Psychometric Analysis of the Spanish-Language Version of the Medicare Health Outcomes Survey

Prepared by:

Health Assessment Lab
Boston, Massachusetts

for:

National Committee on Quality Assurance
Health Care Financing Administration

March 2000

TABLE OF CONTENTS

Executive Summary

A Spanish-language version of the Medicare Health Outcomes Survey (HOS) was made available to Cohort II respondents who requested the form during the 1999 baseline administration. A total of 209 respondents completed the Spanish-language survey. On average, respondents to the Spanish-language form were less educated, less healthy, and younger than Hispanic respondents who completed the English-language form during the Cohort I baseline administration. Spanish-language respondents had mean scores of 34.5 and 44.2 on the PCS and MCS, respectively, compared to scores of 40.2 (PCS) and 49.1 (MCS) for English-language Hispanic respondents ($p<0.0001$ for difference between language groups). Psychometric tests using data from the Spanish-language sample showed that the SF-36 scales generally met scaling assumptions and had satisfactory internal consistency reliability (Cronbach's alpha=0.84 to 0.94). Scales generally appeared to be measures of physical or mental constructs as hypothesized, in very preliminary tests of clinical validity. However, factor analytic results for the Spanish-language respondents differed by education level, and only replicated findings for the English-speaking U.S. general population in the most educated group (high school graduate or greater).

It is noteworthy that about 11 percent of respondents identified as Hispanic (using HCFA race data) did not complete the questionnaire due to language problems in both Cohort I and Cohort II. Additional steps (e.g., development of a telephone version of the Spanish form; including both English and Spanish-language forms in mailings to respondents identified as Hispanic) may encourage greater participation, but may also increase the costs of data collection.

Introduction

This report examines the psychometric properties of the Spanish translation of the Medicare Health Outcomes Survey (HOS) and the characteristics of Cohort II beneficiaries who completed the Spanish form. The report briefly discusses the development of the Spanish-language HOS questionnaire, which was first made available during the Cohort II baseline administration in 1999. It then compares beneficiaries who responded to the Spanish-language form during the Cohort II baseline administration with Hispanic respondents who completed the English-language form during the Cohort I baseline administration. The psychometric properties of the Spanish HOS form are evaluated, focusing on tests of scaling assumptions, internal consistency reliability, factor analysis of the eight SF-36 scales, and preliminary tests of known-groups validity. Scoring algorithms for the eight SF-36 scales and two summary measures are provided for the Spanish form. SF-36 scores for Spanish-language and Hispanic English-language samples are compared. Finally, recommendations for future administrations of the Spanish-language HOS form are discussed.

Development of the English-language HOS questionnaire is discussed elsewhere [1]. The HOS form contains the SF-36 Health Survey, plus additional questions on comorbid conditions, chronic and acute symptoms, and limitations in activities of daily living; a 3-item depression screener; and sociodemographic questions. The SF-36 includes 36 items which are scored as eight multi-item scales measuring: Physical Functioning (PF), Role Limitations due to Physical Health Problems (RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), Role Limitations due to Emotional Problems (RE), and General Mental Health (MH) [2-3]. A one-item measure of change in health over the past year also is included in the SF-36. Physical (PCS) and Mental (MCS) Component Summary measures are calculated from the eight scales [4].

The Spanish-language HOS questionnaire includes the International Quality of Life Assessment (IQOLA) Project translation of the SF-36 [5]. Translation of additional items in the HOS form initially was conducted by New England Research Institute (NERI). Subsequent review of the NERI translation was conducted by the Health Assessment Lab, in conjunction with researchers from California, Florida, Puerto Rico and Texas (who were experienced in the development and testing of health questionnaires for use in Hispanic populations), and a Hispanic linguist who was a major contributor to the development of the US-Spanish SF-36. The revised HOS translation was pilot tested among eleven Spanish-speaking elderly respondents, who did not report problems understanding the form. Probes also were used to make sure that respondents interpreted the questions correctly. Given these positive results, the pilot testing was halted at that point. Further description of the translation and review process is available in a 24-page report prepared by HAL and previously delivered to NCQA [6].

Data

The proportion of Hispanic respondents who had a complete (M10/T10) survey rose from 48.1% in 1998 to 53.6% in 1999 (Table 1). However, the overall response rate for the entire HOS also rose by about 5 percent during this time. It is noteworthy that approximately the same percentage of Hispanic respondents (identified by a HCFA Race code of "5") did not complete the HOS questionnaire due to language problems in 1998 (10.6%) and 1999 (11.0%), despite the availability of a Spanish-language questionnaire during the Cohort II administration.

To better understand the characteristics of the population who did respond to the Spanish-language HOS form in 1999, we compared the Spanish-language group to beneficiaries who were identified as Hispanic during the Cohort I administration in 1998 and completed the English-language form. (English-language Cohort II data was not available when this analysis was conducted). The purpose of this comparison was not to formally test the significance of differences between the two samples, but rather to characterize beneficiaries who responded to the Spanish-language form in 1999. Did these beneficiaries differ from Hispanics who answered the English-language questionnaire in 1998? The two samples were defined as follows.

Spanish-language sample. A total of 209 respondents completed the Spanish HOS questionnaire in 1999. Of these 209 respondents, 46.4% were identified as White, 42.6% were identified as Hispanic, 9.1% were identified as Black and 1.9% were identified as another race, according to the HCFA Race variable. Nearly three-quarters of those respondents who answered the Spanish-language form in 1999 came from four states: California, Florida, New York, and Texas (Table 2). The Spanish-language form was completed by beneficiaries from 53 health plans, including 23 respondents from Maxicare (California), 17 from Health Plan of Nevada, and 14 from FHP of New Mexico. The remaining 50 plans had one to nine respondents each. More than half (53.6%) of the Spanish-language forms were collected by GHS; another 18.7% were collected by MarketFacts. The remaining four vendors had 3.3% to 9.6% of the total Spanish sample. Complete forms (survey disposition of M10) were obtained for 92.8% of respondents. A very small percentage of forms (2.9%) mistakenly were completed by telephone.

