Assessment of the Equivalence of the Spanish and English Versions of the CAHPS® Hospital Survey on the Quality of Inpatient Care

Margarita P. Hurtado, January Angeles, Steven A. Blahut, and Ron D. Hays

Objective. To describe translation and cultural adaptation procedures, and examine the degree of equivalence between the Spanish and English versions of the Agency for Healthcare Research and Quality's (AHRQ) Consumer Assessments of Healthcare Providers and Systems (CAHPS®) Hospital Survey (H-CAHPS®) of patient experiences with care.

Data Sources. Cognitive interviews on survey comprehension with 12 Spanish-speaking and 31 English-speaking subjects. Psychometric analyses of 586 responses to the Spanish version and 19,134 responses to the English version of the H-CAHPS survey tested in Arizona, Maryland, and New York in 2003.

Study Design. A forward/backward translation procedure followed by committee review and cognitive testing was used to ensure a translation that was both culturally and linguistically appropriate. Responses to the two language versions were compared to evaluate equivalence and assess the reliability and validity of both versions.

Data Collection/Extraction Methods. Comparative analyses were carried out on the 32 items of the shortened survey version, focusing on 16 items that comprise seven composites representing different aspects of hospital care quality (communication with nurses, communication with doctors, communication about medicines, nursing services, discharge information, pain control, and physical environment); three items that rate the quality of the nursing staff, physician staff, and the hospital overall; one item on intention to recommend the hospital. The other 12 items used in the analyses addressed mainly respondent characteristics. Analyses included item descriptives, correlations, internal consistency reliability of composites, factor analysis, and regression analysis to examine construct validity.

Principal Findings. Responses to both language versions exhibit similar patterns with respect to item–scale correlations, factor structure, content validity, and the association between each of the seven qualities of care composites with both the hospital rating and intention to recommend the hospital. Internal consistency reliability was slightly, yet significantly lower for the Spanish-language respondents for five of the seven composites, but overall the composites were generally equivalent across language versions.

Conclusions. The results provide preliminary evidence of the equivalence between the Spanish and English versions of H-CAHPS. The translated Spanish version can be used to assess hospital quality of care for Spanish speakers, and compare results across these two language groups.

Key Words. Survey translation and adaptation, patient survey, language equivalence, hospital care quality, Spanish-language survey

The United States is an increasingly diverse country where 45 million people, representing 18 percent of the population, speak a language other than English at home. Among the more than 35 million Hispanics/Latinos who represent the largest minority group, approximately 28 million speak Spanish at home (U.S. Census Bureau 2003). In addition to facing language and cultural barriers, a large proportion of Latinos live below the poverty level and have low literacy skills (Ramirez and De la Cruz 2002), placing them at increased risk for receiving low-quality care. Disparities in access and quality related to language and racial/ethnic differences have been consistently documented (Smedley, Stith, and Nelson 2003), supporting the need for multiple language versions of consumer surveys to accurately assess the quality of care received by those whose main language is not English.

In this article, we describe the cultural adaptation and translation process, and present the results of our analysis of the equivalence between the English- and Spanish-language versions of the Consumer Assessments of Healthcare Providers and Systems (CAHPS®) Hospital Survey (H-CAHPS®) of patient experiences with inpatient care. H-CAHPS was developed in response to a hospital performance public reporting initiative launched by the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) in 2002. It required the development of a standardized survey instrument that would allow for comparisons of quality of care across hospitals. H-CAHPS is a product of AHRQ's CAHPS program to develop surveys that capture patient experiences with health care, design reports to present the results, and design quality improvement ap-

Address correspondence to Margarita P. Hurtado, Ph.D., M.H.S., American Institutes for Research, 10720 Columbia Pike, Suite 500, Silver Spring, MD 20901. January Angeles, M.P.P., is with the American Institutes for Research, Washington, DC. Steven Blahut, Ph.D., M.A., was with the American Institutes for Research when the article was submitted, is now with Olitzky Whittle LLC, Marriottsville, MD. Ron D. Hays, Ph.D., is with the UCLA Department of Medicine/Division of General Internal Medicine and Health Services Research, Los Angeles, CA.

proaches based on the survey results. As was the case for the first CAHPS Health Plan Survey (Weidmer, Brown, and García 1999), the CAHPS team has developed a Spanish version of the hospital survey to allow users to include the many Hispanics in the U.S. who mainly speak Spanish.

The goal of this article is to compare the measurement properties of the Spanish and English versions of the H-CAHPS survey. We precede the psychometric analysis with a brief description of the procedures used for the survey translation and cultural adaptation, including a summary of the results of the cognitive interviews used to test the conceptual equivalence of the Spanish version of the H-CAHPS survey that preceded the field test. Unlike some other studies that examine the measurement properties of a translated survey instrument subsequent to and independently from the original version, in this article we simultaneously assess and compare the measurement properties of the source English version with the translated Spanish version.

METHODS

Development of the Spanish Version of the CAHPS Hospital Survey for the Pilot Study

The Spanish translation and cultural adaptation of the original English-language H-CAHPS survey was conducted following an initial set of procedures established by the CAHPS cultural comparability team. First, a professional translator translated the questionnaire from English to Spanish. Second, another independent professional translator, blinded to the original source questionnaire, was provided with the Spanish translated version and asked to back-translate it into English. Third, a professional translation reviewer examined the products of the two translations, provided comments on the Spanish translation and noted problems identified from the review of the back translation and the original source document. Subsequently, a translation review committee of bilingual researchers with experience in the development of health surveys and translations examined the original Spanish translation and the translation reviewer's comments. The committee reviewed each item with respect to the quality of the translation and its appropriateness for use with culturally diverse Spanish-speaking populations. A consensus process was used to agree on any changes to the items and, when appropriate, the committee provided recommendations for decentering revisions to the English language source document that would result in a better translation and adaptation of the survey. As part of the review, the committee also flagged survey instructions and items that could be difficult for some respondents

because of cultural differences, literacy level, or other reasons. These were investigated using cognitive testing of the Spanish version and addressed by modifying or deleting the problematic items.

