# Evaluating the Effect of Translation on Spanish Speakers' Ratings of Medicare

Carla M. Bann, Ph.D., Vincent G. Iannacchione, M.S., and Edward S. Sekscenski, M.P.H.

*This study examined the equivalence of the English and Spanish versions of the Medicare Consumer Assessment of Health Plans Study (CAHPS®) fee-for-service (FFS) survey among 2,996 Hispanic Medicare beneficiaries. Multigroup confirmatory factor analyses indicated that with few exceptions the factor structures were very similar for the English and Spanish surveys. However, item response theory-based methods for investigating differential item functioning (DIF) revealed that several items demonstrated threshold-related DIF, suggesting that respondents in the two languages utilized the response options for the items differently. The results of this study suggest the need for future qualitative research to understand how respondents comprehend the response options in the two languages.*

## INTRODUCTION

Hispanics have recently become the largest minority group in the U.S. with approximately 37 million people in this country being of Hispanic or Latino origin (U.S. Census Bureau, 2003). The growing number of Hispanics makes it particularly important to understand the quality of health care and services that this population needs and receives. For example, research suggests that Hispanic adults are more likely to be uninsured, less likely to have had a health care visit within the past 2 years, and report being less satisfied with the amount of time spent with their doctors (Doty, 2003).

However, Hispanics may face many obstacles to effectively responding to surveys addressing health care-related issues, such as lower educational achievement (U.S. Census Bureau, 2000a) and less fluency in English (U.S. Census Bureau, 2000b) than non-Hispanic white persons. Furthermore, research suggests that Hispanics, particularly those who speak Spanish, have low levels of health literacy, making it even more difficult for them to respond to questions on health-related issues (Gazmararian et al., 1999).

To increase the probability of obtaining accurate and complete health-related survey information from Hispanic respondents, it is important to remove any possible language barriers by providing a Spanish language version of the survey. Both backward and forward translations should be included when translating items (Cull et al., 1998). In addition, once the Spanish translation has been developed and data have been collected, it is important to analyze responses to the items to ensure their equivalence across the two languages. Relying solely on the translation techniques and failing to quantitatively compare the translations could cause researchers to miss differences that occur when individuals are actually administered the items.

The CAHPS® surveys are one of the leading measures of health plan performance. The National Committee for

Carla M. Bann and Vincent G. Iannacchione are with RTI International. Edward S. Sekscenski is with the Centers for Medicare & Medicaid Services (CMS). The research in this article was supported by CMS under Contract Number 500-95-0061 (TO#7). The statements expressed in this article are those of the authors and do not necessarily reflect the views or policies of RTI International or CMS.

Quality Assurance (NCQA) uses CAHPS® in its accreditation program and the National CAHPS® Benchmarking Database includes information from more than 360,000 individuals enrolled in Medicare, Medicaid, and private insurance plans (Agency for Healthcare Research and Quality, 2003). CMS use the CAHPS® surveys to assess the performance of all Medicare managed care health plans as well as the traditional Medicare FFS plan, in all 50 States, Washington, DC, and Puerto Rico. The Medicare CAHPS® information is disseminated to beneficiaries through the CMS Web site (http://www.cms.gov), the *Medicare & You* handbook, and the 1-800-MEDICARE toll-free telephone line.

The original set of CAHPS® core survey items were translated into Spanish, using both forward and backward translation, and then evaluated through cognitive interviewing techniques. Although a field test of the CAHPS® items was conducted, the number of Spanish surveys was too small to make comparisons to the English surveys (Weidmer, Brown, and Garcia, 1999). Research using CAHPS® in the Medicaid managed care population suggests that Spanish-speaking Hispanics report worse health care experiences than English-speaking Hispanics (Weech-Maldonado et al., 2003), however, it is unclear whether these results represent true differences in experiences or merely differences in interpretation of the items or response options.

To make this distinction, research is needed to assess the equivalence of the English and Spanish version of CAHPS®. Although Marshall and colleagues (2001) used confirmatory factor analysis to test factor invariance of CAHPS® among Latinos and non-Latinos with Medicaid or commercial insurance, to our knowledge, no research has been conducted to establish the factor invariance of different language translations of CAHPS®. The current study conducts a comprehensive evaluation of the equivalence of the Spanish and English versions of the Medicare CAHPS® FFS survey, using information from 2 years of large-scale data collection. Our approach includes confirmatory factor analysis and item response theory techniques for testing possible differential item functioning across the two versions of the survey.

## DIF

DIF analysis can be used to identify items that do not function equivalently across two groups. Investigating possible DIF is an important step in evaluating the psychometric properties of a scale because the presence of DIF reduces the scale's validity. DIF analysis is especially valuable for evaluating the equivalence of items translated into another language. If an item exhibits language-related DIF, it may not be measuring the same concepts for the two translations of the item, making it inappropriate to combine or compare the results across languages. The presence of DIF could signal problems with the translation of the item, the need to use different terms or phrases, or possible cultural differences in the way that respondents interpret the concepts measured by the item.

In classical test theory, cultural equivalence is generally assessed by comparing statistics such as item means, item-total correlations, and Cronbach's (1951) alphas across the various translations. However, the results from these statistics can sometimes be misleading because they are not invariant across samples. For example, differences in item means may simply reflect differences between the groups on the construct being measured rather than differences in how the items are functioning.

Two alternative methods overcome the limitations of classical test theory approaches: confirmatory factor analysis and item response theory (IRT). A detailed comparison of these two approaches is provided by Raju, Laffitte, and Byrne (2002). Briefly, confirmatory factor analysis is based in the structural equation modeling framework and allows us to test whether the factor structure of a scale is invariant across groups. However, as explained by Cooke and colleagues (2001), while factorial invariance suggests that the items are measuring the same construct, it does not necessarily provide evidence that the construct is being measured on the same metric, a necessary condition for equivalence.