English-language Hispanic sample. The English-language group was defined as anyone who had a code of Race=5 (Hispanic) in the Member Level data, or who answered "yes" to the question "Are you of Hispanic or Spanish family background?" on the HOS questionnaire. Of the 8,700 respondents in this group, 8,567 reported that they were of Hispanic family background on the HOS questionnaire. (Only 2,888 of these were identified as Hispanic by the HCFA race variable. The majority, 4,886 or 57%, were identified as white). Another 133 respondents were included because the HCFA race data indicated that they were Hispanic.

In general, the Spanish-language respondents were less educated than the English-language group (61.3% of the Spanish-language respondents and 35.5% of the English-language respondents had eight years or less of education, respectively), had lower household income (35.2% versus 25.5% had income under $10,000), and were more likely to be on Medicaid (26.8% versus 8.8%) (Table 3). The Spanish-language group also had a higher percentage of respondents in the age 18-64 bracket (18.2% versus 9.9%) and were more likely to be entitled to Medicare for reason of disability (18.2% versus 10.5%). This group also was somewhat less likely to self-complete the questionnaire (62.9% versus 74.6%). The two groups did not differ markedly in terms of gender or marital status.

The Spanish-language respondents also reported a greater number of chronic medical conditions (mean=2.7 versus 2.0 for the English-language group); 30% of the Spanish-language group reported four or more chronic conditions compared to 19% of the English-language group (Table 4). The Spanish-language group also was more likely to respond positively to one or more depression screener items (53.6% versus 38.6%), and had a higher proportion of respondents rating their health as "fair" or "poor" compared to others their age (52.4% versus 33.0%). Spanish-language respondents also were less likely to report that they had no difficulty in performing all six activities of daily living included on the HOS questionnaire.

Methods

Tests of Data Quality, Scaling Assumptions and Reliability

Psychometric analyses were used to determine if the assumptions underlying item scoring and the construction of multi-item scales were met in the Spanish HOS data. The goal of this analysis was to confirm that the standard item and scale scoring algorithms used for the English-language SF-36 [3] could be used for the Spanish translation. These tests included data completeness; tests of item-internal consistency and item discriminant validity; and the calculation of scale-level internal consistency reliability [7-8].

Data completeness was evaluated by examining the percent of missing data for each item. A high proportion of missing data may indicate that respondents found an item confusing, offensive, or not applicable, or that there were problems with the translation. A high proportion of missing data for a number of items within a scale may indicate problems with the translation of the response choices for that scale. For this evaluation, the percent of missing data for each item was compared across the Spanish-language and English-language samples.

Three assumptions underlying the construction of multi-item scales were tested in the Spanish-language data, to determine if the items could be grouped into scales and scored as hypothesized under the SF-36 model. First, each item was examined to see if it was substantially linearly related to the scale score computed from all other items in its hypothesized scale (i.e., corrected for overlap between the item and scale) [9]; this is a test of item internal consistency. Item internal consistency was considered substantial and satisfactory if an item correlated 0.40 or more with its hypothesized scale. Item-hypothesized scale correlations also were examined to determine if they were approximately equal within a scale, to enable aggregation without weighting. Finally, the integrity of the hypothesized item groupings was examined using tests of item discriminant validity. If an item is a stronger measure of its hypothesized construct than other constructs, it should have a higher correlation with its hypothesized scale than with other scales; that is, the item should discriminate between scales. A scaling success was counted if an item-hypothesized scale correlation was greater than the correlation between that item and another scale. The significance of the difference between correlations was not taken into account in this analysis, because the number of Spanish HOS records was small. (The standard error of a correlation coefficient decreases as the sample size increases, because the standard error is equal to one divided by the square root of the sample size. In general, the statistical significance of the difference between correlations is not taken into account in tests of item discriminant validity when the sample size is less than 300). The MAP-R for Windows software was used for this analysis [10].

The internal consistency reliability of the eight SF-36 scales was estimated using Cronbach's coefficient alpha [11]. Reliability of measurement refers to the extent to which the measured variance in a scale reflects true score, rather than random error. A minimum reliability coefficient of 0.70 has been suggested for group-level comparisons, while reliability coefficients of 0.90 or 0.95 have been suggested for individual analyses [12].

<u>Psychometric and Clinical Validity</u>

The number of Spanish-language respondents was quite small, and thus the extent of the validity analysis was limited. However, validity analyses were conducted to begin to address the issues of interpreting the eight SF-36 scales and scoring the PCS and MCS in the Spanish data. The psychometric and clinical validity of the eight SF-36 scales was examined following techniques previously used to evaluate the validity of the original English SF-36 and SF-36 translations [13-15]. Specifically, two lines of analyses were pursued. First, the construction and scoring of the PCS and MCS for the U.S.-English questionnaire is based on results of factor analyses that show that there are two (physical and mental) factors in the SF-36. Thus, we repeated the factor analysis previously conducted on English-language data [4, 13] on the Spanish-language HOS data, to see how well the previously observed factor structure was replicated in the Spanish data. Second, the initial interpretation of the PCS and MCS as physical and mental measures, respectively, is also based on findings that the PCS and MCS were valid in discriminating between physical and mental health conditions in cross-sectional and longitudinal tests [4, 15]. Thus, we conducted some preliminary analyses to determine how well the PCS and MCS, and the eight SF-36 scales, discriminated among groups known to differ in external physical and mental criteria. The results of these latter analysis would begin to enable us to interpret the eight scales, and the two summary measures, as measuring physical or mental constructs in the Spanish questionnaire.

More specifically, factor analysis was conducted to evaluate how well the factor structure seen in previous U.S. and European studies was replicated in the Spanish data [4, 13, 16]. Two principal components were extracted from the eight SF-36 scales and rotated to orthogonal simple structure using the varimax method, to facilitate comparisons with published results and for ease of interpretation. We examined whether the eigenvalues for the first two components were greater than unity, as suggested for rotation. Second, we evaluated the proportion of the total variance in SF-36 scale scores explained by the two principal components, to determine if this was 60% or greater, as seen in previous studies. Third, we evaluated the pattern of correlations between the eight SF-36 scales and two rotated components to determine the basis for their interpretation as physical and mental components. A pattern of correlations first seen in the U.S. has been replicated in many Western European countries [16]. Correlations with the PCS are highest for the most physical scales (PF, RP, BP) and lowest for the most mental scales (MH, RE). For the MCS, the opposite pattern is observed; MCS correlations are highest with the most mental scales and lowest for the most physical scales. Factor analysis of the Spanish-language data was conducted to see if there was a similar pattern of correlations. Finally, we looked to see if the proportion of total variance in each SF-36 scale explained by the two factors was above 0.60, as is generally observed. The analysis was conducted on the entire sample, and by gender, age, and educational level.

