA Grant P30-AG0287 (to R.D.H.).

The content of this article does not necessarily represent the official views of the NIA or the NIH.

Reprints: Claude Messan Setodji, PhD, 4570 5th Avenue, Suite 600, Pittsburgh, PA 15213. E-mail: setodji@rand.org.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's web site (www.lww-medicalcare.com).

ISSN: 0025-7079/11/4905–0461

Key Words: measurement equivalence, differential item functioning, item response theory, patient-assessed quality of care, permutation test

(*Med Care* 2011;49: 461–468)

Consumer assessment surveys are used widely by health plans, health care providers, purchasers, and policymakers for quality assessment and improvement, and by consumers to choose the most appropriate health professionals, group practices, and health plans suitable to their needs.[1] The Consumer Assessments of Healthcare Providers and Systems (CAHPS) surveys have been widely used to measure consumer's experiences with providers and to study health disparities among racial-ethnic groups in the United States.[2] For example, Weech-Maldonado et al[3] found that those patients who speak a language other than English at home report less timely care and less adequate staff helpfulness than other patients. This could mean that there are disparities in care between patients who speak English fluently and those who do not. However, these 2 groups may differ in their interpretation of the survey questions or in their proclivities to respond to survey questions. With cultural and linguistic minority groups steadily increasing in the United States, understanding cultural and linguistic differences in responding to surveys among subgroups of the population can ensure the accuracy of inferences made about disparities between them.

Hispanics constitute the largest and fastest growing ethnic minority group in the United States, making up to 12% of the population and they are expected to double in size to 24% of the US population by the year 2050.[4] Data from the 2007 American Community Survey[5] show that Spanish was the primary language spoken at home for more than 34.5 million people, of whom almost half are limited in their English proficiency. National surveys that include geographic areas with a large proportion of Spanish-speakers routinely provide survey questions in both English and Spanish. As the demand for Spanish surveys continues to grow, it will be necessary to ensure that Spanish versions of instruments are culturally appropriate and psychometrically equivalent to their English versions.

Differential item functioning (DIF) occurs when there are group-mediated differences in the response patterns to survey questions (items) even when the individuals in the

different groups are equivalent on the measured construct.[6] For example, DIF occurs when one group (eg, Spanish speakers) receives different scores as compared with another group (eg, English speakers), despite having equal standing on the underlying construct of interest (eg, health care access perception). There are many different possible causes for DIF including poor translations and lack of semantic or conceptual equivalence. Identifying DIF between cultural and linguistic subgroups in CAHPS surveys is important because the outcomes should reflect unbiased differences in experiences with health care providers, and not reflect differences due to other factors such as age, gender, language spoken, or cultural differences.

Few studies have evaluated the psychometric equivalence of CAHPS items in different subgroups. Marshall et al[7] used confirmatory factor analysis (CFA) to show similarity in the underlying factors in the CAHPS 1.0 survey between Latinos and non-Latinos with Medicaid or commercial insurance. Bann et al[8] found DIF in some items between English- and Spanish-speaking fee-for-service beneficiaries on the CAHPS 2.0 Medicare survey, using item response theory (IRT). In this study, we evaluated the equivalence of English and Spanish versions of the CAHPS 2.0 Medicare Survey using a new semiparametric permutation test to identify anchor items in combination with an IRT-based method for DIF detection.

METHODS

Data Sources

The Center for Medicare and Medicaid Services collects information on experiences with health plans from Medicare beneficiaries who are at least 18 years old, living in the United States or Puerto Rico, and enrolled in managed care organizations each year.[9] In 2002, 600 members from each of 321 participating Medicare managed care plans[10] were sent a CAHPS 2.0 Medicare survey in the language of their choice (English or Spanish). A total of 184,782 participants completed the survey (82% response rate), with 84% completed by mail and the rest by telephone. The CAHPS surveys were designed for both English and Spanish, with focus groups and cognitive interviews conducted in both languages to maximize conceptual language equivalence.[11] The readability of the CAHPS surveys is estimated to be seventh grade level for both languages.[12]

Participants

A total of 10,078 Hispanic CAHPS survey respondents completed the survey, with 8496 completing in English and 1582 in Spanish. Hispanic Spanish responders tended to be younger and less educated than the Hispanic English survey responders. As such, if this full sample of 10,078 Hispanic participants was analyzed, any observed language DIF might possibly be attributed to age or education differences between the 2 language groups. To control for any possibility of confounding factors that can affect causal inferences about item functioning,[13] we were able to identically match 703 Spanish-speaking Hispanic survey responders to English-speaking Hispanic responders on self-rated overall health, age, gender, and education distribution. Therefore, this study is limited to the 1406 Hispanic participants in CAHPS (48% by mail and 52% by phone), composed of identical/similar group of 703 Spanish- and 703 English-speaking responders.