The CAHPS Hospital Survey Pilot Study

The H-CAHPS survey was field tested in Arizona, Maryland, and New York. A total of 132 hospitals volunteered to participate. The sample included adult medical or surgical patients with at least one overnight stay, who were discharged between December 2002 and January 2003, and were still alive at the time of the survey; and, obstetric patients discharged between November 2002 and January 2003, who, along with the baby delivered, were still alive at the time of the survey. Patients were excluded from the sample if they were under 18 years old; were admitted for psychiatric or substance abuse treatment, or for observational purposes; died or delivered a baby who died; or, if they were not discharged to a home-setting after the hospital stay.

The data were gathered from the administration of the original 66-item version of the survey, which was longer than the 32-item version analyzed here. The 32 items were selected based on the results of the psychometric analysis of responses to the English version, and other conceptual considerations (Keller et al. 2005). Our analysis of responses to the Spanish version to examine its measurement properties and equivalence to the English version is based on the items retained for the 32-item survey. The items include 16 report items designed to measure seven domains of hospital quality—communication with nurses (three items), communication with doctors (three items), communication about medicines (two items), nursing services (two items), discharge information (two items), pain control (two items), and physical environment (two items). The other items are three rating items on the quality of the nurses, of the physicians, and of the hospital overall; and one item measuring the intention to recommend the hospital. Each report item is measured using a four-point Likert-type scale ranging from "never" to "always," while rating items are on a 0–10 scale. There were 19,134 respondents to the English version and 586 respondents to the Spanish version for a total of 19,720 survey respondents.

Item Characteristics

Data Quality and Item Analysis. After taking into account validly missing responses because they did not apply to that respondent, we calculated the proportion of "inappropriately missing" responses. We also examined the

mean and standard deviation for each report and rating item and the intention to recommend. Finally, we calculated the proportion of respondents in the lowest and highest response categories for each item to examine floor and ceiling effects.

Handling of Missing Data. Individual item analysis was carried out using the original data set in each language version and excluded any missing data. For correlational, factor, and regression analyses, we imputed for missing data by obtaining maximum likelihood estimates of the covariance matrix under the missing at random (MAR) model (Rubin 1987) using *SAS* software (PROC MI). As each missing value was replaced by four imputed values resulting in a dataset that is inflated by a factor of four, we based all statistical tests on the original sample size.

Internal Consistency Reliability

We estimated Cronbach's (1951) coefficient α for each of the seven composites and tested the statistical significance of differences in α (Feldt, Woodruff, and Salih 1987; Cunningham et al. 2003).

Validity

Content Validity. The survey included two open-ended items designed to elicit information on patients' experience of hospital care not covered by the closed-ended items. The first open-ended item asked what the respondent liked most about the care received, and the second asked which aspect of care they would change. We analyzed responses to these items as an indicator of content validity with respect to the comprehensiveness of the survey in both language versions. Responses were coded and analyzed for a random sample of 100 respondents to the Spanish version (18 percent of the total), and 200 respondents to the English version (1 percent of the total). Based on the content, responses were either mapped to one of the H-CAHPS survey items or mapped to new topics that had not been contemplated in the survey.

Item–Scale Correlations. We examined each item's correlation with each of the seven composites to determine whether items correlated more highly with the composite that they were hypothesized to represent, than with other composites.

Factor Structure. We first ran an exploratory factor analysis using a principal axes extraction method and estimated correlations among factors by oblimin rotation (Nunnally and Bernstein 1994). Then, we used confirmatory factor

analysis to test for equivalence in the factor structure of the two language versions. We fit maximum likelihood estimated models to the data from the combined English and Spanish samples, and then fit the model for each of the language groups separately, imposing equality constraints on factor loadings. We also evaluated model fit with and without the constraints. Finally, we examined factorial invariance across language versions by testing loading constraints with the Lagrange multiplier test (Byrne 1994; MacCallum et al. 1994). We used the multiply imputed data to calculate the covariance matrix, but based statistical tests on the original sample size. To assess model fit we relied upon practical fit indices described by Hu and Bentler (1999).

Construct Validity. We examined how the overall hospital rating and intention to recommend item correlated with each of the seven composites. We hypothesized that patients who reported more positive experiences as reflected in the scores for the seven multi-item composites would also rate the hospital higher and would be more likely to recommend it to family and friends. In addition, for each of the language groups, we used general linear models and ran 14 pairs of nested regression models to examine the relationship of each of the seven composites to the overall hospital rating and the intention to recommend item. For each pair, we compared a more complex model that included the language version of the survey with a simpler subset of the model that did not. We examined the significance of language as an explanatory variable of either hospital rating or intention to recommend by comparing the improvement in model fit from the addition of language based on statistics of the change in the model adjusted R^2. We controlled for patient and survey characteristics that are potential confounders because they could be associated with the hospital rating or the intention to recommend independently of language, namely, self-reported health status (overall and mental), age, gender, education level, race/ethnicity, state where surveyed, mode of administration (mail or phone), and the hospital department where the respondent was an inpatient (surgery, internal medicine, or obstetrics).

RESULTS

Cognitive Testing and Cultural Adaptation of the Spanish Version of H-CAHPS

CAHPS team members analyzed the results of 12 cognitive interviews for the Spanish version and 31 for the English version of the survey (see Levine,

Fowler, and Brown 2005). The objective of the interviews was to examine the interpretation and cultural appropriateness of the translated items and identify any other problems in responding to the survey. Interviewees for the Spanish version included four Mexican Americans, one Central American, five South Americans, and two Dominicans to account for potential intra-language variations in usage and understanding because of cultural differences or other reasons.