While IRT also allows us to investigate the relationship between the items and the underlying construct, it has several additional features that make it particularly useful for investigating DIF. For example, IRT allows us to explore differences in functioning at several levels, including the scale, item, and response option level. In addition, IRT produces item parameter estimates that are less likely to be biased when used with unrepresentative samples than classical test theory estimates (Embretson and Reise, 2000) and both the item and person parameters in IRT are placed on the same scale, making them comparable. Finally, features of IRT make it particularly suitable for investigating and presenting DIF graphically.

## METHOD

### Sample

This study utilized data from the 2000 and 2001 administrations of the Medicare CAHPS® FFS survey. To be eligible for the survey, beneficiaries must have met the following criteria: (1) lived in the U.S. or Puerto Rico, (2) been enrolled in Medicare FFS continuously for the prior 6 months, (3) were at least 18 years old, (4) were not currently in a hospice program, and (5) were not enrolled in a group health plan.

As a part of the survey administration, participants in the U.S. received a pre-notification postcard instructing them on how to request a Spanish survey. If they did not request a Spanish survey, they were automatically sent the English version of the survey. Because of the large number of Spanish speakers in Puerto Rico, beneficiaries residing there automatically received the Spanish survey unless they requested an English survey by calling or completing the prenotification postcard. Bilingual interviewers were also available to administer the Spanish survey over the telephone if needed.

As would be expected, a much larger proportion of respondents completed the English survey than the Spanish survey. Across both years, a total of 219,037 respondents completed the English survey while only 2,350 completed the Spanish survey. This large sample size discrepancy could influence the results if we were simply to compare the two groups. Furthermore, individuals who responded to the two surveys may have very different backgrounds and experiences which could influence their responses to the items.

To eliminate possible demographic differences, we included only respondents who indicated that they were of Hispanic origin and did not receive help from a proxy with reading, answering, or translating the questions. Furthermore, because past research has found differences on the CAHPS® FFS survey according to mode of administration (Pugh et al., 2002), we restricted the sample to only participants who responded to the survey through the mail. This resulted in a total of 4,203 English survey respondents and 1,498

Spanish survey respondents who were of Hispanic origin, responded to the survey through the mail, and did not have a proxy responding for them.

We then selected a random sample of 1,498 of these English survey respondents who had demographic characteristics matching those completing the Spanish survey. The following characteristics were used to select the respondents because prior research suggests that they are related to ratings on CAHPS® items: (1) sex, (2) age, (3) education, (4) having a personal doctor, and (5) health status (Uhrig et al., 2002). In addition, many of these variables are included in the case-mix adjustment models used to create the reporting composite scores (Elliott et al., 2001). Table 1 displays the demographic profiles of the English and Spanish respondents included in this study.

## Measures

The Medicare CAHPS® FFS survey is a national survey that has been conducted annually among beneficiaries enrolled in the original Medicare Program since fall 2000. The goal of the survey is to collect and report information on Medicare beneficiaries' experiences with receiving care through the Medicare Program. Questions on the survey items measure beneficiaries' perceptions of physician communication, getting care that is needed, getting care quickly, and ease in seeing a specialist.

## Psychometric Analysis Methods

This study used three different approaches to assessing DIF across the Spanish and English Medicare CAHPS® surveys: comparison of item-level descriptive statistics, multigroup confirmatory factor analyses, and IRT-based DIF analyses. With respect to missing data, other researchers (Marshall et al., 2001) analyzing the CAHPS® items have imputed missing values using methods such as hot-deck imputation. However, the large number of skip patterns in the CAHPS® surveys would require the imputation of a large number of missing values. To avoid possible biases due to imputation of missing data, we chose to use analysis approaches (i.e., full information maximum likelihood method in Mplus® software program [Muthén and Muthén, 2003]) that allowed us to use all available data from a respondent without imputing values[1].

For this study, we analyzed the CAHPS® items used to compute the CAHPS® reporting composites. The CAHPS® ratings items were not included in the analyses because the two types of items appear to be capturing different information. The ratings items are designed to obtain a global perspective of satisfaction with health care while the reports items ask about specific experiences. Furthermore, prior research suggests that Spanish speakers report worse experiences on the reports items, but provide higher scores for the ratings items than English speakers (Weech-Maldonado et al., 2003).

## Item-Level Descriptive Statistics

Item-level descriptive statistics, including means and standard deviations, were computed separately for the two versions of the survey. *T*-tests were used to evaluate the significance of the mean differences.

## Confirmatory Factor Analyses

Based on prior analyses of the CAHPS® FFS items, we expected the reports items to group into two factors representing

[1] The three items comprising CAHPS® customer service (Items 41, 43, and 45 on the CAHPS® FFS survey) were excluded from the analyses because they are relevant to few participants and therefore, had very large missing percentages.

**Table 1**

**Demographic Characteristics of Hispanic Respondents Completing the Medicare CAHPS® FFS Survey, by Language**

| Characteristic | English | Spanish |
|---|---|---|
| | Percent | |
| **Sex** | | |
| Male | 45.7 | 45.4 |
| Female | 54.3 | 54.6 |
| **Age** | | |
| Under 65 Years | 20.6 | 20.4 |
| 65-69 Years | 24.3 | 25.3 |
| 70-74 Years | 24.2 | 23.4 |
| 75-79 Years | 16.4 | 16.6 |
| 80 Years or Over | 14.6 | 14.3 |
| **Education** | | |
| Less than High School Diploma | 59.8 | 58.1 |
| High School Graduate | 17.4 | 18.4 |
| College | 22.8 | 23.5 |
| **Have Personal Doctor** | | |
| Yes | 84.1 | 86.0 |
| No | 16.0 | 14.0 |
| **Health Status** | | |
| Excellent/Very good | 11.7 | 12.3 |
| Good | 20.8 | 20.0 |
| Fair/Poor | 67.5 | 67.8 |

NOTE: *N*=1,498.

SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey (FFS), 2000 and 2001.

satisfaction with provider (Items 14, 16, 18, and 23-29) and access to care (Items 4, 9, 21, and 22). To assess factorial invariance, we computed a series of multigroup confirmatory factor analyses (CFAs). The first set of CFAs allowed all of the loadings to vary across the two languages. We then constrained all of the loadings to be equal across the groups and began freeing one loading at a time and computing the change in model chi-square to determine if the individual factor loadings varied significantly across the groups. The CFA analyses were computed using the MPlus® software program.

## IRT DIF Analyses

We provide a brief description of IRT here for readers not familiar with this technique. IRT uses a statistical model to describe the relationship between an individual's response to an item and the underlying construct (e.g., satisfaction with health plan). For example, Samejima's (1969) graded model is appropriate for items containing ordinal response options (e.g., never, sometimes, usually, always), such as the CAHPS® items. In the graded model, two types of parameters are estimated for each item. The first parameter is the slope or a parameter which quantifies how related the item is to the construct being measured by the scale. In addition to the slope, a set of threshold or b parameters are estimated. The thresholds locate each response option along the continuum of the underlying construct. In other words, the thresholds for the CAHPS® items would indicate the approximate level of satisfaction individuals would need to have before they would endorse

the corresponding response option. The number of thresholds estimated is equal to the number of response options minus one. For example, two threshold parameters would be estimated for the CAHPS® items that contain three response options (e.g., not a problem, small problem, big problem).

Item response theory analyses may be used to uncover differential item functioning with respect to the slope or threshold parameters of an item. Threshold-related DIF indicates that the two groups differ in how they interpret and use the response options. The presence of this type of DIF suggests that the results for the scale should simply be reported separately for the two groups. However, the other type of DIF, slope-related DIF, indicates that the item is differentially related to the underlying construct for the two groups. Slope-related DIF is more detrimental than threshold-related DIF and indicates that the item should ideally not be used to compare the two groups.

The IRT-based DIF analyses in the current study used a method described by Thissen, Steinberg, and Wainer (1993). This method uses likelihood ratio comparisons and has the advantage of allowing us to conduct hypothesis tests to evaluate differences in particular characteristics of the item rather than simply in the item score. Several researchers recommend this approach and suggest that it is a more powerful technique than other available methods (Teresi, Kleinman, and Ocepek-Welikson, 2000; Wainer, 1995). In addition, it has been shown to be effective with very small numbers of items (Thissen, Steinberg, and Wainer, 1988).

As a part of the DIF analyses, a set of items, called anchor items, are used to equate the two groups and to establish an individual's level on the underlying continuum. To our knowledge, possible language-related DIF has not previously been investigated on the CAHPS® FFS items using IRT, so there was no pre-existing information available for designating specific items as anchors. Therefore, we utilized a technique available in the IRTLRDIF program in which all items except the item of interest are used as anchor items (Thissen, 2001).

To test for DIF, a graded IRT model was estimated in which the parameter estimates for the item of interest and the anchor items were constrained to be equal across the two languages. Next, only the threshold (b) parameter estimates for the item under study were allowed to vary between the two groups. Finally, both the threshold (b) and slope (a) parameters were allowed to vary while the parameters for the anchor items were still contained to be equal for the two groups. To test for threshold-related DIF, the fit of the model which allowed the thresholds of the item to vary between groups was compared with the fully constrained model. Slope-related DIF was evaluated by comparing the fit of the model with both the slope and threshold parameters free to the model with only the threshold parameters free. Models were compared by subtracting the value of negative twice the log-likelihood for each model; this value is distributed as a chi-square statistic with degrees of freedom equal to the number of additional parameters estimated by the less-constrained model.

## Estimating the Effect of DIF

To estimate the effect of DIF among the CAHPS® items that demonstrated DIF, we first computed IRT scores for English and Spanish respondents assuming no DIF among the items. This model constrained the parameters for the items to be equal across the English and Spanish surveys.

Next, we computed IRT scores allowing DIF for all of the items that demonstrated DIF. In each of these models, the parameters for the DIF items were allowed to vary across the two languages while the parameters for all other items were constrained to be equal. We then computed the standardized mean difference in scores (i.e., effect size) between English and Spanish respondents for the models.

## RESULTS

### Item-Level Descriptive Statistics

Table 2 displays the means and standard deviations for each item, separately by language. If the items are equivalent, we would expect a similar pattern of means for the two groups. As shown in the table, the ordering of the means generally seems to be consistent across the two languages. One exception is the item concerning waiting in the doctor's office (Item 23). This item had one of the smallest means for the English respondents, but had a mid-range value for the Spanish respondents.

### Factor Analyses

First, we fit CFAs of the items for the two languages separately. The models fit both languages well: English ($\chi^2(2)=5.26$, $p=0.07$, Comparative Fit Index (CFI)=0.997, Tucker-Lewis Index (TLI)=0.991, Root Mean Square Error of Approximation (RMSEA)=0.033) and Spanish ($\chi^2(2)=2.66$, $p=0.26$, CFI=0.999, TLI=0.998, RMSEA= 0.015). Next, we fit a model that constrained all of the loadings to be equal across the two groups. For each item, we then compared the model chi-square for the fully constrained model to a model in which the loadings for that item were free to vary across the two groups. These results suggest that the following three items have factor loadings that vary across the two groups: Item 23 (waiting more than 15 minutes; $\chi^2(1)= 8.59$, $p< 0.05$), Item 25 (office staff were helpful; $\chi^2(1) = 11.66$, $p< 0.05$), and Item 29 (doctors spent enough time; $\chi^2(1) = 18.66$, $p< 0.05$). Reviewing the loadings suggests that whether doctors spend enough time and waiting longer than 15 minutes are more related to satisfaction with providers among English speakers while whether the office staff were helpful was more related to satisfaction among Spanish speakers. Perhaps time concerns are more salient to English speakers than Spanish speakers.