Measures

The CAHPS 2.0 core health plan survey used in this study includes 9 questions asking for reports about specific experiences with doctors' offices (Table, Supplemental Digital Content 1, online only, available at: http://links.lww.com/MLR/A160 which describes the CAHPS items studied). These items represent the following 4 of the CAHPS composites: getting care quickly/timeliness, provider communication, office staff helpfulness, and getting needed care. The items were analyzed using a 3 ordered-category response scale, a big problem, a small problem, not a problem for the getting needed care item and never/sometimes, usually, always for all the other items.

Analysis Plan

We used an exact paired matching method based on ethnicity, age, self-rated overall health, gender, and education to match Hispanics who completed the survey in English to the 703 Hispanic participants who completed the survey in Spanish. χ^2 statistics were used to compare the sample used in the study with the nonmatched group dropped. To determine the presence of DIF in the CAHPS evaluations of providers, responses from Spanish and English speakers were compared using IRT.

We evaluated the extent to which the 9 CAHPS items had sufficient unidimensionality for IRT analyses,[14] using CFA methods described elsewhere.[15,16] Samejima's graded response model[17] (Text, Supplemental Digital Content 2, online only, available at: http://links.lww.com/MLR/A161 which describes the models and additional definitions) was used to estimate 2 types of parameters for each item. The slope or discriminating parameter measures the degree to which an item is related to the latent construct and how quickly the probability of endorsing a response option increases with increasing "trait level," while the location parameter estimates the level on the underlying scale to have a 50% probability of endorsing a particular item response option. The underlying scale can be envisioned as a summary score, where higher values are reflective of more positive experiences with care. Bias or DIF occurs when Spanish and English speakers have different location parameters (uniform DIF) or different slope parameters (nonuniform DIF) indicative of the relationship between an item and the latent construct being stronger in one language group than the other.[18]

DIF analysis requires the identification of anchor items for which the groups perform similarly and that will allow responses from the 2 groups to be linked in a way that parameters can be estimated in a common metric.[19] Multiple methods have been proposed for the selection of anchor items,[20,21] but the "purification" model-based likelihood ratio test method is commonly used.[18] The likelihood-based approach compares 2 models, one constraining the item of interest's parameters to be equal across the 2 groups and the

other estimating those parameters separately for the 2 groups, under the assumption that "all other items are DIF-free." A significant likelihood χ^2 statistic is then used as an indication of potential DIF,[22] and the procedure repeated with potential DIF items dropped.[23] Woods[24] proposed a rank-based approach applicable to any method of DIF testing for the empirical improvement of anchor items selection.

The importance of anchor items detection has been previously discussed[25] and is not trivial when there is a large amount of DIF in the items of interest. Zenisky et al[26] found that when at least 30% of items have DIF, lack of identification of anchor items can result in a substantial change in items that are detected as having DIF. Finch[27] noted that contamination of anchor items was an important limitation, particularly for the likelihood ratio-based approach, a finding echoed by Finch and French[28] in a simulation study. Wang[29] inferred that these limitations arise when the number of items with DIF increases because the assumption of "all other items are DIF-free" in the purification procedure will be incorrect in some cases. Shih and Wang[30] on the other hand reported that purification can result in a nearly perfect rate in selecting up to 4 DIF-free items. In this article, we propose a new iterative method for the identification of anchor items based on semiparametric permutation tests.

In anchor item selection procedures, a significance test evaluates whether the difference between the 2 models posited (parameter constraint vs. no parameter constraint) could reasonably occur "just by chance" in a selection of a random sample. If such evidence is not found, an inference can be made about the observed difference being present in the population. The sampling distribution of the difference between parameters in a DIF-free environment can be used to evaluate whether the observed difference is likely to have only occurred by chance. Our proposed permutation test follows 3 steps:

1. First, for each item, fit the 2 models (constrained vs. not constrained) under the assumption that "all other items are DIF-free" and then estimate the difference between the parameters obtained from both models, a statistic that represents the model difference (S_{data}).
2. Second, randomly assign survey responders to pseudo Spanish or English groups (ie, permutation). This will be the ideal case scenario where there is no differential functioning of the items by language. With this new assignment of responders to the 2 groups, models in Step 1 are then fitted again and the model difference will now be referred to as S_k, where (k) indicates the permutated sample. The random assignment to group or permutation will be done P times (say $P = 1000$), and the empirical distribution of the difference statistic obtained by S_1, S_2, …, S_P, will represent what the distribution of the statistic would have been, if there were no DIF in the item between the groups of interest.
3. In a third step, using the permutation empirical distribution S_1, S_2, …, S_P, a 2-sided test statistic is obtained and a P value more than 0.05 was used to infer that the item does not have DIF and can be considered a potential anchor item.

