Published in final edited form as:
Med Care. 2000 May ; 38(5): 517–527.

Evaluating the Equivalence of Health Care Ratings by Whites and Hispanics

Leo S. Morales, MD, MPH, **Steve P. Reise, PhD**, and **Ron D. Hays, PhD**
From the University of California at Los Angeles and RAND, Santa Monica, California.

Abstract

Purpose—This study was designed to assess the equivalence of a/health care ratings scale administered to non-Hispanic white and Hispanic survey respondents.

Methods—We sent 18,840 questionnaires to a random sample of patients receiving medical care from a physician group association concentrated in the western United States; 7,093 were returned (59% adjusted response rate). Approximately 90% of survey respondents self-identified as white/Caucasian (n = 5,508) or Hispanic/Latino (n = 713). Interpersonal and technical aspects of medical care were assessed with 9 items, all administered with a 7-point response format: the best, excellent, very good, good, fair, poor, and very poor, with a "not applicable" option. Item response theory procedures were used to test for differential item functioning between white and Hispanic respondents.

Results—Hispanics were found to be significantly more dissatisfied with care than whites (effect size=0.27; P <0.05). Of the 9 test items, 2 had statistically significant differential item functioning (P <0.05): reassurance and support offered by your doctors and staff and quality of examinations received. However, summative scale scores and test characteristic curves for whites and Hispanics were similar whether or not these items were included in the scale.

Conclusions—Despite some differences in item functioning, valid satisfaction-with-care comparisons between whites and Hispanics are possible. Thus, disparities in satisfaction ratings between whites and Hispanics should not be ascribed to measurement bias but should be viewed as arising from actual differences in experiences with care.

Keywords

Hispanics; satisfaction; psychometrics; quality of care

As the health care system continues to evolve, consumers have increasingly turned to cost and quality-of-care information to guide their health care choices. Demand for such information, in turn, has fueled the number of consumer surveys conducted each year. Although such surveys can provide important information about how well health plans and clinicians are meeting the needs of their various patient populations,[1,2] a number of researchers have raised methodological concerns about their use in culturally and linguistically diverse patient populations. In addition to concerns about providing adequate translations into multiple languages,[3] there is concern that cultural differences in the interpretation of questions [4–8] and in response styles [9] may limit direct comparisons between members of different racial/ethnic

Address correspondence to: Leo S. Morales, UCLA Medicine/GIM, 911 Broxton Plaza, Box 951736, Los Angeles, CA 90095-1736. E-mail: moralesr@rand.org.

Supported by the Agency for Health Care Policy Research to RAND (U18HS09204) (Dr Hays, principal investigator; Dr Morales, co-principal investigator) and an unrestricted research grant from The Medical Quality Commission to RAND (Dr Hays, principal investigator).

groups. As a result, the quality of care provided to members of vulnerable population groups may prove difficult to monitor, evaluate, and improve. Hence, there is a need to determine the equivalence of patient satisfaction measures in different cultural and linguistic groups.

This article addresses the comparability of ratings by Hispanic and white consumers. In a prior study, we reported greater dissatisfaction with provider communication among Hispanics than among whites and raised the concern that undetected measurement bias may have affected our results.[10] In this study, we assess the equivalence of satisfaction-with-care questions administered to white and Hispanic respondents in that study.[11] More specifically, we test for the measurement equivalence of a 9-item satisfaction with care scale using multigroup item response theory (IRT) procedures. Because no prior empirical work has addressed the comparability of patient satisfaction with care ratings for whites and Hispanics, we had no a priori hypotheses regarding particular items that might be expected to display bias. Thus, this research is exploratory in nature.

Methods

Setting

This study was based on survey data obtained from randomly selected patients receiving medical care from an association of 48 physician groups. The survey asked individuals about their satisfaction with care, health status, and use of health services during the past 12 months. Sixty-three physician group practices located primarily in the western United States participated in the study.

Patients >=18 years of age who made >=1 provider visit during the 365 days before the study were eligible for the study. The field period began in October 1994 and ended in June 1995. Each patient selected was mailed both Spanish and English versions of the survey, along with a $2 cash payment and a return envelope. Survey nonrespondents were followed up with reminder postcards and telephone calls. We mailed 18,840 surveys; 7,093 were returned, for an overall response rate of 59%, adjusted for undeliverable surveys, ineligible respondents, and deceased individuals. Response rates across medical groups ranged from 46% to 73% and were not significantly associated with ratings of health care.[11]

Survey Instrument

A detailed description of the survey, including a full description of its contents and psychometric properties, has been reported elsewhere.[11] Briefly, the survey included 153 items and took ~27 minutes to complete. The Spanish version of the survey was created through a process of independent forward (English to Spanish) and back (Spanish to English) translation followed by reconciliation.