### IRT Analyses

An assumption of IRT is that the items form a single underlying construct. Therefore, to conduct IRT DIF analyses, we considered the factors included in the prior CFAs to be separate scales containing only the items that loaded on the corresponding factor. Table 3 presents the final parameter estimates for the items. In cases where an item demonstrated threshold-related DIF, the final parameter estimates were computed using a model where the thresholds were allowed to vary between the two languages. When the item demonstrated slope-related DIF, the parameter estimates were obtained from a model in which both the slope and threshold parameters varied between the two languages.

All items except Items 21 (problem getting necessary care) and 22 (delays while waiting for approval) demonstrated some form of DIF with most having threshold-related DIF. The presence of threshold-related DIF suggests that respondents in the two languages used the response options for the items differently. For example, as shown in Table 3, the threshold parameters for Item 4 (problem getting

**Table 2**

**Means of Items on the Medicare CAHPS® Fee-For-Service (FFS) Survey, by Language and CAHPS® Reporting Composite**

| Reporting Composite/Item | English | Spanish | Significance of *t*-test |
|---|---|---|---|
| **Needed Care Composite** | | | |
| 4. Problem getting personal doctor happy with | 2.7 (0.7) | 2.9 (0.4) | *** |
| 9. Problem seeing specialist | 2.6 (0.7) | 2.6 (0.7) | NS |
| 21. Problem getting necessary care | 2.8 (0.5) | 2.8 (0.5) | NS |
| 22. Delays while waiting for approval | 2.9 (0.4) | 2.9 (0.3) | NS |
| **Good Communication Composite** | | | |
| 26. Doctors listened carefully | 3.7 (0.6) | 3.6 (0.6) | ** |
| 27. Doctors explained things | 3.6 (0.7) | 3.5 (0.7) | ** |
| 28. Doctors showed respect | 3.7 (0.6) | 3.6 (0.6) | * |
| 29. Doctors spent enough time | 3.5 (0.7) | 3.3 (0.7) | *** |
| **Getting Care Quickly Composite** | | | |
| 14. Received help when calling during office hours | 3.6 (0.7) | 3.4 (0.7) | ** |
| 16. Got appointment as soon as wanted | 3.5 (0.8) | 3.4 (0.6) | NS |
| 18. Received care for illness or injury | 3.6 (0.8) | 3.3 (0.7) | *** |
| 23. Waited in doctor's office more than 15 minutes | 2.4 (1.0) | 3.0 (0.9) | *** |
| **Respectful Treatment Composite** | | | |
| 24. Were treated with respect by office staff | 3.8 (0.6) | 3.7 (0.6) | *** |
| 25. Office staff were helpful | 3.6 (0.7) | 3.5 (0.6) | *** |
| **Medicare Customer Service Composite** | | | |
| 41. Problem with paperwork | 2.3 (0.8) | 2.6 (0.7) | ** |
| 43. Problem understanding written materials | 2.5 (0.7) | 2.8 (0.5) | *** |
| 45. Problem getting help from customer service | 2.4 (0.8) | 2.6 (0.7) | NS |

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

NOTES: For clarity, only item numbers from the 2000 Medicare CAHPS® FFS survey are presented. NS is non-significant. Standard deviations are shown in parentheses.

SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

personal doctor) were lower for the English respondents than the Spanish respondents, indicating that English respondents required a lower level of satisfaction before endorsing the same response option. In other words, among English and Spanish respondents who had the same overall satisfaction (as measured by the anchor items), the English respondents were less likely to report having difficulty finding a personal doctor they were happy with. Several items also had evidence of slope-related DIF. This type of DIF is a more serious problem than threshold-related DIF and indicates that the items are related to the underlying construct differently in the two language versions, possibly suggesting that the items are actually measuring two different constructs. For example, Item 16 (got appointment as soon as wanted) is more related to experiences with providers among English respondents (slope = 1.46) than Spanish respondents (slope = 0.98). Perhaps English-speaking respondents regard this concept as a more important aspect of care from providers than do Spanish-speaking respondents.

## Effect of DIF

To illustrate the effect of DIF across the two languages, Figures 1-6 present the item response functions for several of the items demonstrating DIF. For each of the graphs, the solid lines represent English and the dotted lines represent Spanish. An

**Table 3**

**Item Response Theory Parameter Estimates for Provider and Access to Care Items on the Medicare CAHPS® Fee-for-Service (FFS) Survey**

| Item | Slope a | Thresholds b1 | b2 | b3 |
|---|---|---|---|---|
| **Provider** | | | | |
| **14. Received help when calling during office hour**[1] | | | | |
| English | 0.89 | 1.00 | 1.17 | 1.97 |
| Spanish | 0.60 | 0.32 | 0.69 | 1.49 |
| **16. Got appointment as soon as wanted**[1] | | | | |
| English | 1.46 | -0.65 | -0.46 | 0.70 |
| Spanish | 0.98 | -0.68 | -0.20 | 0.69 |
| **18. Received care for illness or injury**[1] | | | | |
| English | 0.83 | 1.10 | 1.20 | 2.49 |
| Spanish | 0.54 | 1.35 | 1.62 | 2.32 |
| **23. Waited in doctor's office more than 15 minutes**[2] | | | | |
| English | 1.51 | 0.17 | 1.39 | 2.51 |
| Spanish | 1.51 | -0.04 | 0.43 | 1.87 |
| **24. Were treated with respect by office staff**[1] | | | | |
| English | 5.16 | -0.42 | -0.34 | 0.16 |
| Spanish | 4.38 | -0.48 | -0.28 | 0.06 |
| **25. Office staff were helpful**[1] | | | | |
| English | 6.67 | -0.44 | -0.25 | 0.40 |
| Spanish | 4.53 | -0.48 | -0.18 | 0.32 |
| **26. Doctors listened carefully**[1] | | | | |
| English | 10.02 | -0.46 | -0.28 | 0.26 |
| Spanish | 7.06 | -0.48 | -0.18 | 0.26 |
| **27. Doctors explained things**[1] | | | | |
| English | 7.21 | -0.40 | -0.22 | 0.29 |
| Spanish | 6.03 | -0.47 | -0.13 | 0.34 |
| **28. Doctors showed respect**[2] | | | | |
| English | 9.48 | -0.43 | -0.26 | 0.21 |
| Spanish | 9.48 | -0.47 | -0.19 | 0.26 |
| **29. Doctors spent enough time**[2] | | | | |
| English | 5.19 | -0.45 | -0.26 | 0.58 |
| Spanish | 5.19 | -0.44 | -0.05 | 0.43 |
| **Access to Care** | | | | |
| **4. Problem getting personal doctor**[2] | | | | |
| English | 0.5 | -1.13 | -0.84 | — |
| Spanish | 0.57 | -0.26 | 0.27 | — |
| **9. Problem seeing specialist**[3] | | | | |
| English | 0.98 | 0.16 | 0.53 | — |
| Spanish | 1.25 | 0.22 | 0.48 | — |