After all potential anchor items are identified, the above procedure is repeated using only the potential anchor items until no additional item with DIF is observed. The final "DIF-free" items obtained are used as anchor items in the DIF analysis.

To test for DIF in the nonanchor items, we used the likelihood-based approach, but the identified anchor items are used as the only anchors and remain the same throughout the tests. To adjust for multiple comparisons, the Benjamini-Hochberg procedure[31] was used. The permutations were done using SAS version 9.1 and all the other analyses were conducted using MULTILOG 7[32] and Item Response Theory Likelihood-Ratio tests for Differential Item Functioning (IRTLRDIF).[22] Results from the permutation method were qualitatively contrasted with IRTLRDIF estimates.

RESULTS

Sample Description

Table 1 provides descriptive characteristics of the different Hispanic samples. The matched and unmatched samples were different with respect to several characteristics. The matched sample had more women, was less educated, and reported poorer health. These differences on the studied characteristics suggest that if the full sample of Hispanics were used for DIF detection, any of the detectable DIF might not necessarily be attributable to the difference in language. But with the matched sample of 1406 Hispanics (703 English and 703 Spanish), a more definite link can be made between observed DIF and survey language.

Items Description and Dimensionality

The means and standard deviations of the 9 CAHPS items and the total score (averaged over items), as well as a brief description of the items, are displayed in Table 2. The observed total scores had an average score of 2.49 in the Spanish group, statistically smaller than the 2.54 average in the English group. For all items except items 2 and 9, the English group reported more positive experiences.

An exploratory factor analysis conducted in Mplus16 showed a first eigenvalue that was 5.2 and a second of 0.90, providing support for unidimensionality for the items. Within the language groups, internal consistency reliability (alpha) was 0.85 for English speakers and 0.83 for Spanish speakers. CFAs using weighted least square mean and variance adjusted also supports unidimensionality (Table 2): comparative fit index (>0.98), Tucker-Lewis Index (>0.99), and residual correlations (<0.07).

Identification of Anchor and Study Items

Using Step 1 described above on item 1, the discrimination parameter was 1.34 from the constrained model, and 1.37 and 1.29 (ie, difference of 0.08) for Spanish and English, respectively, from the unconstrained model (Table, Supplemental Digital Content 3, online only, available at: http://links.lww.com/MLR/A162 which reported all parameter estimates). Across all the items, the magnitude of the difference in the discrimination parameter varied from 0.02 to 0.83. For the first and second location parameters, the difference varied from 0.01 to 1.50 and from 0.03 to 2.33,

TABLE 1. Characteristics (%) the Sample Included and the One Dropped From the Analysis Due to Nonmatching

Variable	All Hispanics	Matched Sample Used	English Speakers Not Matched	Spanish Speakers Not Matched
Sample size	10,078	703	7793	879
Age (%)			*	*
18-44 y	1	0	1	1
45-64 y	8	0	8	11
65-69 y	25	29	24	21
70-74 y	30	34	29	30
75-79 y	20	24	20	20
80+ y	16	14	16	16
Missing	1	0	1	1
Sex (%)			*	
Male	45	43	45	44
Female	54	57	54	55
Missing	1	0	1	0
Education (%)			*	*
Eight grade or less	36	60	28	61
Some high school	19	14	21	13
High school graduate	23	17	25	11
Some college	12	6	14	7
College graduate	4	2	4	4
More than 4 y of college	3	1	4	1
Missing	3	0	4	4
Self-rated health (%)			*	*
Excellent	8	7	8	15
Very good	17	8	20	10
Good	33	35	34	27
Fair	32	45	28	38
Poor	9	6	9	9
Missing	1	0	1	1

As the matched sample was based on an identical one-to-one match of these characteristics, the Spanish and English matched sample have the same characteristics. Test statistics conducted are for comparison of the nonmatched Spanish and English speakers to the matched sample.