This study evaluates the 9 survey items relevant to ratings of interpersonal and technical aspects of care. Five items (items 1 –5) asked about interpersonal care (medical staff listening, answers to your questions, explanations about prescribed medications, explanations about tests and medical procedures, and reassurance and support offered), and 4 items (items 6–9) asked about technical care (quality of examinations, quality of treatment, thoroughness and accuracy of diagnosis, and comprehensiveness of exams). All 9 survey items were asked with a 7-point response format (the best, excellent, very good, good, fair, poor, very poor), with a “not applicable” response option.

Seventy-nine percent of respondents were white/Caucasian (white) (n = 5,508), and 10% were Hispanic/Latino (Hispanic) (n = 713). The remaining 11% were Asian/Pacific Islander, African-American/black, Native American/American Indian, or other ethnic groups. Because precise item parameter estimation with IRT requires a large number of respondents across the

trait level continuum, [12] we retained only white and Hispanic respondents for this analysis. Although the white and Hispanic groups were similar about gender and health status, Hispanics were significantly younger (P <0.01), more likely to be married (P <0.01), and less likely to have graduated from high school (P <0.01) (Table 1).

Unidimensionality

Because the typical IRT model assumes sufficient unidimensionality,[13] we evaluated the dimensionality of our 9-item scale. First, we conducted principal components factor analysis for the white and Hispanic groups separately using the SAS FACTOR procedure [14] For both whites and Hispanics, we examined the magnitude of the eigenvalues, the ratio of the first and second eigenvalues, the component loadings, the Tucker and Lewis coefficient,[15] the average residual correlations (absolute values), and the SD of the residual correlations, In addition, we computed item-scale correlation coefficients and internal consistency reliability for the white and Hispanic groups.

Overview of IRT Models

IRT models posit a nonlinear monotonic function to account for the relationship between the examinee's position on a latent trait ([THETA]) and the probability of a particular set of item responses.[16] In this study, [THETA] refers to a respondent's level of satisfaction with care. The curves specified by this function are referred to as category response curves (CRCs). We used the generalized partial credit model as implemented in Parscale 3.5 [17] to estimate the relationship between [THETA] and the item response probabilities. This model was developed for scales composed of items with polytomous response formats and defines the CRCs for each item (i) and response category (k) as follows: EQUATION (1) where each item is represented by 3 parameters (a_i, $[lambda]_i$, $[tau]_k$) and the examinee trait level is represented by 1 parameter, [THETA]. For identification purposes, the latent trait scale is specified to have a mean of 0 and an SD of 1.0. The $[tau]_k$ parameters are called category intersection parameters; there are 6 such parameters for an item with 7 response options.

$$P_{ik} = \exp\left[\sum_{v=1}^{k} a_i(\theta - \lambda_i + \tau_k)\right] \Big/ \sum_{c=1}^{K} \exp\left[\sum_{v=1}^{c} a_i(\theta - \lambda_i + \tau_k)\right] \qquad \text{Equation 1}$$

The $[lambda]_i$ parameter is called an item location parameter. It indicates the difficulty of an item and can be thought of as shifting the intersection parameters up and down the latent trait scale. Large positive values of $[lambda]_i$ indicate a difficult item in which few examinees respond in the highest categories. Negative $[lambda]_i$ values indicate an easy item in which many examinees respond in the highest category. The slope parameter (a_i) indicates how fast the probability of responding in a higher category changes as a function of increases in the trait level. Items with large a_i are more discriminating than items with smaller slopes.

Assessing Goodness of Fit

There is no widely accepted goodness-of-fit statistic or index available for polytomous IRT models. To assess fit, we computed the difference between the observed and expected response frequencies by item and response category for whites and Hispanics. Parscale 3.5 does produce an item-fit $[chi]^2$ statistic based on these cell frequencies, but this test is too sensitive to sample size to produce a good gauge of model fit.[18,19]

Assessing Measurement Invariance With IRT

Measurement invariance (no bias) occurs when the CRCs for each item of a scale are identical for the groups of examinees in question (eg, whites and Hispanics).[20] Conversely, when particular item CRCs are not identical, measurement invariance is notobtained. The IRT literature uses the term differential item functioning (DIF) to describe items with nonidentical CRCs across groups.