Most problems identified were either common in both language versions or idiosyncratic to the respondent and unlikely to be related to language or culture. Only a few were specific to the Spanish survey (Levine et al. 2004). Consistent with previous study findings (Carrasco 2003), the item on educational level was problematic for some Spanish-language respondents because of differences in the educational system between the U.S. and their country of origin. This item is based on the question used in the U.S. census. No better alternatives were found to replace it, given the differences in educational systems across Spanish-speaking countries. As hypothesized by the translation review committee, some of the respondents also had difficulty with the item that asked how often did hospital staff "introduce themselves to you" (Q28). The Spanish translation of "introduced themselves" ("*se le presentaban*") can be confused with "they were present" ("*se presentaban*"). This item was dropped for the 32-item version of the H-CAHPS survey, which is the one tested and reported on here. For two of the Spanish-speaking interviewees, there were discrepancies between their descriptions of their experiences and their responses to the "doctor courtesy and respect" (Q11) and doctor rating (Q15) items. The two interviewees reported negative experiences, yet provided favorable responses to these items indicating that they might have used different criteria for judging quality. Finally, some Spanish-language respondents did not expect to be involved in treatment decisions and had difficulty with the item on this topic (Q25: "Hospital staff involved you in decisions about your treatment as much as you wanted"). This item was also dropped for the 32-item version of the survey, which was developed based on the results of the cognitive interviews, public comments, discussions by the H-CAHPS instrument development team, and recommendations of the Spanish translation and cultural adaptation review committee.

Characteristics of Respondents to the Spanish and English Versions of H-CAHPS

Respondents to the English version represent 97 percent (19,134) of all completed surveys, while Spanish version respondents represent the other 3

percent (586). Compared with the English version, Spanish version respondents tend to be much younger (the 50th percentile was in the 25–34-year-old group versus the 55–64-year-old group for English), have less education (40 percent had less than a ninth grade education versus 5 percent for English) and include a higher proportion of women (86 versus 67 percent for English) (Table 1). Most of the Spanish version respondents identified themselves as Hispanic (97 versus 7 percent for English) and about half also identified themselves as white (52 versus 74 percent for English); or, of "other race" or more than one race (46 versus 14 percent for English).

A greater proportion of respondents to the Spanish version identified themselves as being in "fair" or "poor" overall health as compared with English version respondents (27 versus 20 percent, respectively; $\chi^2 = 47.8, p < .01$). Similarly, a greater proportion of Spanish respondents reported that they were in "fair" or "poor" overall mental health than English respondents (16 versus 12 percent, respectively, $\chi^2 = 49.6, p < .01$). More than two-thirds of Spanish version respondents (72 percent), compared with approximately one-fifth of English version respondents (22 percent), were drawn from hospitals in Arizona (72 percent) rather than Maryland or New York ($\chi^2 = 767.5, p < .01$). Spanish version respondents were also more likely than English version respondents to respond by phone (57 versus 17 percent, respectively, $\chi^2 = 591.3, p < .01$) rather than mail. Lastly, consistent with the gender and age differences noted above, Spanish respondents included a much higher proportion of patients who received obstetrical services (63 versus 22 percent, $\chi^2 = 471.5, p < .01$), rather than internal medicine or surgical services.

Characteristics of Survey Items and Composites

In both language versions, the same two items had the highest proportions of "inappropriately missing" responses (i.e., items that were applicable and should have been answered but were not). For the discharge item asking if hospital staff "talk with you about whether you would have the help you needed when you left the hospital" (Q48), 35 percent of responses were missing for the Spanish group and 29 percent for the English group. For the item on whether staff described medication side effects to the patient (Q41), the percentage of missing responses was 9 percent for the Spanish group and 14 percent for the English one (Electronic Appendix—Table A1).

Responses to report and rating items were skewed towards higher values in both language versions, and exhibited ceiling effects (Table 2). The magnitude of the ceiling effects for the report items did not have a consistent

Table 1:   Respondent Characteristics for the English and Spanish H-CAHPS Versions

Characteristic	English Survey (%)	Spanish Survey (%)	χ^2 stat. (p)
Overall health	n = 18,638	n = 571	47.80 (<.01)
Excellent	20.83	24.34	
Very good	29.97	19.44	
Good	28.72	28.90	
Fair	16.28	24.34	
Poor	4.21	2.98	
Overall mental health	n = 17,661	n = 557	49.57 (<.01)
Excellent	32.57	30.88	
Very good	31.41	19.93	
Good	24.45	33.03	
Fair	9.81	14.00	
Poor	1.76	2.15	
Current age	n = 17,810	n = 557	620.96 (<.01)
18–24	6.03	23.34	
25–34	16.44	34.47	
35–44	12.77	15.26	
45–54	12.23	5.03	
55–64	14.83	8.26	
65–74	17.03	7.54	
75–79	9.13	2.69	
80 or older	11.54	3.41	
Gender	n = 17,818	n = 560	88.34 (<.01)
Male	33.20	14.29	
Female	66.80	85.71	
Educational level	n = 17,622	n = 555	1,293.11 (<.01)
Eighth grade or less	4.63	39.46	
Some high school, but did not graduate	9.95	14.05	
High school graduate or GED	28.60	26.31	
Some college or 2-year degree	29.32	14.59	
Four-year college graduate degree	12.36	2.88	
More than 4-year college degree	15.14	2.70	
Ethnicity	n = 17,167	n = 561	4,808.50 (<.01)
Hispanic	7.34	97.33	
Non-Hispanic	92.66	2.67	
Race	n = 14,013	n = 584	451.81 (<.01)
White	73.62	51.71	
Black	9.07	1.88	
Asian	1.92	0.00	
Native Hawaiian or other Pacific Islander	0.11	0.00	
American Indian or Alaskan Native	0.84	0.34	
Other (including multiple race)	14.45	46.06	

Table 1:   *Continued*

Characteristic	English Survey (%)	Spanish Survey (%)	χ^2 stat. (p)
State	n = 19,134	n = 586	767.47 (<.01)
Arizona	22.47	71.84	
Maryland	33.22	7.68	
New York	44.31	20.48	
Survey mode	n = 19,134	n = 586	591.35 (<.01)
Mail	82.54	42.83	
Telephone	17.46	57.17	
Service line	n = 19,134	n = 586	471.55 (<.01)
Medical	37.08	16.55	
Surgical	40.75	20.65	
Obstetrics	22.17	62.80	

Note: H-CAHPS, Hospital Survey-Consumer Assessments of Healthcare Providers and Systems; GED, General Educational Development.

pattern across language versions; the proportion at the highest response category was larger in the English group for about half of the items, and larger in the Spanish group for the other half. In comparison, ceiling effects for the three rating items and the intention to recommend item were consistently higher for the Spanish than for the English group, with the proportion of responses at the highest category ranging from 53 to 68 percent for the Spanish group, and, 35 to 64 percent for the English group. There were minimal floor effects for both language groups (Electronic Appendix—Table A1).