*Statistical significance at the 0.01 significance level.

respectively. Results of the permutation tests described in step 3 (tests of the likelihood of the differences described above observed just by chance) are reported in Table 3 (also see Figure, online only, Supplemental Digital Content 4, available at: http://links.lww.com/MLR/A163 which illustrates an example of the permutation distribution).

TABLE 2. Sample Size, Mean and Standard Deviation (SD) of Items on Patient Satisfaction and Comparison Test

		Spanish			English			
Item	Abbreviated Item Content	N	Mean	SD	N	Mean	SD	Comparison Test
Q19	1. Get advice needed	251	2.32	0.80	309	2.46	0.77	$\chi^2=6.5$*
Q30	2. See provider within 15 min	616	1.81	0.75	604	1.75	0.84	$\chi^2=40.4$†
Q33	3. Provider listens to you	630	2.58	0.63	624	2.70	0.59	$\chi^2=19.6$†
Q34	4. Provider explains things	632	2.53	0.70	628	2.64	0.64	$\chi^2=10.0$†
Q35	5. Provider respects your opinion	632	2.62	0.62	625	2.68	0.60	$\chi^2=5.9$
Q36	6. Provider spent enough time with you	631	2.31	0.69	618	2.53	0.70	$\chi^2=68.0$†
Q31	7. Office staff treat you with respect	631	2.69	0.58	622	2.79	0.53	$\chi^2=16.6$†
Q32	8. Office staff helpful as you thought	632	2.56	0.65	627	2.65	0.63	$\chi^2=13.8$†
Q12	9. Problem getting personal doctor	284	2.86	0.45	463	2.71	0.56	$\chi^2=27.4$†
Q12	9. Problem getting personal doctor	284	2.86	0.45	463	2.71	0.56	$\chi^2=27.4$†
Total score		703	2.49	0.46	703	2.54	0.47	$t_{702}=-2.01$*
Cronbach alpha		0.83			0.85			
CFA fit statistics								
χ^2		51.55 (df = 20, P<0.001)			80.11 (df = 21, P<0.001)			
CFI		0.987			0.978			
TLI		0.992			0.988			
RMSEA		0.047			0.063			

The Item comparison test was obtained contrasting the 3 outcome level choices (1, 2, or 3) to the language using a χ^2 test. The total scores were compared using a t test.

*P value <0.05.

†P value <0.01.

CFA indicates confirmatory factor analysis; CFI, comparative fit index; RMSEA, root mean square error of approximation; TLI, Tucker-Lewis index.

TABLE 3. Anchor Item Identification. *P* Values of Step by Step Permutation Tests

	Round 1 *P*			Round 2 *P*			Round 3 *P*		
Item	**a**	**b_1**	**b_2**	**a**	**b_1**	**b_2**	**a**	**b_1**	**b_2**
1	0.82	0.74	0.44	0.39	0.27	0.82	0.35	0.24	0.80
2	0.84	0.00	0.72						
3	0.36	0.89	0.07	0.12	0.84	0.04			
4	0.50	0.73	0.66	0.85	0.77	0.96	0.77	0.71	0.96
5	0.81	0.14	0.05						
6	0.06	0.03	0.00						
7	0.36	0.99	0.06	0.44	0.13	0.64	0.46	0.12	0.69
8	0.05	0.92	0.43	0.31	0.22	0.96	0.27	0.21	0.90
9	0.44	0.28	0.01						

In the IRT graded response model used, a is the discrimination parameter and b_1 and b_2 are the item difficulty or location parameters.
IRT indicates item response theory.

On the first round of anchor item detection, items 1, 3, 4, 7, and 8 showed no significant DIF in parameters (columns 2 to 4); on the second round, item 3 was also found to potentially have DIF. When this procedure was conducted again (round 3), none of the remaining items (1, 4, 7, and 8) showed DIF. Therefore, these 4 items were retained as anchor items and the remaining 5 items (2, 3, 5, 6, and 9) were evaluated for DIF.

DIF Analysis Among Items 2, 3, 5, 6, and 9

The last 3 columns of Table 4 list the likelihood ratio χ^2 test statistics and significance for the DIF analyses. The omnibus tests indicated that only item 9 is DIF-free at a 0.05 significance level. After adjusting for multiple comparisons, one more item (item 3) can be considered DIF-free. CAHPS item 2 ("see provider within 15 min of an appointment"), item 5 ("provider respects your opinion"), and item 6 ("provider spent enough time with you") do not perform similarly for English and Spanish speakers. For these items, DIF is observed in only the location parameters.