In this study, DIF is determined by contrasting the item parameters, ie, a_i and [lambda] parameters, that determine the CRCs for whites and Hispanics.[21] Because the CRCs are completely determined by their corresponding item parameters, CRCs can be identical only if the item parameters that determine them are equal.

To guard against finding item DIF by chance alone, we conducted our analyses in a stepwise fashion. First, we contrasted a multigroup model in which the slope and location parameters were freely estimated between groups (unconstrained model) with a multigroup model in which the slope and location parameters were constrained to equality across groups (fully constrained model). A significant difference in the likelihood function value for the 2 models was interpreted as indicating the presence of DIP without identifying the particular items accounting for it.[22]

Subsequently, we fit 2 additional multigroup models to test individual items for DIF. In the first model, we freely estimated the slope parameters across ethnic groups while constraining the location parameters to equality. Then, we compared the slope parameters for each item using the following effect size statistic: EQUATION (2) where DIF = $\hat{a}_{i(white)} - \hat{a}_{i(Hispanic)}$. SDIF refers to standardized differential item functioning and is evaluated as $[chi]^2$ with 1 *df*.[21]

$$SDIF = DIF \Big/ \sqrt{Var\,\hat{a}_{i(white)} + Var\,\hat{a}_{i(Hispanic)}} \qquad \text{Equation 2}$$

In the second model, we freely estimated the location parameters across ethnic groups while constraining the slope parameters to equality. We computed a similar statistic that contrasted the location parameters for each item: EQUATION (3) where DIF = [Greek small letter lambda with circumflex accent]$_{i(white)}$−[Greek small letter lambda with circumflex accent]$_{i(Hispanic)}$. Note that in both models, the category intersection parameters ($[tau]_k$) are constrained to equality across ethnic groups. For this study, an item was considered to display DIF if its test-statistic $[chi]^2$ value was significant at the 0.05 level.

$$P_{ik} = \exp\left[\sum_{v=1}^{k} a_i(\theta - \lambda_i + \tau_k)\right] \Big/ \sum_{c=1}^{K} \exp\left[\sum_{v=1}^{c} a_i(\theta - \lambda_i + \tau_k)\right] \qquad \text{Equation 3}$$

Results

Descriptive Results and Unidimensionality of Scale

Table 2 shows the raw score descriptive statistics (ie, means and SD) and inter-item correlation coefficients for each ethnic group. Also shown is the ethnic group effect size (the group mean difference divided by the pooled SD) for each item and for the scale. A total scale score was computed by summing across the 9 items (possible 0 to 100 range). The total score was 67.86 (SD = 16.11) for whites (n = 5,508) and 63.54 (SD = 16.34) for Hispanics (n = 713). The difference between the mean scores was significant (t = 6.74, P <0.01) and resulted in an effect size of 0.27 (pooled SD = 16.14). Thus, with no item bias (measurement invariance) assumed,

Hispanics scored nearly one third of an SD lower than whites on this satisfaction-with-care scale.

The inter-item correlation coefficients ranged from 0.66 to 0.83 for whites and from 0.69 to 0.84 for Hispanics (Table 2). The results of the principal-components analysis of the 9 items indicated 1 dimension for whites and Hispanics. For both whites and Hispanics, only 1 eigenvalue was >1; it accounted for 78% of the total variance for whites and 77% of the total variance for Hispanics. The ratio of the first and second eigenvalues was 7.1/0.4 = 17.8 for whites and 6.9/0.5 = 17.3 for Hispanics. The mean residual correlation (absolute value) after extraction of 1 factor was 0.03 (SD = 0.03) for whites and 0.03 (SD = 0.03) for Hispanics. The Tucker and Lewis coefficients for a 1 -factor solution were 0.96 and 0.94 for whites and Hispanics, respectively. Principal-components loadings were >=0.83 for both whites and Hispanics, and item-scale correlation coefficients (corrected for overlap) ranged from 0.81 to 0.89 for whites and from 0.79 to 0.89 for Hispanics (Table 3). Alpha coefficients for both whites and Hispanics were 0.96. By any standard factor analytic/psychometric criterion, this 9-item scale is unidimensional.[13]

Goodness of Fit

Table 4 shows the difference between the observed and expected response frequencies by item and response category for whites and Hispanics as evidence of data-model fit. The mean discrepancy (absolute values) across all items and all response categories was 0.04 (SD = 0.03) for whites and 0.02 (SD = 0.02) for Hispanics. The item-fit [chi]^2 statistics generated by Parscale were significant (P <0.05) for both groups across all items.