While factor analysis is a traditional psychometric approach to validation, in which the congruence between hypothesized constructs and the scales constructed to measures these constructs is evaluated, tests of clinical validity – that is the relation of SF-36 scales to external measures – also are important in evaluating the underlying interpretation of the scales, and ultimately, the scoring of the physical and mental summary measures [13-14]. Very preliminary comparisons were conducted between Spanish-language groups differing in self-reported medical conditions and on the depression screener. Four groups were constructed: (1) respondents who did not report any of 12 medical conditions (all of the conditions in the chronic condition checklist, excluding hypertension. Ideally respondents who had hypertension also would have been excluded from this group, but doing so would have reduced the sample size for Group 1 considerably. Because hypertension has been shown to have a relatively small impact on most SF-36 scale scores, respondent who had hypertension only were left in this group); (2) respondents who reported that they had angina, congestive heart failure, a stroke, or arthritis of the hip or knee (all conditions with a demonstrated impact on physical health), and who did not screen positively on the depression screener; (3) respondents who screened positively on the depression screener ("yes" to any question) but did not have any of the 12 medical conditions; and (4) respondents who had one or more of the physical conditions in Group 2 and who screened positively on the depression screener. We hypothesized that respondents in Group 2 would score lower on predominantly physical measures (e.g., PF, RP, BP, PCS) than respondents in either Groups 1 or 3; and that respondents in Group 3 would score lower on predominantly mental measures (e.g., MH, RE, SF, MCS) than respondents in Groups 1 or 2. Respondents in Group 4 were expected to have low scores on both physical and mental measures, relative to Group 1.

Calculation of Summary Measures and SF-36 Descriptive Statistics

We calculated scores for the Physical Component Summary (PCS) and Mental Component Summary (MCS), using standard scoring algorithms [4]. The reliability of the PCS and MCS was calculated, following techniques used to calculate the reliability of the summary measures in U.S.-English data [4]. We also compared unadjusted scores for the Spanish-language and English-language Hispanic groups across all eight scales and the two summary measures. The purpose of this analysis was primarily to characterize the Spanish-language population; that is, to determine if beneficiaries who responded to the Spanish-language form had significantly different health status scores on the SF-36 than Hispanic beneficiaries who completed the form in English.

Results

Tests of Scaling Assumptions and Scale-Level Reliability

The Spanish SF-36 data generally met scaling assumptions. The rate of missing data per item was somewhat higher in the Spanish-language sample than in the English-language Hispanic group (Table 5, Spanish-language data is summarized in Table 7). There was a higher percentage of missing data for items which were included in grids (e.g., the physical functioning items) as opposed to items which were printed individually (e.g., the bodily pain items). Of the 209 Spanish-language respondents, one or more SF-36 scale scores could not be calculated for 24 respondents, and the health transition item was missing for an additional respondent. These 25 respondents were not included in the tests of scaling assumptions and reliability.

Item-internal consistency (the correlation of an item with its hypothesized scale) was above the 0.40 level recommended as a minimum to demonstrate a linear relationship between an item and its underlying construct, for all items (Table 6, summarized in Table 7). In general, items within a scale had roughly equivalent item-scale correlations. There were some exceptions, which have been seen in previous studies. Within the Physical Functioning scale, the most difficult (Vigorous Activities) and least difficult (Bathing/Dressing) items had lower item-scale correlations than the other Physical Functioning items, which measure activities which are more similar in difficulty. The two General Health items which ask respondents to rate their health as excellent to poor, and to agree or disagree with the statement that their health was excellent, had higher item-scale correlations than the other General Health items. Within the Mental Health scale, the two positive affect items (calm and peaceful; happy person) had lower item-scale correlations than the other three items which measure negative affect.

Items generally had higher correlations with their hypothesized scales than with other scales (Table 6, summarized in Table 7). There were a few exceptions to this finding, for one Mental Health, one Vitality, and two General Health items. The Mental Health item "Happy person" and the Vitality item "Lot of energy" had high correlations with both the Mental Health and Vitality scales. Similar results have been seen in some other foreign-language populations. Two General Health items ("Get sick easier than others" and "As healthy as anyone I know") also had high correlations with other scales. Because general health is a concept that is meant to incorporate respondents' perceptions of their overall health, the General Health items often have high correlations with other scales.

Internal consistency reliability (as measured by Cronbach's coefficient alpha) ranged from 0.84 to 0.94, and met accepted minimum standards for group-level comparisons (Table 7). Reliability was greater than 0.90 for the Physical Functioning and Role Physical scales, and approached 0.95 for the Physical Functioning scale. A minimum level of 0.90 has been recommended for person-level analyses.

<u>Psychometric and Clinical Validity</u>

Principal components analysis of the Spanish-language data was conducted overall and by gender, age and education level. Results by gender and age did not differ from the overall results, and are not reported here.

For the sample as a whole, there was an extremely weak second factor, which had an eigenvalue well below 1.0 (at 0.65) and percentage of total variance explained below 10% (Table 8). Thus, conventional guidelines would indicate that there was one general health factor in the data. However, when two factors were forced, the factor loadings (correlations between a scale and a factor) indicated that there was an overall health factor and a role limitations factor in the total sample (see columns titled "Factor 1" and "Factor 2" under the "All Respondents" portion of the table. Higher factor loadings (closer to 1.0) indicate that a scale is loading more highly on a factor (i.e., that the correlation between that scale and the principal component is greater)). The Bodily Pain, General Health, Vitality, Social Functioning, and Mental Health scales loaded highly on one factor, while the Role Physical and Role Emotional scales loaded highly on a second factor. The Physical Functioning scale had similar loadings on both factors.