Standard deviations for most composite scores were slightly higher for the English group than for the Spanish group, with differences ranging from 0.01 to 0.11. Standard deviations for the rating items and intention to recommend items were also higher for the Spanish group than for the English one (Table 2).

Although equality of item means is not an indicator of the equivalence of the instrument's measurement properties, it is interesting to note that only two of the seven composite means showed any statistically significant differences across the two language versions. The Spanish group gave significantly higher scores for the nursing services composite ($\Delta = 0.12$ on a 0–4 scale, $t = -3.42$, $p < .01$) and physical environment composite ($\Delta = 0.20$ on a 0–4 scale, $t = -7.93$, $p < .01$), for all global ratings (Δ ranged from 0.52 to 0.71 on a 0–10 scale, $p < .01$), and for the intention to recommend item ($\Delta = 0.08$ on a 1–4 scale, $t = -3.07$, $p < .01$) (Table 2).

Table 2: Item and Composite Characteristics for English and Spanish H-CAHPS

Composite/Items	Mean (Standard Deviation)			Ceiling Effects— Proportion at the Highest Response Category (%)		
	t-Statistic (p)	English	Spanish	χ² (p)	English	Spanish
Nurse communications	−0.52 (.60)	3.52 (0.63)	3.53 (0.56)			
Nurse respect (Q4)	1.54 (.12)	3.63 (0.63)	3.59 (0.65)	3.35 (.07)	70.70	67.18
Nurse listen (Q5)	−4.36 (<.01)	3.46 (0.73)	3.58 (0.65)	14.31 (<.01)	58.82	66.67
Nurse explain (Q6)	1.14 (.25)	3.48 (0.76)	3.44 (0.75)	2.84 (.09)	61.76	58.30
Doctor communications	−0.23 (.82)	3.59 (0.62)	3.60 (0.54)			
Doctor respect (Q11)	0.56 (.58)	3.68 (0.61)	3.67 (0.59)	3.24 (.07)	75.06	71.77
Doctor listen (Q12)	−1.59 (.11)	3.55 (0.72)	3.59 (0.64)	0.00 (.98)	66.49	66.44
Doctor explain (Q13)	0.60 (.55)	3.55 (0.72)	3.53 (0.68)	3.80 (.05)	66.26	62.37
Communication about meds	−1.83 (.07)	3.20 (0.90)	3.32 (0.89)			
Allergies to Rx (Q40)	−0.05 (.96)	3.54 (0.86)	3.54 (0.87)	0.01 (.91)	73.17	72.78
Rx side effects (Q41)	−2.54 (.01)	2.86 (1.17)	3.09 (1.12)	5.54 (.02)	41.74	50.57
Nursing services	−3.42 (<.01)	3.22 (0.80)	3.34 (0.74)			
Call button response (Q9)	−3.95 (<0.01)	3.16 (0.85)	3.31 (0.79)	9.78 (<.01)	40.92	48.65
How often bathroom (Q22)	−1.24 (.22)	3.29 (0.87)	3.35 (0.79)	0.00 (.98)	52.18	52.28
Discharge information	−1.71 (.08)	1.82 (0.34)	1.84 (0.30)			
Help after discharge (Q48)	−1.62 (.11)	1.83 (0.38)	1.79 (0.41)	2.98 (.08)	82.67	79.27
Symptoms in writing (Q49)	2.81 (.01)	1.82 (0.38)	1.87 (0.34)	6.44 (.01)	82.39	86.54
Pain control	−1.43 (.15)	3.46 (0.70)	3.50 (0.59)			
Pain controlled (Q32)	−1.57 (.12)	3.39 (0.76)	3.44 (0.67)	0.06 (.81)	53.78	54.36
MD pain help (Q33)	−0.82 (.41)	3.53 (0.74)	3.56 (0.64)	1.35 (.25)	65.52	62.84
Physical environment	−7.93 (<.01)	3.35 (0.68)	3.55 (0.59)			
Room clean (Q17)	−3.62 (<.01)	3.49 (0.77)	3.60 (0.70)	9.44 (<.01)	63.42	69.67
Room quiet (Q18)	−8.96 (<.01)	3.21 (0.87)	3.50 (0.76)	75.71 (<.01)	45.52	63.84
Global nurses rating	−8.61 (<.01)	8.35 (2.03)	8.97 (1.70)	63.56 (<.01)	37.39	53.71
Global doctors rating	−8.57 (<.01)	8.67 (1.92)	9.19 (1.43)	40.70 (<.01)	47.33	60.76
Global hospital rating	−10.99 (<.01)	8.36 (2.01)	9.07 (1.51)	72.97 (<.01)	35.30	52.60
Recommend hospital	−3.07 (<.01)	3.52 (0.75)	3.60 (0.67)	7.62 (.05)	63.67	68.35

Note: Composites, composite items, and the recommend hospital item are on a 1–4 scale; rating items are on a 0–10 scale.
H-CAHPS, Hospital Survey-Consumer Assessments of Healthcare Providers and Systems; Rx, prescription medicines.

Internal Consistency Reliability

Internal consistency reliability of the composites ranged from 0.51 to 0.88 for the English group and from 0.38 to 0.81 for the Spanish group. Patterns with respect to the magnitude of the αs were similar across language groups, with the nurse communication and doctor communication composites exhibiting the highest αs, and the discharge information and physical environment composites exhibiting the lowest αs. For five of the seven composites, the αs were significantly higher for the English group than the Spanish group (Table 3).