The final parameter estimates and standard errors of all the items, with items 1, 4, 7, and 8 set as anchor items, are presented in columns 4 to 6 of Table 4. The estimates for the items with identified DIF were generated allowing for uniform DIF by specifying separate location parameter estimates for the 2 language groups, while keeping the discrimination parameters the same between groups. For all other items, the parameter estimates for the 2 groups were constrained to be equal using MULTILOG. With the exception of low slopes values of 0.64 and 0.68, for items 2 and 9, respectively, the slopes of the remaining 7 items were all above 1.0, ranging from 1.35 to 3.31, indicating that most of the items have a strong relationship, with the patient evaluations of the provider construct being measured in these CAHPS survey items. Item 5 had the highest discrimination slope and, thus, was the most salient indicator of patients' experiences with providers.

As the CAHPS items are scored with 3 response categories,[1–3] a boundary response function (BRF), a plot which represents each item with 2 curves, one tracing the probability of scoring at or above 2 (ie, usually or always) and the other the probability of scoring 3 (ie, always), was used to display DIF (Fig. 1). Inspection of the BRFs illustrates how the 2 language groups are not using the response categories in the same manner for the DIF items. For example, a Spanish speaker with an evaluation of provider score of -1 on the IRT scale is expected to answer "usually or always" for item 2 about 47% of the time, whereas an English speaker with the same level of evaluation of care will likely give the same answer only 30% of the time. This is equivalent to an expected item score of 1.58 and 1.42 for the Spanish and English speakers, respectively, a difference equal to a standardized effect of size 0.20 at the construct level -1 (Supplemental Digital Content 2, online only, available at: http://links.lww.com/MLR/A161 for the definition of effect size). Even at the average construct score of 0, these percentages are 62% and 45% for the Spanish and English speakers, respectively, or the same standardized effect of size 0.20. For item 6, at the average latent score of 0, the probability of selecting "always" goes from 39% for Spanish to 60% for English speakers or a standardized effect of size -0.27. For this item, at the latent score of -1, this effect size was really small at only 0.03. When it comes to item 5, the effect size was 0.09 and 0.22 at the latent scores of 0 and -1, respectively. Figure 1D also shows the scale score plot, and the expected item score plots for the item with DIF are also presented in Supplemental Digital Content 5 (online only, available at: http://links.lww.com/MLR/A164, a figure that shows the plots).

For qualitative comparison, using the standard IRTLR-DIF (Table, Supplemental Digital Content 6, online only, available at: http://links.lww.com/MLR/A165 which presents these results), item 9 was flagged as the sole anchor item and the DIF analysis itself revealed that only items 1 and 2 showed DIF. Because the permutation method identified more anchor items, it will possibly have higher power for detecting DIF.[29]

DISCUSSION

Our results indicate that the Spanish version of CAHPS is not entirely psychometrically equivalent to the English version. Using a semiparametric method, we found that all the items studied were equivalent in discrimination parameter, indicating that analogous items in the different languages are equally related to the construct of patient evaluations of providers. However, 3 of the 9 items displayed DIF in their location parameters, indicating that responses for these 3 items do not reflect the same degree of experience at the doctor's office for Spanish compared with English speakers.

The patterns in observed DIF were inconsistent. In item 2 (seeing a provider within 15 min of an appointment) and item 6 (provider spending enough time with responders), the pattern of the BRFs revealed that English speakers were more likely to endorse the extreme response options (ie, never/sometimes or always) than Spanish speakers with equal level on the construct. However, in item 5 (provider respecting what the patient says), Spanish speakers were more likely to endorse the higher end of the scale compared

TABLE 4. Estimated Item Parameters and Their Standard Errors From Graded Response Model With Differential Item Functioning Tests (χ^2 Estimate and *P* Value)

			Parameters and Standard Error			DIF Tests: χ^2 and *P*		
	Item	Language	a (se)	b_1 (se)	b_2 (se)	Omnibus DIF: χ^2 (3)	a DIF: χ^2 (1)	b DIF: χ^2 (2)
Anchor	1	Both	1.35 (0.15)	−1.34 (0.17)	−0.21 (0.11)			
	4		2.59 (0.16)	−1.36 (0.08)	−0.46 (0.05)			
	7		2.23 (0.16)	−1.82 (0.12)	−0.9 (0.06)			
	8		2.57 (0.16)	−1.49 (0.09)	−0.44 (0.05)			
DIF free	3	Both	3.22 (0.22)	−1.48 (0.08)	−0.48 (0.04)	11.2 (0.011)	3.4 (0.065)	7.8 (0.020)
	9		0.68 (0.13)	−4.56 (0.86)	−2.3 (0.45)	1.0 (0.801)		
Showed DIF	2	English	0.64 (0.04)	0.31 (0.18)	2.11 (0.21)	98.2 (0.000)	4.9 (0.027)	93.2 (0.000)
		Spanish		−0.8 (0.18)	2.29 (0.23)			
	5	English	3.31 (0.20)	−1.38 (0.09)	−0.45 (0.06)	26.5 (0.000)	0.1 (0.752)	26.3 (0.000)
		Spanish		−1.56 (0.09)	−0.56 (0.05)			
	6	English	2.42 (0.13)	−1.17 (0.09)	−0.17 (0.07)	63.7 (0.000)	0.1 (0.752)	63.5 (0.000)
		Spanish		−1.33 (0.08)	0.18 (0.06)			