Results for the least educated group (8^{th} grade or less) were similar to those for the total sample. The group with some high school education had mixed results, although there was some emergence of separate physical and mental factors. However, results for Spanish-language respondents who had completed high school or had education beyond high school were similar to those seen in the general U.S. (English-speaking) population and in Western European populations. For the most educated Spanish-language group, there were two factors with an eigenvalue of 1.0 or greater, and both factors explained more than 10 percent of the total variance. In addition, the factor loadings were similar to those for the English-speaking population in the U.S. There was a gradient of factor loadings going from the least mental scale (PF) to the most mental scale (MH) for one factor, and an opposite gradient for the other factor.

The proportion of variance in the SF-36 scales explained by the two factors generally was above 0.70, although it was low for the Physical Functioning scale in the total sample and the two less educated samples.

Many hypotheses in the tests of clinical validity were confirmed in the data (Table 9). Respondents who did not report any of the 12 medical conditions and did not screen positive for depression had the highest mean scores on all eight scales and the two summary measures.

Respondents who had one or more medical conditions and screened positive for depression had the lowest scores on all measures. The Medical Only group did score lower on the PCS than the Depression Only group, and the Depression Only group scored lower than the Medical Only group on the MCS, as hypothesized. However, mean scores on the eight SF-36 scales for the Medical Only and Depression Only groups did not always fit hypothesized patterns, particularly for the Physical Functioning and Role Emotional scales. These results provide some indication that the SF-36 summary measures are measuring physical and mental health, as hypothesized, in the Spanish-language population, but also raise some questions about the interpretation of some individual scales. However, given the extremely small size of the sample, particularly of Group 3, these results need to be examined in other senior populations.

Calculation of Summary Measures and SF-36 Descriptive Statistics

When results of validity tests are mixed (with the clinical validity tests generally supporting the hypothesized interpretation of scales as physical or mental, and the factor analysis presenting mixed results within subgroups of the population), we generally recommend scoring the PCS and MCS using standard scoring algorithms. In general, we also recommend using language-specific scoring algorithms, and monitoring results using both scoring methods across a number of datasets, to better understand the interpretation of the scales as physical or mental measures. However, the HOS Spanish-language sample is too small to derive language-specific scoring of the PCS and MCS. Therefore, at this time we just recommend using the standard scoring algorithms for the PCS and MCS. Using these algorithms, the reliability of the PCS was 0.93 and the MCS was 0.90 in the Spanish-language data.

Unadjusted scores on the Physical Component Summary (PCS) and Mental Component Summary (MCS) were compared across the Spanish-language Cohort II data and a 10% random sample of the Hispanic English-language Cohort I data. As would be expected given the sociodemographic and self-reported clinical makeup of the two samples, the Spanish-language respondents scored significantly lower on all eight SF-36 scales and the two summary measures (Table 10). This was true for the total sample, and for samples limited to respondents aged 65 and older.

Discussion and Recommendations

According to the 1990 Census, approximately 17.3 million Americans spoke Spanish at home. Of these 17.3 million, 8.5% did not speak English at all, and another 17.5% did not speak English well [17]. Although these statistics are outdated, they do give some indication of the proportion of Hispanics who may find it impossible to respond to an English-language HOS questionnaire. Researchers also have noted that even among English-speaking Hispanics who have achieved a high degree of acculturation, a notable percent prefer to use Spanish when completing a survey [18].

Although a Spanish-language version of the Health Outcomes Survey was offered in 1999, the proportion of Hispanic Medicare beneficiaries who did not answer the HOS due to language problems did not decline between 1998 and 1999. Approximately 11 percent of Hispanic respondents did not complete the survey due to language problems both years. If even a portion of this group had answered the Spanish-language form, the response rates among Hispanics would have begun to approach those for other ethnic groups. In Cohort I (1998), the percentage of Hispanic respondents with any type of complete survey (Complete or Partial Complete) was 51 percent, compared to 65-66 percent for White and Asian respondents and 54 percent for Black respondents. Results were similar for Cohort II (1999), in which 57 percent of Hispanic, 59 percent of Black and 69 percent of White and Asian respondents had any type of complete (Complete or Partial Complete) survey.

A second notable finding is that Cohort II respondents who did complete the Spanish-language survey tended to be different than Hispanic respondents who completed the English-language survey during Cohort I. On average, the Spanish-language respondents were less educated, had lower household income, were more likely to be on Medicaid, reported more chronic medical conditions and limitations in activities of daily living, were more likely to screen positively for depression, and had lower SF-36 scores. In other words, the Spanish-language sample was sicker and had lower socioeconomic status, on average.

Particularly in light of the differences among Hispanic respondents completing Spanish and English versions of the HOS, efforts should be made to encourage greater use of the Spanish HOS, if it is possible to do so. These efforts could include development of a telephone version of the Spanish form, and including both English and Spanish-language forms in mailings to respondents identified as Hispanic in the HCFA enrollment data. In addition, if major revisions of the pre-notification postcard and survey letters are being made in subsequent survey administrations, we suggest that one reviewer be a specialist in conducting research among Hispanic populations. It would be preferable if this person had experience in conducting federal government surveys. There may be some minor changes that could be made to the survey letter that might increase respondent comfort with the survey. For example, the cover letter includes a sentence "after the study is completed, your individual responses will be made available to your health plan", which may be threatening to some respondents. Some Hispanics have come to the United States from countries in which governments have used information about individuals to violate human rights, and thus may be somewhat wary of this sort of information sharing. Thus, minor changes to the wording of the HOS document may affect response rates.

While the amount of Spanish-language HOS data from Cohort II was relatively small, the psychometric properties of the form were satisfactory overall. The rate of missing data was somewhat higher in the Spanish data than in English-language groups, but this may reflect the lower average educational level in the Spanish sample. Missing data tends to be higher among less educated populations. Other psychometric criteria, such as item internal consistency, item discriminant validity, and internal consistency reliability, generally were met in the Spanish data. Thus, standard scoring algorithms for the eight SF-36 scales can be used for the Spanish HOS form. However, the amount of missing data, tests of scaling assumptions, and reliability of the eight SF-36 scales should continue to be monitored.