Item–Scale Correlations

The Spanish and English versions of the instrument exhibited a similar, though not identical, pattern of item–composite correlations (Table 3). The correlations between individual items and their hypothesized composite for the Spanish group were slightly lower or identical to the correlations for the English group. This held true for every item except "doctor explains things in a way you could understand" (Q13) where the correlation for the Spanish group was 0.76 versus 0.73 for the English group. For both language groups, items belonging to the communication with doctors, communication with nurses, and pain control composites had the highest set of item–composite correlations ($r = 0.69$–0.82 for English, and $r = 0.50$–0.76 for Spanish). Items for the discharge information and physical environment composites had the lowest item–composite correlations in both language groups, with correlations ranging from 0.25 to 0.35. Nursing services items exhibited somewhat higher correlations with the nurse communication composite than with their hypothesized composite, for both language groups. Physical environment items also exhibited higher correlations with other composites rather than with their hypothesized composite, for both the Spanish and English groups.

Validity

Content Validity. For both language groups, the majority of responses to the two open-ended items included in the H-CAHPS survey either mapped to an existing item or were left blank, indicating that the survey is comprehensive. One of the few aspects not captured in the closed-ended items that appears to differ across language versions, but was seldom mentioned was the importance of language (mentioned by 13 percent of Spanish version respondents and less than 1 percent of English version respondents; e.g., "They have translators for people who speak different languages"). Another

Table 3: Internal Consistency Reliability, and Composite and Item Intercorrelations for the English and Spanish H-CAHPS

	α*	Nurse Communications	Doctor Communications	Communications about Meds	Nursing Services	Discharge Information	Pain Control	Physical Environmental
Composite								
Nurse communications (NC)	0.86* (0.78)	1.00 (1.00)						
Doctor communications (DC)	0.88* (0.81)	0.50 (0.51)	1.00 (1.00)					
Communication about medicines (Med)	0.67 (0.68)	0.46 (0.35)	0.38 (0.33)	1.00 (1.00)				
Nursing services (NS)	0.72 (0.65)	0.67 (0.62)	0.41 (0.41)	0.44 (0.35)	1.00 (1.00)			
Discharge information (D)	0.52* (0.38)	0.27 (0.23)	0.26 (0.20)	0.33 (0.28)	0.25 (0.19)	1.00 (1.00)		
Pain control (P)	0.83* (0.77)	0.56 (0.47)	0.43 (0.42)	0.41 (0.34)	0.55 (0.46)	0.23 (0.20)	1.00 (1.00)	
Physical environment (Env)	0.51* (0.44)	0.47 (0.38)	0.32 (0.34)	0.35 (0.24)	0.51 (0.38)	0.18 (0.18)	0.40 (0.40)	1.00 (1.00)
Items by composite								
NC—nurse respect (Q4)		**0.73 (0.61)**	0.40 (0.42)	0.38 (0.26)	0.59 (0.53)	0.22 (0.16)	0.50 (0.43)	0.42 (0.33)
NC—nurse listen (Q5)		**0.77 (0.66)**	0.44 (0.41)	0.42 (0.28)	0.63 (0.53)	0.24 (0.16)	0.52 (0.40)	0.44 (0.33)
NC—nurse explain (Q6)		**0.69 (0.50)**	0.47 (0.41)	0.42 (0.30)	0.57 (0.46)	0.26 (0.25)	0.46 (0.33)	0.40 (0.29)
DC—doctor respect (Q11)		0.43 (0.42)	**0.76 (0.75)**	0.31 (0.24)	0.35 (0.34)	0.21 (0.14)	0.38 (0.34)	0.28 (0.27)
DC—doctor listen (Q12)		0.46 (0.41)	**0.82 (0.70)**	0.35 (0.27)	0.39 (0.36)	0.24 (0.15)	0.40 (0.38)	0.30 (0.33)
DC—doctor explain (Q13)		0.45 (0.45)	**0.73 (0.76)**	0.36 (0.32)	0.38 (0.35)	0.24 (0.22)	0.38 (0.35)	0.28 (0.28)
Med—allergies to Rx (Q40)		0.36 (0.27)	0.28 (0.25)	**0.51 (0.51)**	0.34 (0.30)	0.23 (0.16)	0.32 (0.27)	0.27 (0.20)
Med—Rx side effects (Q41)		0.43 (0.33)	0.37 (0.31)	**0.51 (0.51)**	0.41 (0.30)	0.33 (0.30)	0.38 (0.31)	0.33 (0.22)
NS—call button response (Q9)		0.63† (0.57)†	0.37 (0.38)	0.39 (0.30)	**0.56 (0.48)**	0.21 (0.16)	0.50 (0.41)	0.44 (0.30)
NS—how often bathroom (Q22)		0.56 (0.50)†	0.36 (0.33)	0.38 (0.30)	**0.56 (0.48)**	0.23 (0.17)	0.47 (0.39)	0.46 (0.35)
D—help after discharge (Q48)		0.23 (0.14)	0.20 (0.13)	0.28 (0.21)	0.22 (0.15)	**0.35 (0.25)**	0.21 (0.16)	0.17 (0.12)
D—symptoms in writing (Q49)		0.21 (0.24)	0.22 (0.20)	0.27 (0.23)	0.19 (0.15)	**0.35 (0.25)**	0.17 (0.16)	0.13 (0.17)
P—pain controlled (Q32)		0.46 (0.40)	0.36 (0.36)	0.35 (0.31)	0.47 (0.38)	0.19 (0.22)	**0.71 (0.63)**	0.34 (0.33)
P—MD pain help (Q33)		0.57 (0.44)	0.43 (0.40)	0.40 (0.30)	0.54 (0.45)	0.24 (0.14)	**0.71 (0.63)**	0.39 (0.39)
Env—room clean (Q17)		0.43† (0.34)	0.28 (0.28)	0.33 (0.23)	0.45† (0.34)†	0.17 (0.14)	0.34† (0.31)	**0.34 (0.28)**
Env—room quiet (Q18)		0.35* (0.28)	0.25 (0.27)	0.25 (0.16)	0.39 (0.27)†	0.13 (0.15)	0.31† (0.33)	**0.34 (0.28)**

Note: Spanish results are in parentheses next to the English ones.