IRT Results

The mean score difference between whites and Hispanics on the latent trait scale was 0.27 (SD = 0.99), which is consistent with the raw score effect size noted above. The difference in likelihood function value between the unconstrained model and the fully constrained model was statistically significant at the P <0.05 level, indicating the presence of item-level DIF.

Table 5 shows item slope parameter estimates and the slope parameter DIF statistics. It is worth noting that the mean item slopes were 2.86 for whites and 2.88 for Hispanics, indicating good model fit at the item level and suggesting that the items in the scale are highly discriminating. Slope parameter values >2.0 are generally regarded as high.[12] The DIF test results show that the slope parameter estimates for items 5 ([chi]2 = 4.11, P = 0.04) and 6 ([chi]2 = 11.94, P <0.01) were statistically different between the 2 groups. The slope parameter estimates for item 5 were 2.84 for whites and 2.53 for Hispanics. Similarly, the slope parameter estimates for item 6 were 3.09 for whites and 3.70 for Hispanics.

Table 6 shows item location parameter estimates and the location parameter DIF statistics. The DIF statistics indicate that no items demonstrated DIF about item location. Only item 6, for which the location parameter estimates were −0.83 for whites and −0.76 for Hispanics, had a nearly significant DIF statistic ([chi]2 = 3.65, P = 0.05).

Assessing the Impact of Items With DIF

To evaluate the impact of the item-level DIF on raw scale scores, we dropped the biased items from the scale and recomputed the effect size for whites' versus Hispanics' satisfaction ratings. The effect sizes were computed based on a summative scale (0 to 100 possible range). After dropping item 5, we obtained scale scores of 67.8 (SD = 16.0) for whites and 63.6 (SD = 16.3) for Hispanics and an effect size of 0.26 (pooled SD = 16.0). After dropping item 6, we obtained scale scores of 67.7 (SD = 16.5) for whites and 63.4 (SD = 16.5) for Hispanics and an effect size of 0.26 (pooled SD = 16.5). Finally, dropping items 5 and 6 from the scale simultaneously,

we obtained scale scores of 67.7 (SD = 16.4) for whites and 63.5 (SD = 16.6) for Hispanics and an effect size of 0.26 (pooled SD = 16.4). Recall that with all 9 items, the effect size was 0.27.

To further assess the effect of the detected item bias on our measure of satisfaction with care, we compared test response curves for whites and Hispanics using the following procedure. The test response curves show the relationship between the underlying level of satisfaction and the expected raw score on the 9-item scale. First, we estimated the IRT item parameters for the 9-item satisfaction scale independently for whites and Hispanics. This is equivalent to estimating a simultaneous multigroup model without between-group constraints on any of the parameters. However, because the 2 sets of item parameters may not be on the same scale, we resealed the item parameter estimates for Hispanics to those for whites by estimating linking constants and performing the appropriate transformations. Using the 2 sets of commonly scaled item parameters, we then computed the test response curves for whites and Hispanics.

Figure 1 shows the test response curves for whites and Hispanics. Deviations between the test response curves for whites and Hispanics show the degree of differential scale functioning due to items 5 and 6.

Figure 2 shows the results of subtracting the Hispanics' test response curve from the whites' test response curve. At low satisfaction levels, whites tend to score higher than Hispanics, whereas at middle levels of satisfaction, Hispanics tend to score higher than whites. However, the largest differential scale functioning (bias) is 1.5, which occurs at the −2.0 satisfaction level. A differential of 1.5 (on the 0 to 100 score range) represents < 1/10 of an SD difference between whites and Hispanics with the same latent trait level.

Discussion

This study examined a satisfaction with care scale for equivalence among 2 demographically important groups in the United States: whites and Hispanics. Our study found that valid comparisons between whites and Hispanics are possible, despite detection of statistically significant differences in the slope parameters for 2 of 9 scale items. More specifically, we found that items 5 (reassurance and support) and 6 (quality of examinations) showed statistically significant DIF ($P < 0.05$) but that the DIF did not have a meaningful impact on the expected scores of whites and Hispanics responding to these items. As a result, Hispanics' significantly lower rating of care in this study should be viewed as representing actual differences in experiences with care and should not be attributed to biased measurement.