Results from factor analysis and tests of clinical validity were somewhat mixed. On the one hand, SF-36 summary measures showed expected relationships to external criteria, although the sample size for one group (Depression Only) was extremely small. Respondents who reported having one or more medical conditions scored lower on the PCS than a control group that did not report any of 12 chronic conditions or depression. Respondents who reported depression only scored higher on the PCS than the group with medical conditions only. Respondents who reported medical conditions and screened positive for depression had the lowest PCS scores, as would be expected given the more complicated disease situation of these respondents. Similarly, respondents who screened positive for depression only scored lower on the MCS than the healthy control group or the group with medical conditions only. The group with both medical conditions and a positive screen for depression scored lowest on the MCS. Thus, in many respects the PCS and MCS are performing as would be expected in the Spanish-language group.

However, results for some individual SF-36 scales did not fit expected patterns. The results of the factor analysis on the less educated Spanish-language groups were not similar to those for English-language respondents in the U.S. or for the more educated Spanish-language group. In addition, the Medical Only group and the Depression Only group had nearly identical mean scores on the Role Emotional scale. If Role Emotional were a strong measure of mental health, the RE scores for the Depression Only group would be expected to be lower than those for the Medical Only group. These results are based on very small sample sizes (especially for the Depression Only group), and thus no conclusions about the interpretation of individual scales should be made on the basis of this data. However, these results do raise questions to be examined in future analyses. Do respondents who complete the HOS form in Spanish and who have less education continue to show a different pattern of results in factor analyses, compared to more educated Hispanic respondents? What would SF-36 scale scores be in a larger group of respondents who do not have any reported medical conditions, but who screen positive for depression? Would the Role Emotional scores for this group continue to be similar to scores for a group with medical conditions only, indicating that respondents may not be making a distinction between role limitations due to physical and emotional problems?

The results for the Spanish-language sample are not unique, and should not be cause for concern at this point. There have been some indications, primarily from a number of unpublished international datasets, that there may be a link between how respondents think about some SF-36 items and their educational levels, language, and general exposure to Western philosophical ideas. (One published study that discusses these issues is Fukuhara et al. [19]). Findings from the U.S. and Western Europe may not generalize across all cultures. For example, if respondents are not used to making a distinction between limitations in their work due to physical problems or emotional problems, or if cultural norms indicate that work should <u>not</u> be limited by emotional problems, then the Role Emotional items may not be a good measure of mental health. Those Hispanic respondents who are answering the HOS in Spanish, and who have little formal education, may be viewing health and health's relationship to life, in a different way than Hispanics who are responding to the HOS in English, or who have more formal education. The latter group may have made a cultural shift in thinking that the former group has not made.

Because all of the results from the tests of psychometric and clinical validity did not conform to hypotheses, it is unclear at this point whether the PCS and MCS can be interpreted in the Spanish-language group in exactly the same manner as in the general English-speaking U.S. population. The amount of data available for this analysis was limited, and thus all conclusions are preliminary. In future administrations of the Spanish HOS questionnaire, additional analyses of the psychometric and clinical validity of the data should be conducted, to better understand the interpretation of the eight scales as physical and mental measures of health. For the current time, we recommend continuing to score the PCS and MCS using standard scoring algorithms. However, results from future analyses of the psychometric and clinical validity of the Spanish translation may have implications for the scoring and interpretation of the PCS and MCS summary measures in this population.

Table 1 – HOS Survey Disposition Variable for Hispanic Respondents, Cohort I and II

Code	Disposition	Cohort I (n=5960)	Cohort II (n=6893)
M10/T10	Complete survey	48.1%	53.6%
M11/T11	Partial complete survey	2.6	3.4
M20/T20	Deceased	0.7	0.6
M21/T21	Ineligible: Not in HMO	0.2	0.2
M22/T22	Ineligible: ESRD	0.1	0.0
M23/T23	Ineligible: Language problem	10.6	11.0
T24	Ineligible: No address/phone	0.8	1.2
M31/T31	Nonresponse: breakoff	0.6	0.5
M32/T32	Nonresponse: refusal	8.6	8.3
M33/T33	Nonresponse: unavailable	0.5	1.8
M34/T34	Nonresponse: incapacitated	0.6	0.7
M35/T35	Nonresponse: institutionalized	0.0	0.2
M36/T36	Nonresponse: maximum attempts	26.6	18.5

Note: Hispanic respondents are defined as having a HCFA Member Level Record Race variable equal to “5”.

Table 2 – Geographic and Survey Characteristics of Respondents to the Spanish-Language HOS (n=209)

	Percent
State	
Arizona	1.9
California	34.9
Florida	12.9
Illinois	4.3
Massachusetts	0.5
Nevada	1.0
New Jersey	7.2
New Mexico	2.4
New York	11.0
Oregon	0.5
Pennsylvania	8.1
Puerto Rico	1.0
Texas	14.4
Vendor	
DSS	5.7
GHS	53.6
HCIA	9.6
Market Facts	18.7
NRC	3.3
Response Analysis	9.1
Survey Disposition	
M10 – complete survey	92.8
M11 – partial complete survey	6.7
M31 – Nonresponse: break-off	0.5
Round Survey Completed	
1st mailing	71.3
2nd mailing	19.1
1st telephone	1.9
2nd telephone	1.0
Not completed	6.7

Table 3 – Sociodemographic Characteristics of Spanish-Language Cohort II and Hispanic English-Language Cohort I Respondents

	Spanish-language respondents, Cohort II (n=209)	English-language Hispanic respondents, Cohort I (n=8700)
Mean age (SD)	68.3 (9.0)	71.0 (8.3)
18-64	18.2%	9.9%
65-74	60.7	60.7
75+	21.1	29.4
% Female	52.6%	54.5%
Education:		
8th grade or less	61.3%	35.5%
Some high school, did not graduate	21.0	20.1
High school graduate/GED	9.1	23.9
Some college/2 year degree	3.8	13.1
4 year college degree	2.7	3.5
More than 4 years of college	2.2	3.9
Marital Status		
Married	58.3%	59.8%
Divorced	10.2	10.4
Separated	5.8	2.1
Widowed	20.9	23.6
Never married	4.9	4.2
Household Income		
Less than $5,000	8.2%	7.6%
$5,000-$9,999	27.0	17.9
$10,000-$19,999	29.1	30.0
$20,000-$29,999	9.2	15.6
$30,000-$39,999	6.1	7.2
$40,000-$49,999	0.5	3.7
$50,000-$79,999	1.0	3.0
$80,000-$99,999	0.0	0.5
$100,000 or more	0.0	0.5
Don’t know	18.9	14.0
On Medicaid	26.8%	8.8%
Reason for Entitlement		
Aged without ESRD	81.8%	89.5%
Disabled without ESRD	18.2	10.5
Who Completed Survey?		
Person sampled	62.9%	74.6%
Family member/relative	33.1	22.8
Friend	2.9	1.9
Professional caregiver	1.1	0.7