In columns 4 to 6, standard errors are in parenthesis and in columns 7 to 9, *P* values are in parenthesis. In the IRT graded response model used, a is the discrimination parameter and b_1 and b_2 are the item difficulty or location parameters.

DIF indicates differential item functioning; IRT, item response theory.

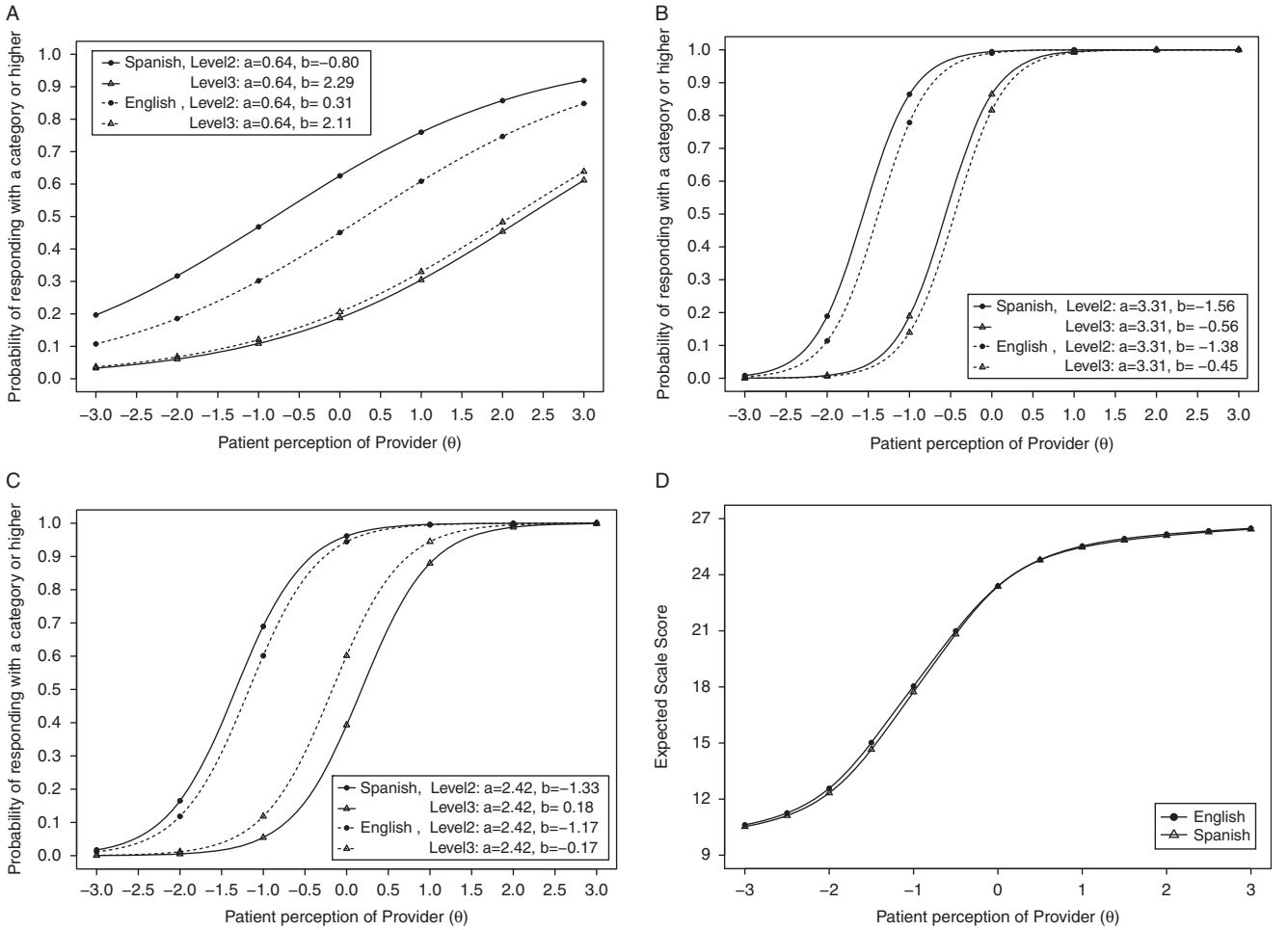

FIGURE 1. Boundary response function curve for each item with DIF and total expected response function curve. A, Item 2: see provider within 15 minutes of an appointment. B, Item 5: provider respects your opinion. C, Item 6: provider spent enough time with you. D, Total expected response function language group comparison.

with English speakers. These differences in response patterns to items 2, 6, and 5 may be culturally driven. Late and hurried doctors are normative in Latin American countries and respect (respeto) is a very important value in Hispanic cultures. Although Latino patients regard physicians as authority figures to be given respect, they also expect respect in return.[33] Spanish speakers are more tolerant of busy and late doctors than English speakers and therefore less likely to endorse extreme response options in items 2 and 6, but less tolerant of disrespect and therefore more likely to endorse the extremes in item 5.

As some studies have suggested that iterative purification method can be less effective at pure anchor detection,[34,35] the same limitations will be applicable to the permutation method. But as demonstrated by Wang and Shih,[25] there is also evidence about the effectiveness of the iterative purification method. For the permutation method we used for the detection of anchor items, it is known to be sometimes conservative,[36] thus it will take more extreme test statistics to reject a null hypothesis, and the actual error rate is much less than the prescribed alpha level.