**z for English version was significantly higher than for Spanish version, *p* < .05.

†Item correlates more highly with a different composite from one hypothesized. Pairwise correlations conducted with imputed data, *n* = 76,696 for English responses and *n* = 2,344 for Spanish responses.

Boldface is used to denote the diagonal of the item intercorrelation matrix, which lists the correlation between the items and the scale that each item is induced in.

H-CAHPS, Hospital Survey-Consumer Assessments of Healthcare Providers and Systems; Rx, prescription medicines.

aspect that was not included in the survey and differs by language, but was rarely noted, is coordinating care among hospital staff or with outside providers (mentioned by 3 percent of English version respondents but not at all by Spanish version respondents; e.g., "The doctor that was needed for my problem was on call").

Factor Structure. For the Spanish version responses, the results of the exploratory factor analysis in which seven correlated factors were rotated showed that the items hypothesized for each composite loaded together onto distinct factors and had the highest loadings on the factor corresponding to that composite. Responses to the English version exhibited a similar pattern for the items belonging to five of the seven composites. However, items for the nursing services and physical environment composites loaded together on one factor, and the highest item loading for the seventh factor was only 0.21.

Factor loadings for the first-order model to confirm the seven-factor model were above 0.70 for all composite (factor) items, except those items referring to discharge information, physical environment, or communication about medicines. This was true for both language groups (Table 4). All loadings were statistically significant at $p < .01$. The analysis also shows that while four of the 16 report items had significantly higher loadings on their respective factors in the English version than in the Spanish version [Nurse respect (Q4), 0.81 versus 0.77, $\chi^2 = 4.67$, $p = .03$; Nurse listens (Q5), 0.87 versus 0.85, $\chi^2 = 4.75$, $p = .03$; Doctor listens (Q12), 0.91 versus 0.88, $\chi^2 = 8.26$, $p = .004$, and Doctor's help with pain (Q33), 0.92 versus 0.91, $\chi^2 = 6.30$, $p = .01$] (Table 4), the magnitude of the difference in the loadings was very small (ranging from 0.01 to 0.04).

For both language versions, the results also revealed that the first-order factors were highly correlated with each other, indicating a potential second-order factor structure. This was confirmed in the subsequent analysis, which showed that the correlations among the first-order factors could be accounted for by a single second-order factor. Multiple population simultaneous fit statistics for both the English and Spanish versions showed satisfactory model fit (Table 4). With the unstandardized coefficients constrained to be equal in each language group, the fit indices for both the first-order factor structure (CFI = 0.981, SRMR = 0.116, RMSEA = 0.027) and the second-order structure (CFI = 0.972, SRMR = 0.052, RMSEA = 0.032) were adequate (Hu and Bentler 1999).

Table 4: Confirmatory Factor Analysis of H-CAHPS Survey Report Items—First-Order Factor Loadings

Construct	Survey Item Variable	English Version Loading (Uniqueness of Error)	Spanish Version Loading (Uniqueness of Error)	Equality of Factor Loadings across Language, χ^2 Stat (p)
Nurse communications	Nurse respect (Q4)	0.812 (0.584)	0.766 (0.643)	4.668 (.031)*
	Nurse listen (Q5)	0.874 (0.486)	0.853 (0.522)	4.746 (.029)*
	Nurse explain (Q6)	0.772 (0.635)	0.709 (0.705)	3.011(.083)
Doctor communications	Doctor respect (Q11)	0.833 (0.554)	0.791 (0.611)	0.250 (.617)
	Doctor listen (Q12)	0.910 (0.414)	0.876 (0.482)	8.257 (.004)*
	Doctor explain (Q13)	0.793 (0.610)	0.778 (0.628)	0.3540 (.552)
Communication about medicines	Allergies to Rx (Q40)	0.632 (0.775)	0.639 (0.769)	0.343 (.558)
	Rx side effects (Q41)	0.798 (0.603)	0.819 (0.574)	0.128 (.721)
Nursing services	Call button response (Q9)	0.778 (0.628)	0.771 (0.637)	0.015 (.903)
	How often bathroom (Q22)	0.720 (0.694)	0.718 (0.696)	0.116 (.733)
Discharge information	Help after discharge (Q48)	0.603 (0.798)	0.551 (0.834)	3.150 (.076)
	Symptoms in writing (Q49)	0.577 (0.817)	0.621 (0.784)	1.358 (.244)
Pain control	Pain controlled (Q32)	0.772 (0.635)	0.785 (0.619)	0.342 (.559)
	MD pain help (Q33)	0.916 (0.402)	0.907 (0.422)	6.303 (.012)*
Physical environment	Room clean (Q17)	0.633 (0.774)	0.651 (0.759)	1.244 (.265)
	Room quiet (Q18)	0.538 (0.843)	0.582 (0.813)	0.326 (.568)

Note: Model fit statistics for both models simultaneously: CFI = 0.981, standard. RMR = 0.116, RMSEA = 0.027. Significance tests are based on the original sample size.

*χ^2 statistic for unequal item level factor loadings across the Spanish and English versions has $p < .05$. All factor loadings significant at $p < .01$ level.
H-CAHPS, Hospital Survey-Consumer Assessments of Healthcare Providers and Systems; Rx, prescription medicines.

Construct Validity. Correlations between the global hospital rating and each of the seven composites range from 0.29 to 0.69 for the English group, and from 0.24 to 0.52 for the Spanish group. Correlations between the intention to recommend item and each of the composites range from 0.26 to 0.60 for the English group, and 0.28 to 0.51 for the Spanish group. The pattern of correlations was very similar across language versions. The communication with nurses and nursing services composites are most highly correlated with the overall hospital rating and intention to recommend item, while the discharge information and communication about medicines composites were the least correlated with either hospital rating or intention to recommend it (Table 5).