Table 4 – Selected Health/ADL Characteristics of Spanish-Language Cohort II Respondents and Hispanic English-Language Cohort I Respondents

	Spanish-language respondents, Cohort II (n=209)	English-language Hispanic respondents, Cohort I (n=8700)
Mean # of medical conditions (SD)	2.7 (2.0)	2.0 (1.9)
Number of medical conditions		
0	15.5%	23.7%
1	16.4	22.7
2	18.4	19.8
3	19.8	14.7
4+	29.9	19.1
Medical Conditions:		
Hypertension	59.2%	51.5%
Angina/coronary artery disease	15.9	13.7
Congestive heart failure	13.2	6.8
Myocardial infarction	9.9	9.2
Other heart conditions	20.2	16.0
Stroke	5.9	8.3
Emphysema, asthma, COPD	12.1	10.9
Crohn's/colitis/inflammatory bowel	21.0	6.3
Arthritis of hip or knee	56.6	39.5
Arthritis of hand or wrist	46.8	36.1
Sciatica	40.3	25.6
Diabetes	27.2	25.4
Any cancer (other than skin cancer)	6.9	9.1
Positive answer to any depression screener question	53.6%	38.6%
No difficulty in ADL's		
Bathing	70.7%	82.7%
Dressing	73.3	84.0
Eating	85.6	90.8
Getting in or out of chairs	56.3	72.7
Walking	54.2	64.1
Using the toilet	82.0	88.6
Health compared to others		
Excellent	10.2%	11.8%
Very good	6.8	24.9
Good	30.6	30.4
Fair	36.9	24.6
Poor	15.5	8.4

Table 5 – Percent of Missing Data by Item: Spanish-Language Cohort II Respondents and English-Language Hispanic Cohort I Respondents

Label	Item content	Spanish-language respondents (n=209)	English-language Hispanic respondents (n=8700)
PF01	Vigorous activities	2.4	0.5
PF02	Moderate activities	2.9	0.5
PF03	Lifting groceries	2.4	0.5
PF04	Climb several flight	8.6	0.5
PF05	Climb one flight	4.8	0.5
PF06	Bending, kneeling	3.8	0.5
PF07	Walk mile	3.8	0.5
PF08	Walk several blocks	5.3	0.5
PF09	Walk one block	5.3	0.5
PF10	Bath dress	2.9	0.5
RP1	Cut down time	6.2	2.5
RP2	Less done	7.2	2.5
RP3	Limited in kind	8.6	2.5
RP4	Difficulty in work	6.2	2.5
BP1	Pain severity	1.9	1.9
BP2	Pain limitation	1.9	2.4
GH1	EVGFP	2.9	0.7
GH2	Sick easier others	3.8	0.0
GH3	Healthy as anyone	4.8	3.2
GH4	Expect worse health	4.3	0.0
GH5	Excellent health	4.8	2.8
VT1	Pep	5.3	4.9
VT2	Energy	4.8	2.8
VT3	Worn out	7.7	1.6
VT4	Tired	2.9	1.6
SF1	Social extent	1.9	2.3
SF2	Social frequency	4.3	1.4
RE1	Cut down time	6.2	3.1
RE2	Less done	4.3	3.1
RE3	Less careful	5.7	3.1
MH1	Nervous	3.3	1.5
MH2	Down in dumps	3.3	1.5
MH3	Calm and peaceful	5.7	3.1
MH4	Blue	6.2	1.5
MH5	Happy	6.7	2.5

PF= Physical Functioning
RP=Role Limitations due to Physical Health
BP=Bodily Pain
GH=General Health Perceptions
VT=Vitality
SF=Social Functioning
RE=Role Limitations Emotional
MH=Mental Health

Table 6 – Item Descriptive Statistics and Pearson Item-Scale Correlations Corrected for Overlap: Spanish-Language Cohort II Respondents (n=184)