This study has limitations. The proposed permutation method yielded a larger number of anchor items than the IRTLR method, and identified only 1 item with DIF in common (item 2). It is not possible to evaluate the accuracy of these findings. A more comprehensive simulation study is needed for the evaluation of the performance of the new test. An additional limitation is that the study focused on Spanish speakers, and did not examine other fast-growing groups of native speakers. Also, the results of the sample characteristics comparison suggested that sample of Spanish speakers in CAHPS were different from the general population when it comes to education, health status, gender, and age. This means that the results of these language differences may only generalize to Hispanics, similar to the Spanish speakers from which our sample is drawn. Even among the population sampled, because CAHPS surveys are self-administered, low-literacy plan members might have been excluded,[2] and a greater proportion of such persons might have been Spanish-speaking Hispanics.

In summary, with the rapidly growing population of Spanish speakers, Hispanic or not, making the ethnic makeup of the United States more complex, accurately translated patient surveys from English to Spanish are becoming a necessity. Equivalence in such translated instruments will allow researchers and policy makers to alleviate doubts about disparities in patients' experience with health care providers that are being observed, and that might be attributed to cultural difference in evaluating care received. The results of this study suggest that a few of the CAHPS items display location DIF by language, and the current practice of using these items for disparity inference without controlling for the survey language may result in biased conclusions[37] and thus should be changed. If modification of the translation of items with DIF is possible, it should be made to eliminate DIF in these items.

Although language was the only cultural characteristics we studied as source of DIF in the CAHPS survey items, such differences can also exist in the matching variables (self reported health status, age, gender, and education level) and similar analyses should be conducted on these potential sources of differences as well. Researchers are challenged to promote equivalence in measurements by using the technique employed in this study, to examine tools they have translated for use among diverse population groups.