Nested regression models show that when the global hospital rating and intention to recommend item were regressed on each of the seven composites, controlling for potential confounders (overall and mental health status, age, gender, education, race/ethnicity, state where surveyed, survey mode, and service line), the addition of the language version to the model has a negligible effect on the regression coefficient for the corresponding composite and the adjusted R^2 for the model (language increases R^2 by less than 1 percent in all cases) (Table 5).

DISCUSSION

A variety of approaches are needed to ensure the proper translation and adaptation of survey instruments. Cognitive interviewing is important in the development of translations as a way of identifying potential differences between language and cultural groups up front, and ensuring that survey items measure the same construct irrespective of the language version and the culture of the intended respondents (Carrasco 2003). The results of the cognitive interviews for the Spanish version of the CAHPS Hospital Survey revealed that few problems with the questionnaire were language specific and suggested most items measured the same construct regardless of language version.

Despite marked differences between the two language groups in sociodemographic, health status, and sampling-related characteristics, the results of this analysis provide evidence for the equivalence between the Spanish and English language versions of the H-CAHPS survey. Similar patterns of item–composite correlations across language versions were found, with most correlations being equal or differing by less than 0.03 between the two language versions. With respect to data quality, one of the items on being asked

Table 5:   Correlations between Composites and Hospital Rating, and Intention to Recommend Item by Language, with Nested Linear Regression Models Examining the Effect of Language for the Same Items Regressed on the Composites

Item Hypothesized to be Associated with the Composites (Composites)	Correlations		Nested Linear Regression Models		
	Eng.	Spa.	Unstandardized β Coefficient for Composite (SE) (w/out Language)*	Unstandardized β Coefficient for Composite (SE) (with Language)*	Change in Model Adj. R² with Language†
Hospital rating					
Nurse communications	0.69	0.52	2.131 (0.009)	2.130 (0.009)	0.001
Doctor communications	0.49	0.40	1.495 (0.011)	1.496 (0.011)	0.011
Communication about meds	0.45	0.33	0.961 (0.008)	0.961 (0.008)	0.001
Nursing services	0.61	0.45	1.564 (0.008)	1.564 (0.008)	0.001
Discharge information	0.29	0.24	1.799 (0.023)	1.801 (0.023)	0.001
Pain control	0.55	0.42	1.485 (0.009)	1.484 (0.009)	0.001
Physical environment	0.50	0.34	1.402 (0.010)	1.401 (0.010)	0.001
Recommend hospital					
Nurse communications	0.60	0.51	0.708 (0.004)	0.708 (0.004)	<0.001
Doctor communications	0.44	0.40	0.506 (0.004)	0.506 (0.004)	<0.001
Communication about meds	0.39	0.28	0.311 (0.003)	0.311 (0.003)	<0.001
Nursing services	0.52	0.41	0.509 (0.003)	0.509 (0.003)	<0.001
Discharge information	0.26	0.29	0.606 (0.009)	0.606 (0.009)	<0.001
Pain control	0.48	0.38	0.496 (0.004)	0.496 (0.004)	<0.001
Physical environment	0.42	0.32	0.448 (0.004)	0.448 (0.004)	<0.001

Note: Correlation and regression analyses were run using the multiply imputed sample, English $n = 76,696$ and Spanish $n = 2,344$. Hypothesis testing for the nested regression models was based on the original sample size.

*Adjusted for overall and mental health status, age, gender, education, race/ethnicity, state where surveyed, survey mode, and service line.

†All were statistically significant at $p < .001$.

about having the help needed after discharge (Q48) exhibited problems in both language versions as evidenced by the high proportion of missing responses (29 and 35 percent for the English and Spanish groups, respectively). Both discharge-related items (Q48 and Q49) were also problematic based on the results of the factor analysis. However, they were retained in the 32-item survey given the importance of receiving proper discharge instructions for successfully transitioning from the hospital to other settings (Keller et al. 2005).

This study also provides preliminary evidence for the validity of both language versions of the survey. With respect to content validity, analysis of the responses to the two open-ended items showed that the close-ended survey items already covered most aspects of hospital care that patients care about. This held true for both the English and Spanish survey versions.

The confirmatory factor analysis indicated satisfactory model fit for both language versions indicating construct validity. The analysis for the first-order model demonstrated satisfactory fit for a seven factor structure. Factor loadings were equal across language versions for 12 of the 16 report items, and for the other four report items loadings differed by less than 0.05. As was the case for many of the analyses, a large number of findings are statistically significant due in part to the very large sample (19,134 for English and 586 for Spanish), but effect sizes tend to be negligible. The analysis for a second-order model shows that a single factor can account for the correlations among first-order factors. This second-order single factor can be interpreted as a generalized construct of hospital-based quality of care underlying the seven first-order factors. Further evidence of construct validity is provided by the results of the correlational and regression analyses: the composites exhibit the same hypothesized relationships with the hospital rating and the intention to recommend item across language versions. In addition, language version does not have a significant main effect in regressions of hospital rating and the intention to recommend item on each of the composites. Both sets of results suggest that the two language versions are equivalent in this respect.

The internal consistency reliability of five of the seven composites is significantly lower for the Spanish version than for the English one. However, the differences in α are small, ranging from 0.06 to 0.14. For the physical environment composite, where the difference is the largest (0.14), the difference may be because of mode effects, as those who responded by telephone tended to provide more favorable responses for these items (see de Vries et al. 2005), and respondents to the Spanish version did so by telephone, rather than mail, 40 percent more often than those responding to the English version.

For other areas where we found differences in certain measurement properties across the two language versions, it is unclear how much is because of language as compared with mode of administration or other factors. There is some indication that the rating items may be working somewhat differently for some Spanish version respondents. The cognitive interviewing results show that two subjects for the Spanish version provided higher ratings than expected given their reported experiences with care. The means for all three rating items are also significantly higher for the Spanish group than for the English group. This difference was documented previously for the CAHPS 3.0 Health Plan Survey (Weech-Maldonado et al. 2003) and a patient satisfaction hospital survey (Miceli 2004). In the case of H-CAHPS, however, differences in ratings are probably not because of mode effects as none were identified for these items in a separate analysis (de Vries et al. 2005). This suggests that there may be cultural differences in the way that ratings are ascertained. This merits further investigation. In comparison, intention to recommend the hospital, which is also higher for the Spanish group seems to be associated with the tendency to provide more favorable responses by phone as mode effects were identified (see de Vries et al. 2005).