Label	Item content	Mean	Std Dev	Pearson Item-Scale Correlation							
				PF	RP	BP	GH	VT	SF	RE	MH
PF01	Vigorous activities	1.54	0.67	0.55*	0.46	0.41	0.50	0.50	0.40	0.33	0.34
PF02	Moderate activities	1.85	0.74	0.77*	0.51	0.49	0.47	0.53	0.51	0.41	0.40
PF03	Lifting groceries	2.05	0.75	0.80*	0.49	0.51	0.49	0.54	0.51	0.40	0.44
PF04	Climb several flight	1.70	0.71	0.80*	0.47	0.51	0.48	0.57	0.50	0.38	0.43
PF05	Climb one flight	2.05	0.75	0.77*	0.46	0.51	0.49	0.52	0.48	0.36	0.45
PF06	Bending, kneeling	1.76	0.75	0.70*	0.42	0.48	0.39	0.52	0.46	0.32	0.38
PF07	Walk mile	1.91	0.84	0.85*	0.57	0.56	0.51	0.61	0.53	0.47	0.47
PF08	Walk several blocks	1.92	0.83	0.87*	0.56	0.52	0.52	0.58	0.53	0.47	0.46
PF09	Walk one block	2.29	0.75	0.74*	0.38	0.46	0.44	0.48	0.47	0.36	0.44
PF10	Bath dress	2.46	0.73	0.60*	0.33	0.49	0.37	0.42	0.44	0.31	0.40
RP1	Cut down time	1.34	0.48	0.47	0.78*	0.55	0.48	0.51	0.53	0.62	0.46
RP2	Less done	1.29	0.45	0.51	0.83*	0.58	0.50	0.52	0.47	0.61	0.43
RP3	Limited in kind	1.29	0.45	0.57	0.88*	0.57	0.53	0.57	0.53	0.58	0.46
RP4	Difficulty in work	1.29	0.46	0.58	0.87*	0.63	0.58	0.64	0.55	0.58	0.51
BP1	Pain severity	3.39	1.52	0.54	0.59	0.81*	0.69	0.71	0.64	0.52	0.67
BP2	Pain limitation	3.04	1.29	0.64	0.62	0.81*	0.63	0.69	0.68	0.49	0.64
GH1	EVGFP	2.35	0.91	0.54	0.49	0.60	0.72*	0.65	0.59	0.41	0.59
GH2	Sick easier others	3.13	1.39	0.46	0.42	0.56	0.58*	0.58	0.47	0.41	0.60
GH3	Healthy as anyone	2.70	1.17	0.47	0.41	0.51	0.62*	0.64	0.57	0.44	0.59
GH4	Expect worse health	2.99	1.16	0.37	0.41	0.50	0.66*	0.50	0.50	0.35	0.52
GH5	Excellent health	2.47	1.36	0.46	0.54	0.59	0.71*	0.68	0.58	0.51	0.62
VT1	Pep	3.54	1.53	0.55	0.50	0.61	0.62	0.69*	0.59	0.52	0.65
VT2	Energy	3.40	1.50	0.57	0.51	0.60	0.65	0.70*	0.57	0.48	0.71
VT3	Worn out	3.78	1.41	0.56	0.53	0.67	0.69	0.78*	0.63	0.49	0.76
VT4	Tired	3.36	1.37	0.56	0.55	0.63	0.67	0.76*	0.60	0.46	0.72
SF1	Social extent	3.52	1.36	0.56	0.55	0.64	0.63	0.64	0.74*	0.62	0.65
SF2	Social frequency	3.29	1.20	0.57	0.52	0.65	0.64	0.67	0.74*	0.56	0.67
RE1	Cut down time	1.44	0.50	0.43	0.56	0.46	0.44	0.50	0.58	0.74*	0.53
RE2	Less done	1.39	0.49	0.50	0.65	0.57	0.57	0.60	0.63	0.79*	0.62
RE3	Less careful	1.57	0.50	0.33	0.51	0.37	0.40	0.41	0.45	0.64*	0.46
MH1	Nervous	3.91	1.46	0.36	0.38	0.58	0.56	0.58	0.50	0.42	0.74*
MH2	Down in dumps	4.27	1.54	0.50	0.48	0.63	0.67	0.73	0.68	0.58	0.79*
MH3	Calm and peaceful	3.77	1.46	0.47	0.43	0.54	0.61	0.67	0.58	0.52	0.69*
MH4	Blue	4.09	1.48	0.46	0.44	0.57	0.65	0.78	0.64	0.53	0.81*
MH5	Happy	3.74	1.49	0.42	0.41	0.57	0.63	0.71	0.55	0.51	0.68*

Table 7 – Summary of Tests of Scaling Assumptions and Reliability: Spanish-Language Cohort II Respondents (n=184)

	% Missing Data (Range Across Items Within Scale)	Item Internal Consistency (Range of Correlations)[a]	Item Discriminant Validity (Percent of Scaling Successes)[b]	Internal Consistency Reliability[c]
PF	2.4-8.6%	0.55-0.87	100.0	0.94
RP	6.2-8.6	0.78-0.88	100.0	0.93
BP	1.9	0.81	100.0	0.89
GH	2.9-4.8	0.58-0.72	95.0	0.84
VT	2.9-7.7	0.69-0.78	96.9	0.87
SF	1.9-4.3	0.74	100.0	0.84
RE	4.3-6.2	0.64-0.79	100.0	0.85
MH	3.3-6.7	0.68-0.81	97.5	0.89

[a] Correlations between items and the scale to which they are expected to belong, corrected for overlap.
[b] Percent of item-hypothesized scale correlations greater than item-competing scale correlations.
[c] Cronbach's coefficient alpha.

PF= Physical Functioning
RP=Role Limitations due to Physical Health
BP=Bodily Pain
GH=General Health Perceptions
VT=Vitality
SF=Social Functioning
RE=Role Limitations due to Emotional Problems
MH=Mental Health

Table 8 – Summary of Results from Principal Components Analysis of Spanish-Language Data, Overall and by Education

	All Spanish-Language Respondents (n=185)			8th Grade or Less (n=99)			Some High School (n=38)			High School Graduate or Greater (n=29)		
	Rotated Factor Loadings			Rotated Factor Loadings			Rotated Factor Loadings			Rotated Factor Loadings		
Scale	Factor 1	Factor 2	h^2	Factor 1	Factor 2	h^2	Factor 1	Factor 2	h^2	Factor 1	Factor 2	h^2
PF	0.55	0.53	0.58	0.72	0.28	0.59	0.62	0.58	0.72	0.18	0.92	0.87
RP	0.32	0.85	0.83	0.39	0.78	0.77	0.26	0.90	0.87	0.32	0.81	0.77
BP	0.72	0.45	0.72	0.76	0.34	0.70	0.58	0.68	0.80	0.59	0.67	0.80
GH	0.83	0.31	0.79	0.85	0.25	0.78	0.49	0.76	0.81	0.84	0.28	0.78
VT	0.85	0.36	0.86	0.85	0.36	0.85	0.65	0.63	0.82	0.78	0.41	0.77
SF	0.69	0.49	0.72	0.66	0.50	0.68	0.84	0.31	0.80	0.77	0.44	0.78
RE	0.33	0.81	0.76	0.27	0.90	0.88	0.78	0.33	0.71	0.73	0.27	0.61
MH	0.86	0.29	0.82	0.78	0.35	0.75	0.80	0.42	0.82	0.95	0.14	0.91
Unrotated Eigenvalue	5.44	0.65		5.29	0.71		5.81	0.56		5.29	1.00	
Total Variance Explained	68.0%	8.1%		66.1%	8.9%		72.6%	7.0%		66.1%	12.5%	

Note: The rotated factor loading is the correlation between each SF-36 scale and the rotated principal component. h^2 is the proportion of total variance in each SF-36 scale explained by the two principal components.