REFERENCES

1. Hargraves JL, Hays RD, Cleary PD. Psychometric properties of the Consumer Assessment of Health Plans Study (CAHPS) 2.0 adult core survey. *Health Serv Res.* 2003;38(6 pt 1):1509–1527.
2. Morales LS, Elliott MN, Weech-Maldonado R, et al. Differences in CAHPS adult survey reports and ratings by race and ethnicity: an analysis of the National CAHPS benchmarking data 1.0. *Health Serv Res.* 2001;36:595–617.
3. Weech-Maldonado R, Morales LS, Spritzer K, et al. Racial and ethnic differences in parents' assessments of pediatric care in Medicaid managed care. *Health Serv Res.* 2001;36:575–594.
4. Guzmán B. *The Hispanic Population: Census 2000 Brief*. Washington, DC: US Department of Commerce, Economics and Statistics Administration, US Census Bureau; 2001. Brief No. C2KBRO1-3.
5. US Census Bureau. American Community Survey. Selected Social Characteristics in the United States; 2007. Available at: http://factfinder.census.gov/servlet/ADPTable?_bm=y&-geo_id=01000US&-qr_name=ACS_2007_1YR_G00_DP2&-context=adp&-ds_name=ACS_2007_1YR_G00_&-tree_id=306&-_lang=en&-redoLog=false&-format. (Accessed April 15, 2010).
6. Lord FM. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.
7. Marshall GN, Morales LS, Elliott M, et al. Confirmatory factor analysis of the Consumer Assessment of Health Plans Study (CAHPS) 1.0 core survey. *Psychol Assess.* 2001;13:216–229.
8. Bann CM, Iannacchione VG, Sekscenski ES. Evaluating the effect of translation on Spanish speakers' ratings of Medicare. *Health Care Financ Rev.* 2005;26:51–65.
9. Goldstein E, Cleary PD, Langwell KM, et al. Medicare managed care CAHPS: a tool for performance improvement. *Health Care Financ Rev.* 2001;22:101–107.
10. Zaslavsky AM, Zaborski LB, Cleary PD. Plan, geographical, and temporal variation of consumer assessments of ambulatory health care. *Health Serv Res.* 2004;39:1467–1485.
11. Weidmer B, Brown J, Garcia L. Translating the CAHPS 1.0 survey instruments into Spanish. *Med Care.* 1999;37:MS89–MS96.
12. Morales LS, Weidmer BO, Hays RD. "Readability of the CAHPS 2.0 Child and Adult Surveys." In: Cynamon ML, Kulka, RA, eds. Seventh Conference on Health Survey Research Methods: Conference Proceedings; Hyattsville, MD: US Department of Health and Human Services; 2001:83–90. DHHS Publication No. (PHS) 01-1013.
13. Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat.* 1985;39:33–38.
14. McLeod LD, Swygert KA, Thissen D. Factor analysis for items scored in two categories. In: Thissen D, Wainer H, eds. *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates; 2001:185–209.
15. Byrne B. *Structural Equation Modeling With EQS and EQS/Windows: Basic Concepts Applications and Programming*. Thousand Oaks, CA: Sage; 1994.
16. Muthén LK, Muthén BO. *Mplus User's Guide*. 2nd ed. Los Angeles, CA: Muthén & Muthén; 2001.
17. Samejima F. Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika.* 1996;23:17–35.
18. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993:67–113.
19. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
20. Miller TR, Spray JA. Logistic discriminant function analysis for DIF identification of polytomously scored items. *J Educ Meas.* 1993;30:107–122.

21. Orlando M, Marshall GN. Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. *Psychol Assess.* 2002;14:50–59.
22. Thissen D. IRTLRDIF (software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning) Version v. 20b. 2001. Available at: www.unc.edu/∼dthissen/dl.html.
23. Edelen MO, Thissen D, Teresi JA, et al. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the mini-mental state examination. *Med Care.* 2006;44(suppl 3):S134–S142.
24. Woods CM. Empirical selection of anchors for tests of differential item functioning. *Appl Psychol Meas.* 2009;33:42–57.
25. Wang WC, Shih CL. MIMIC methods for assessing differential item functioning in polytomous items. *Appl Psychol Meas.* 2010;34:166–180.
26. Zenisky AL, Hambleton RK, Robin F. Detection of differential item functioning in large-scale state assessments: a study evaluating a two-stage approach. *Educ Psychol Meas.* 2003;63:51–64.
27. Finch H. The MIMIC model as a method for detecting DIF: comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Appl Psychol Meas.* 2005;29:278–295.
28. Finch WH, French BF. Detection of crossing differential item functioning. A comparison of four methods. *Educ Psychol Meas.* 2007; 67:565–582.
29. Wang W. Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *J Exp Educ.* 2004;72:221–261.
30. Shih CL, Wang WC. Differential item functioning detection using the multiple indicators, multiple causes MIMIC method with a pure short anchor. *Appl Psychol Meas.* 2009;33:184–199.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995;57:289–300.
32. du Toit M. *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International; 2003.
33. Lassetter JH, Baldwin JH. Health care barriers for Latino children and provision of culturally competent care. *J Pediatr Nurs.* 2004; 19:184–192.
34. Wang WC, Su YH. Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Appl Psychol Meas.* 2004;28:450–480.
35. French BF, Maller SJ. Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educ Psychol Meas.* 2007;67:373–393.
36. Berger VW. Pros and cons of permutation tests in clinical trials. *Stat Med.* 2000;19:1319–1328.
37. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care.* 2000;38(suppl 9):II28–II42.