This study shows that the preference for telephone administration is higher among Spanish language respondents, even when the survey is available in Spanish. This may be because of greater difficulties responding to the mail version because of lack of familiarity with surveys or with the U.S. health care system (for those who are foreign born), and/or low literacy, all of which make it easier to respond by phone. As non-English speakers also tend to face greater barriers to care, it is essential that they be included in surveys to identify disparities in quality of care and improve it where it is possible. Thus, reaching populations for whom English is not their main language, may require not only developing culturally appropriate translated versions of the survey, but also making it available by telephone, even if it is generally more expensive to do so.

The results of this study provide preliminary evidence of the equivalence of the Spanish and English language versions of the H-CAHPS survey and indicate that the translated version can be used to assess experiences with hospital care for patients who mainly or only speak Spanish, and compare them with the experiences of those who mainly or only speak English. Gathering further evidence from a larger and more varied sample of respondents would be useful as our study was limited to 586 Spanish respondents, most of whom were from Arizona.

Until recently, there were no standardized guidelines for survey translation and adaptation procedures, as noted in a recent review conducted by the U.S. Census Bureau (2004). The cultural adaptation and translation of the H-CAHPS survey into Spanish is one of a series of translations of the new set of CAHPS surveys being developed to assess consumer reports of quality of care at different sites including hospitals, physician offices, dialysis facilities, and nursing homes. As part of this process, the CAHPS team is developing guidelines for the translation and cultural adaptation of CAHPS surveys. They are being designed so that they are useful to AHRQ as well as other organizations wishing to develop other translations of CAHPS surveys. The development and testing of the Spanish version of the CAHPS Hospital Survey is part of an iterative process that the CAHPS team is undertaking to define the guidelines.

REFERENCES

Byrne, B. M. 1994. *Structural Equation Modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage Publications.

Carrasco, L. 2003. *The American Community Survey en Español: Using Cognitive Interviews to Test the Functional Equivalence of Questionnaire Translations*. Washington, DC: U.S. Census Bureau. Retrieved May 5, 2005 from www.fcsm.gov/03papers/carrasco_Final.pdf.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16: 297–334.

Cunningham, W. E., R. D. Hays, T. M. Burton, D. B. Reuben, and R. S. Kington. 2003. "Correlates of Social Function: A Comparison of a Black and White Sample of Older Persons in Los Angeles." *Journal of Applied Gerontology* 22 (1): 3–18.

de Vries, H., M. Elliot, K. A. Hepner, S. D. Keller, and R. D. Hays. 2005. "Equivalence of Mail and Telephone Responses to the Hospital CAHPS (H-CAHPS) Survey." *Health Services Research.* DOI: 10.1111/j.1475-6773.2005.00479.x. Available at www.blackwell-synergy.com.

Feldt, L. S., D. J. Woodruff, and F. A. Salih. 1987. "Statistical Inference for Coefficient Alpha." *Applied Psychological Measurement* 11 (1): 93–103.

Hu, L., and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6: 1–55.

Keller, S., J. O'Malley, K. Hepner, R. D. Hays, A. M. Zaslavsky, and P. Cleary. (2005). "Methodological Approach to Deriving the Shortened H-CAHPS Survey." *Health Services Research.* DOI: 10.1111/j.1475-6773.2005.00478.x. Available at www.blackwell-synergy.com.

Levine, R. E., F. J. Fowler, and J. A. Brown. (2005). "Role of Cognitive Testing in the Development of the CAHPS® Hospital Survey." *Health Services Research.* DOI: 10.1111/j.1475-6773.2005.00472.x. Available at www.blackwell-synergy.com.

Levine, R., R. González, B. Weidmer, and P. Gallagher. 2004. "Cognitive Testing of English and Spanish Versions of Health Questionnaires." Presented at the American Association of Public Opinion Researchers 59th Annual Research Conference, Phoenix, 14 May.

MacCallum, R. C., M. Roznowski, C. M. Mar, and J. V. Reith. 1994. "Alternative Strategies for Cross-Validation of Covariance Structure Models." *Multivariate Behavioral Research* 29: 1–32.

Miceli, P. J. 2004. "Validating a Patient Satisfaction Survey Translated into Spanish." *Journal of Healthcare Quality* 26 (4): 4–13.

Nunnally, J. C., and I. H. Bernstein. 1994. *Psychometric Theory.* New York: McGraw-Hill.

Ramirez, R. R., and G. P. de la Cruz. 2002. "The Hispanic Population in the United States: March 2002." Current Population Reports, P20-545. Washington, DC: U.S. Census Bureau.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Smedley, B. D., A. Y. Stith, and A. R. Nelson, (eds.). 2003. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Washington, DC: National Academies Press.

U.S. Census Bureau. 2003. "Table 1. Language Use, English Ability, and Linguistic Isolation for the Population 5 Years and Over by State: 2000." Retrieved May 2, 2005 from www.census.gov/population/cen2000/phc-t20/tab01.xls.

U.S. Census Bureau. 2004. "The Translation of Surveys: An Overview of Methods and Practices and the Current State of Knowledge." In *Census Bureau Guideline for the Translation of Data Collection Instruments and Supporting Materials: Documentation on*

How the Guideline Was Developed. Washington, DC: U.S. Census Bureau. Available at http://www.census.gov/srd/papers/pdf/rsm2005-06.pdf, retrieved September 24, 2005.

Weech-Maldonado, R., L. Morales, M. Elliott, K. Spritzer, G. Marshall, and R. D. Hays. 2003. "Race/Ethnicity, Language, and Patients' Assessments of Care in Medicaid Managed Care." *Health Services Research* 38 (3): 789–808.

Weidmer, B., J. Brown, and L. García. 1999. "Translating the CAHPS® 1.0 Survey Instruments into Spanish." *Medical Care* 37: MS89–96.