Table 9 – Mean SF-36 Scores (Standard Error) for Groups Differing in Self-Reported Medical and Psychiatric Conditions, Spanish-Language Cohort II Respondents

	No Reported Condition (n=21)[a]	Medical Condition Only (n=47)[b]	Depression Screener Only (n=6)[c]	Medical and Depression (n=68)[d]
PF	74.1	54.0	40.0	33.7
	(5.9)	(3.9)	(11.0)	(3.3)
RP	65.1	33.2	62.5	7.7
	(7.4)	(5.0)	(13.8)	(4.1)
BP	81.0	51.2	65.8	27.3
	(4.6)	(3.1)	(8.7)	(2.6)
GH	69.0	50.2	45.3	28.4
	(4.1)	(2.7)	(7.6)	(2.3)
VT	75.5	57.5	50.8	34.9
	(4.2)	(2.8)	(7.8)	(2.3)
SF	86.3	75.0	62.5	38.2
	(5.1)	(3.4)	(9.5)	(2.8)
RE	82.5	67.4	66.7	21.1
	(7.7)	(5.1)	(14.4)	(4.2)
MH	81.0	71.4	56.7	41.2
	(4.1)	(2.7)	(7.6)	(2.3)
PCS	46.1	34.6	38.2	28.5
	(2.0)	(1.3)	(3.8)	(1.1)
MCS	55.2	51.8	45.2	34.7
	(2.1)	(1.4)	(3.9)	(1.2)

Note: Any interpretation of results should take the small sample size into account.

[a] No condition: Responded “no” to all of the following: angina, CHF, MI, other heart disease, stroke, emphysema/asthma/COPD, GI disorders, arthritis of hip/knee, arthritis of wrist/hand, sciatica, diabetes, cancer.

[b] Medical condition: Responded “yes” to angina, CHF, stroke, and/or arthritis of hip or knee.

[c] Depression screener: Responded “yes” to any depression screener items.

[d] Medical and depression: Has a Medical Condition and answered “yes” to any of the items on the depression screener.

Table 10 – Unadjusted Mean SF-36 Scores (Standard Error) for Spanish-Language Cohort II Respondents and Hispanic Cohort I English-Language Respondents

	Total Sample		Age 65+	
	Spanish-Language Respondents (n=209)	English-Language 10% Random Sample (n=883)	Spanish-Language Respondents (n=171)	English-language 10% Random Sample (n=792)
PF	47.8 (2.1)	60.5 (1.0)	51.7 (2.3)	62.9 (1.1)
RP	30.6 (3.0)	53.8 (1.5)	35.1 (3.3)	57.3 (1.6)
BP	46.5 (2.0)	60.1 (1.0)	50.2 (2.2)	62.9 (1.0)
GH	42.9 (1.7)	57.6 (0.8)	47.2 (1.8)	60.7 (0.8)
VT	50.1 (1.7)	55.7 (0.8)	53.7 (1.9)	58.4 (0.8)
SF	59.9 (2.1)	71.9 (1.0)	63.8 (2.3)	74.9 (1.0)
RE	45.8 (3.1)	64.9 (1.5)	51.0 (3.4)	68.0 (1.5)
MH	58.2 (1.8)	71.6 (0.7)	62.0 (1.8)	73.8 (0.7)
PCS	34.5 (0.8)	40.2 (0.4)	35.8 (0.9)	41.3 (0.4)
MCS	44.2 (0.9)	49.1 (0.4)	45.8 (1.0)	50.2 (0.4)

Note: Spanish and English-language groups differ significantly for all measures at $p<0.0001$, with the exception of Vitality ($p<0.01$ for Total Sample, $p<0.05$ for Age 65+).

REFERENCES

1. National Committee for Quality Assurance. *Medicare Health Outcomes Survey Manual. HEDIS® 1999 Volume 6.* Washington DC: National Committee for Quality Assurance, 1999.

2. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Med Care* 1992; 30:473-483.

3. Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey manual and interpretation guide*. Boston: The Health Institute, 1993.

4. Ware JE, Kosinski M, Keller SD. *SF-36 physical and mental health summary scales: A user's manual*. Boston: The Health Institute, 1994.

5. Gandek B, Ware JE. Translating functional health and well-being: International Quality of Life Assessment (IQOLA) Project studies of the SF-36 Health Survey. *J Clin Epidemiol* 1998; 11:891-1214.

6. Health Assessment Lab. *Linguistic review and pilot testing of the Spanish translation of the Health Outcomes Survey*. Report prepared for the Health Care Financing Administration and the National Committee for Quality Assurance. Boston, MA: Health Assessment Lab, March 1999.

7. McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* 1994; 32:40-66.

8. Ware JE, Gandek B. Methods for testing data quality, scaling assumptions, and reliability: The IQOLA Project approach. *J Clin Epidemiol* 1998; 11:945-952.

9. Howard KI, Forehand GG. A method for correcting item-total correlations for the effect of relevant item inclusion. *Educ Psychol Measmt* 1962; 22:731-735.

10. Ware JE, Harris WJ, Gandek B, Rogers BW, Reese PR. *MAP-R for Windows: Multitrait/multi-item analysis program-revised user's guide*. Boston, MA: Health Assessment Lab, 1997.

11. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16:297-334.

12. Nunnally JC, Bernstein IR. *Psychometric Theory, 3rd Edition*. New York: McGraw-Hill, 1994.

13. McHorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993; 31: 247-263.

14. Gandek B, Ware JE. Methods for validating and norming translations of health status questionnaires: The IQOLA Project approach. *J Clin Epidemiol* 1998; 11: 953-959.

15. Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: Summary of results from the Medical Outcomes Study. *Med Care* 1995; 33(Suppl. 4): AS264-AS279.

16. Ware JE, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, et al. The factor structure of the SF-36 Health Survey in 10 countries: Results from the IQOLA Project. *J Clin Epidemiol* 1998; 11: 1159-1165.

17. U.S. Census Bureau, Population Division, Education and Social Stratification Branch. *1990 United States Census, Language Use, Table 5: Detailed Language Spoken at Home and Ability to Speak English for Persons 5 Years and Over --50 Languages with Greatest Number of Speakers*. (http://www.census.gov:80/population/socdemo/language/table5.txt)

18. Marín G, Marín BV. *Research with Hispanic populations*. Applied Social Research Methods Series, Volume 23. Newbury Park, CA: Sage Publications, 1991.

19. Fukuhara S, Ware JE, Kosinski M, Wada S, Gandek B. Psychometric and clinical tests of validity of the Japanese SF-36 Health Survey. *J Clin Epidemiol* 1998; 11: 1045-53.