[1] Demonstrated slope and threshold differential item functioning (DIF) ($p<0.05$).

[2] Demonstrated threshold DIF only ($p<0.05$).

[3] Demonstrated slope DIF only ($p<0.05$).

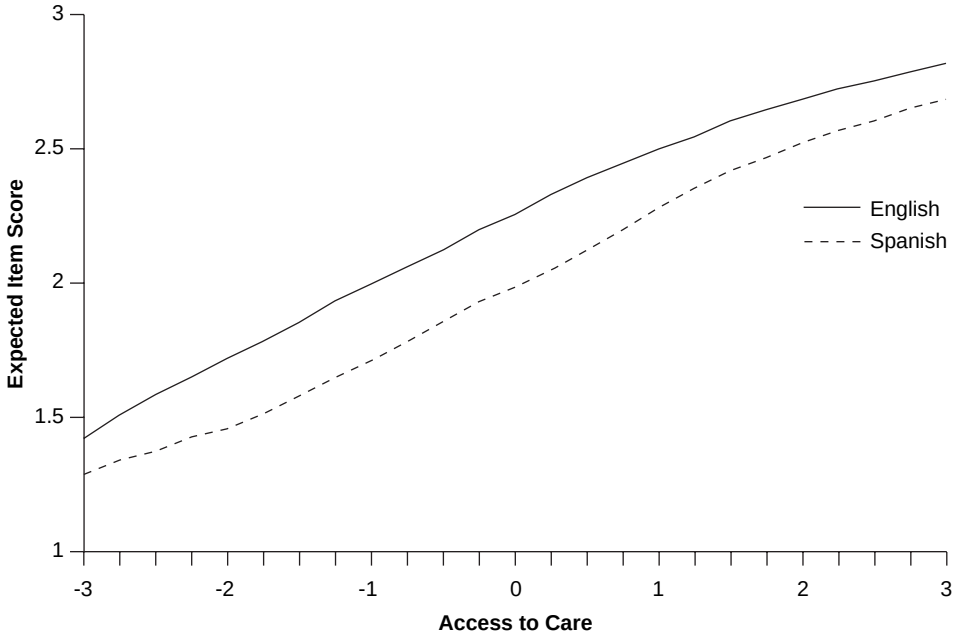
SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

item response function shows the expected item scores for each of the levels of the underlying construct. For Items 14, 16, 18, and 23, the construct is satisfaction with health care provider while the underlying construct for Items 4 and 9 is access to care. When interpreting the curves for items 4 (problem getting personal doctor), 9 (problem seeing specialist) and 23 (waited more than 15 minutes), it is important to note that the items are reverse coded so that high scores represent greater satisfaction (i.e., fewer problems).

If the curves for the two languages are very close or identical, English and Spanish survey participants are expected to receive the same scores, suggesting little or no effect of DIF. For example,

## Figure 1

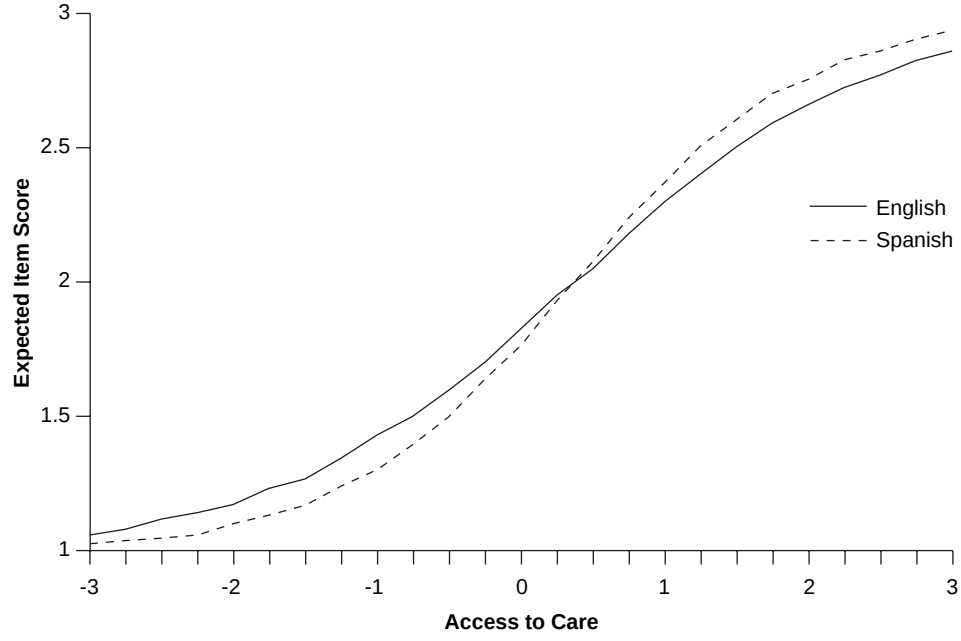
**Item Characteristic Curve for CAHPS® Item 4 (Problem Getting Personal Doctor) Demonstrating Differential Item Functioning, by Language**




SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

## Figure 2

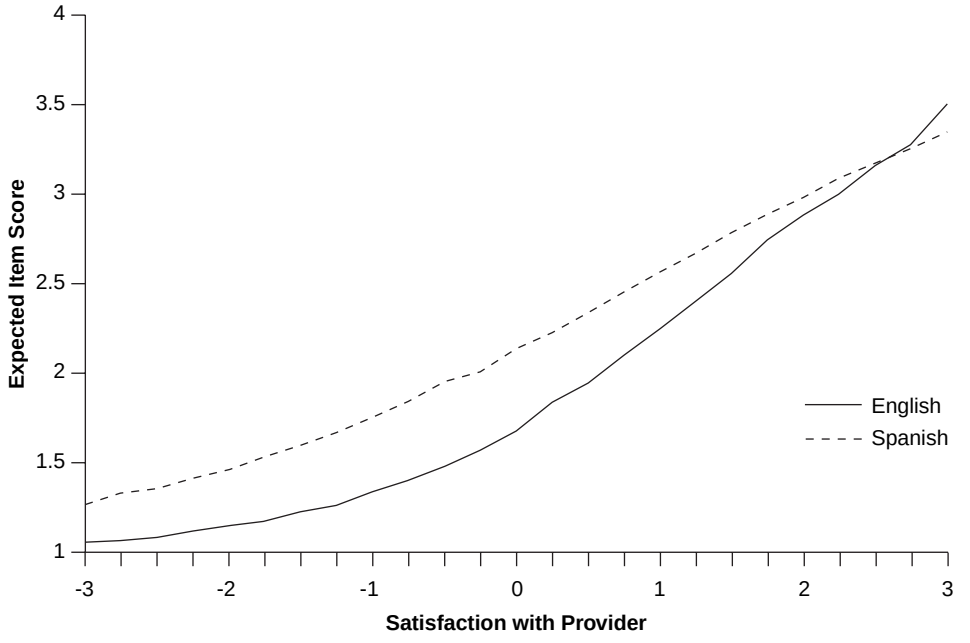
**Item Characteristic Curve for CAHPS® Item 9 (Problem Seeing Specialist) Demonstrating Differential Item Functioning, by Language**




SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

**Figure 3**

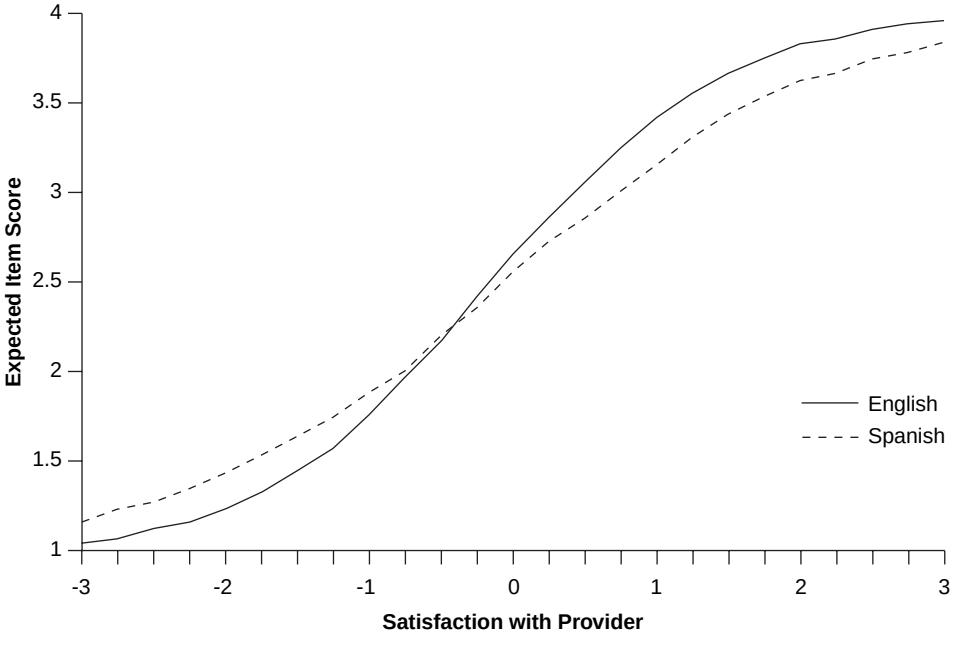
**Item Characteristic Curve for CAHPS® Item 14 (Received Help When Calling During Office Hours) Demonstrating Differential Item Functioning, by Language**




SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

**Figure 4**

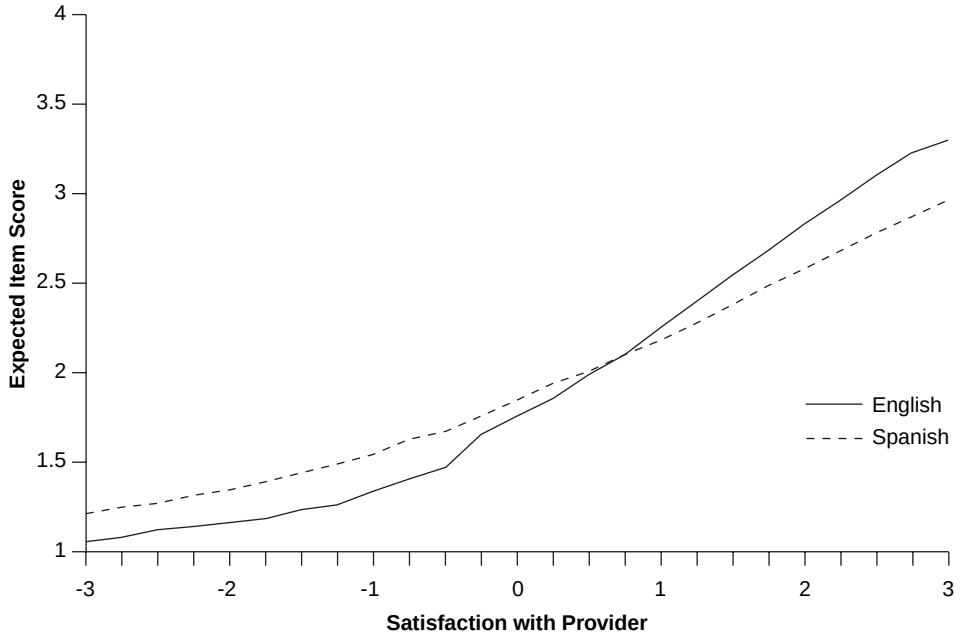
**Item Characteristic Curve for CAHPS® Item 16 (Got Appointment As Soon As Wanted) Demonstrating Differential Item Functioning, by Language**




SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

## Figure 5

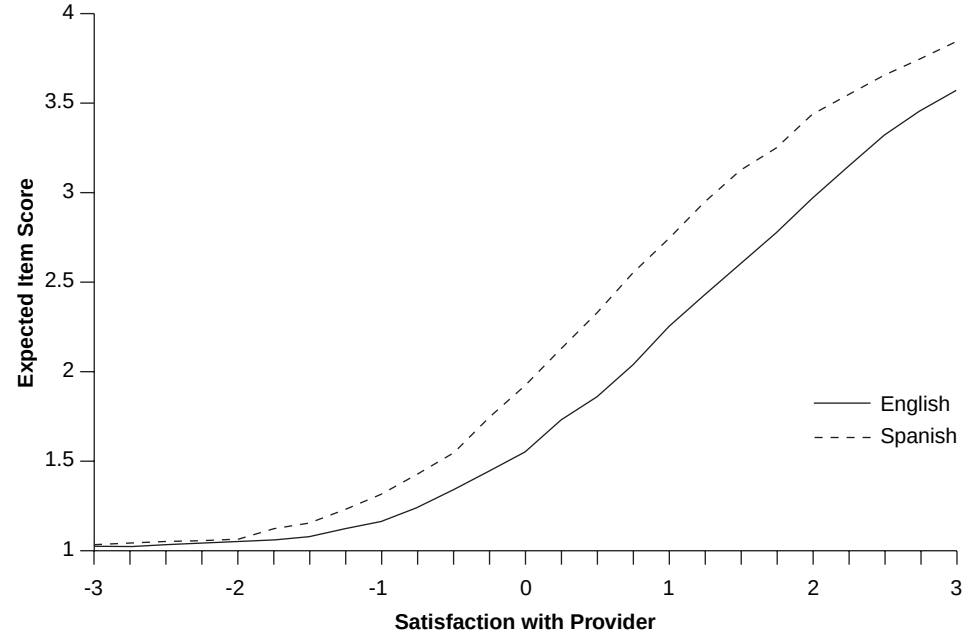
**Item Characteristic Curve for CAHPS® Item 18 (Received Care for Illness or Injury) Demonstrating Differential Item Functioning, by Language**




SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

## Figure 6

**Item Characteristic Curve for CAHPS® Item 23 (Waited in Doctor's Office More Than 15 Minutes) Demonstrating Differential Item Functioning, by Language**




SOURCE: Centers for Medicare & Medicaid Services, Medicare CAHPS® Fee-for-Service Survey, 2000 and 2001.

although not presented due to space constraints, the item response functions for the English and Spanish respondents for Items 24-29 are very similar, suggesting that DIF has little effect for items related to treatment by doctors and office staff. With respect to access to care, it appears that Spanish speakers were more likely to report problems getting a personal doctor. Regarding satisfaction with provider, Spanish speakers were consistently more likely to report receiving help when calling during office hours and less likely to report waiting more than 15 minutes in the doctor's office. The patterns for Items 16 and 18 differed across the continuum with Spanish speakers who were more satisfied with their providers being less likely to report getting an appointment as soon as they wanted or receiving care for an illness or injury; the opposite is true for those at the lower ends of the continuum. As another measure of the effect of DIF, we computed standardized mean differences between English and Spanish speakers assuming no DIF and then accounting for DIF for the items that demonstrated DIF in the IRT analyses. The standardized mean difference was computed by subtracting the mean IRT scores for Spanish respondents from the mean for English respondents and dividing by the pooled standard deviation. Therefore, positive values indicate that English respondents had higher scores while negative values indicate lower scores for English respondents. For satisfaction with provider, the effect size assuming no DIF was 0.13, however, after accounting for DIF, the effect size increased to 0.37. The effect size for access to care assuming no DIF was large (-0.95), but decreased when adjusting for DIF (-0.14).

## CONCLUSION AND DISCUSSION

In summary, various psychometric analyses were utilized to investigate the equivalence of the Spanish and English versions of the Medicare CAHPS® FFS report items. Overall, the results of this study support the equivalence of the Spanish and English versions of the CAHPS® FFS survey among the Medicare population. The ordering of item means and the factor structures for the two languages were very similar. Although the IRT DIF analyses revealed that almost all items demonstrated differences in how participants in the two languages used the response options (i.e., threshold-related DIF), the graphs of the item response functions for most items, particularly those related to treatment by providers and office staff, suggest very little difference in expected item scores across the two languages.

The results suggest a few items that could perhaps be revised to minimize the effect of DIF. In particular, most of the items that differed across the two languages related to time issues. Some examples are Items 16 (got appointment as soon as wanted), 23 (waiting more than 15 minutes), and 29 (doctors spent enough time) which seem to be more related to satisfaction among English-speaking respondents than Spanish-speaking respondents. Perhaps these results simply reflect cultural differences in the salience of time factors with respect to receiving health care.

The findings of this study suggest that in most cases when the CAHPS® items exhibiting DIF are combined with other items to compute an overall score (e.g., CAHPS® composite scores), the presence of DIF should not have a substantial effect. However, DIF can have a larger effect when

making comparisons at the individual item level. For example, a particular geographic region with a high proportion of Spanish-speaking beneficiaries could receive different scores on an item simply due to the tendency of Spanish speakers to give different ratings than English speakers. However, it is important to note that very few Spanish language CAHPS® surveys are collected relative to the English language surveys (approximately 1 percent) and therefore, the DIF found here should have very little effect on national estimates.

Future research could be conducted to try to uncover the source of the differences in how English and Spanish survey respondents use the items demonstrating DIF. For example, indepth cognitive interviewing could be used to gain qualitative information about how individuals in the two languages interpret the items, particularly the response options, and whether the concepts addressed by items are consistent in the two languages. In particular, it may be informative to include bilingual individuals in these interviews to obtain their insights on whether there appear to be differences in the interpretations of the response options used in the two versions of the questions.

Finally, a limitation of this study is that Hispanic respondents were treated as a homogeneous group. In fact, Hispanics from different countries of origin (e.g., Puerto Rico, Mexico) may have very different health-related experiences which could potentially lead to differences in their survey responses. For example, Doty (2003) found that Hispanics of Puerto Rican origin were significantly more likely to be satisfied with the quality of their health care than those of Mexican or Central American origin. Future research should examine the CAHPS® survey responses for these groups separately.

## REFERENCE


Agency for Healthcare Research and Quality: *CAHPS® and the National CAHPS® Benchmarking Database: Fact Sheet*. AHRQ Publication Number 03-P001. Rockville, MD. 2003.

Cooke, D.J., Kosson, D.S., and Michie, C.: Psychopathy and Ethnicity: Structural, Item, and Test Generalizability of the Psychopathy Checklist-Revised (PCL-R) in Caucasian and African American Participants. *Psychological Assessment* 13(4):531-542, 2001.

Cronbach, L.J.: Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16(3):297-334, 1951.

Cull, A., Spangers, M., Bjordal, K., et al.: *On Behalf of the EORTC Quality of Life Study Group*: *Translation Procedure*. EORTC Monograph. 1998.

Doty, M.M.: *Hispanic Patients' Double Burden: Lack of Health Insurance and Limited English*. The Commonwealth Fund. The Commonwealth Fund Publication Number 592. 2003. Internet address: http://www.cmwf.org/publications/publications_show.htm?doc_id=221326 (Accessed 2005.)

Elliott, M.N., Swartz, R., Adams, J., et al.: Case-Mix Adjustment of the National CAHPS® Benchmarking Data 1.0: A Violation of Model Assumptions? *Health Services Research* 36(3):555-573, 2001.

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

Gazmararian, J.A., Baker, D.W., Williams, M.V., et al.: Health Literacy Among Medicare Enrollees in a Managed Care Organization. *Journal of the American Medical Association* 281(6):545-551, 1999.

Marshall, G.N., Morales, L.S., Elliott, M., et al.: Confirmatory Factor Analysis of the Consumer Assessment of Health Plans Study (CAHPS®) 1.0 Core Survey. *Psychological Assessment* 13(2):216-229, 2001.

Muthén, L.K. and Muthén, B.O.: *Mplus User's Guide* (2nd Ed.). Muthén & Muthén. Los Angeles, CA. 2003.

Pugh, N., Iannacchione, V., Lance, T., et al.: Evaluating Mode Effects in the Medicare CAHPS® Fee-for-Service Survey. *Proceedings of the American Statistical Association, Section on Survey Research Methods*. New York, NY. August 2002.

Raju, N.S., Laffitte, L.J., and Byrne, B.M.: Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory. *Journal of Applied Psychology* 87(3):517-529, 2002.

Samejima, F.: Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement* 4, Part 2, Whole #17, 1969.

Teresi, J., Kleinman, M., and Ocepek-Welikson, K.: Modern Psychometric Methods for Detection of Differential Item Functioning: Application to Cognitive Assessment Measures. *Statistics in Medicine* 19(11-12):1651-1683, 2000.

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

Thissen, D., Steinberg, L., and Wainer, H.: Use of Item Response Theory in a Study of Group Differences in Trace Lines. In Wainer, H., and Braun, H. (eds.): *Test Validity*. Lawrence Erlbaum Associates. Hillsdale, NJ. 1988.

Thissen, D., Steinberg, L., and Wainer, H.: Detection of Differential Item Functioning Using the Parameters of Item Response Models. In Holland, P.W., and Wainer, H. (eds.): *Differential Item Functioning*. Lawrence Erlbaum Associates. Hillsdale, NJ. 1993.

Uhrig, J.D., Bernard, S., Tornatore, D., et al.: *The Performance of Fee-for-Service Medicare on CAHPS® Measures: An Analysis of Beneficiary Subgroups*. Presented at the Eighth Annual CAHPS® User Group Meeting in Nashville, TN. 2002.

U.S. Census Bureau: *Age by Language Spoken at Home by Ability to Speak English for the Population 5 Years and Older (Hispanic or Latino)*. 2000a. Internet address: http://factfinder.census.gov. (Accessed 2005.)

U.S. Census Bureau: *The Hispanic Population in the United States*. 2000b. Internet address: http://www.census.gov/prod/2001pubs/p20-535.pdf. (Accessed 2005.)

U.S. Census Bureau: *Census Bureau Releases Population Estimates by Age, Sex, Race, and Hispanic Origin*. 2003. Internet address: http://www.census.gov/Press-Release/www/2003/cb03-16.html. (Accessed 2005.)

Wainer, H.: Precision and Differential Item Functioning on a Testlet-Based Test: The 1991 Law School Admissions Test as an Example. *Applied Measurement in Education* 8:157-186, 1995.

Weech-Maldonado, R., Morales, L.S., Elliott, M., et al.: Race/Ethnicity, Language, and Patients' Assessments of Care in Medicaid Managed Care. *Health Services Research* 38(3):789-808, 2003.

Weidmer, B., Brown, J., and Garcia, L.: Translating the CAHPS® 1.0 Survey Instruments into Spanish: Consumer Assessment of Health Plans Study. *Medical Care* 37(3):MS89-MS96, 1999.

Reprint Requests: Carla M. Bann, Ph.D., P.O. Box 12194, Research Triangle Park, NC 27709-2194. E-mail address: cmb@rti.org