Previous methodological studies of survey questions have found evidence that whites and Hispanics may not respond similarly. Johnson et al [4] found qualitative differences in whites' and Hispanics' interpretation of health status questions from widely used health surveys. Hayes and Baker [9] found that the reliability and validity of a Spanish version of a patient satisfaction with communication scale differed significantly from that of the English version. Aday and colleagues [23] noted that Hispanics were more likely to respond "yes" to patient satisfaction questions than non-Hispanics, regardless of whether the question indicated greater satisfaction or dissatisfaction, providing support for the contention that Hispanics are prone to more acquiescent responses than non-Hispanics or are biased toward more favorable responses.[9]

Unlike many prior studies, we conducted analyses to assess the effect of differences in scale functioning among whites and Hispanics on comparisons between the groups. Specifically, we examined the effect of the 2 biased items on the group mean scale scores and computed the effect size with and without including the items showing DIF. When all 9 items were included in the scale, the effect size was 0.27, with whites rating care significantly more positively than

Hispanics (P <0.05). When the biased items-items 5 and 6-were dropped from the scale, the effect size changed to 0.26 and the mean scale scores remained significantly different (P<0.05).

Furthermore, we examined the test response curves for whites and Hispanics. These curves plot the expected raw scale scores of each group over the underlying satisfaction continuum. At worst, our 9-item scale resulted in a 1.5 raw score differential (bias) between whites and Hispanics. Together, these results show that at all levels of satisfaction, whites and Hispanics have nearly identical expected raw scale scores despite 2 items with statistically significant DIF.

Our study uses a relatively new procedure for detecting DIF that is based on polytomous IRT model procedures. Prior studies have relied primarily on classic psychometric methods (eg, reliability, validity, and item-scale correlations), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) for the identification of item and survey bias in multiethnic settings. Although these methods can yield useful information about item and scale bias, IRT models are theoretically more appropriate for survey scales that use categorical response formats. Although EFA and CFA models typically assume continuous indicators that have linear relationships with the latent variable(s), IRT models do not make these assumptions. Furthermore, IRT models do not assume multivariate normality, which is an assumption made by most CFA estimation routines. IRT models also offer practical approaches to quantifying the effect size of statistically significant DIF. As other studies have illustrated [24] and as we have demonstrated in this study, statistically significant DIF does not necessarily invalidate comparisons between groups of interest. EFA and CFA models do not offer a similarly practical approach to assessing the impact of DIF when it is detected. For more detailed discussions of IRT and factor analytic approaches to item and scale bias detection, the reader is referred to McDonald [25] and Reise et al, [26] and Widaman and Reise.[27]

Explaining why the items asking about quality of examinations and reassurance and support demonstrated DIF is beyond the scope of this study and thus remains speculative. Item bias occurs when an instrument measures one thing for one group and something else for the other group. Items 5 and 6 may have been interpreted differently by white and Hispanic respondents because of between-group differences in age, gender, income, education, or cultural background. Although we found significant differences in the sociodemographic characteristics of the whites and Hispanic respondents in our study, our purpose was not to identify factors that explain the DIF we detected. On the basis of the results of this study, we cannot attribute the DIF in these items to ethnicity per se or to any other particular background or health status variable. Future studies may be needed to explain the influence of background characteristics on differences in item functioning.

The moderate response rate (59%) in this study may pose some risk of nonresponse bias. To threaten the validity of this study, however, respondents and nonrespondents would have to differ about their interpretations of the meanings of the survey questions. This might occur, for example, if Hispanic respondents were more acculturated than Hispanic nonrespondents. Acculturation refers to the processes of acquisition the host culture by an ethnic minority.[28] In this scenario, the Hispanic respondents in our study would be culturally more similar to the white respondents than a truly representative sample of the Hispanic patients would be; therefore, our study would be less likely to find measurement bias than a study with a more representative Hispanic sample. Unfortunately, our data sources do not allow us to compare respondents and nonrespondents along such dimensions as acculturation. On the basis of the available data, the differences between the sampling frame and those responding to the survey were minimal. Specifically, those returning the questionnaire had a mean age of 51 years (median = 49 years), whereas the mean age of the sampling frame was 46 years (median = 43 years). Sixty-five percent of the responders were women; 58% in the sampling frame were

women. The last medical visit for the study participants was, on average, 119 days (median = 88 days) before the beginning of the study. For those in the sampling frame, the average was 130 days (median = 112 days).[11] Unfortunately, our data sources prevented us from computing ethnic group-specific response rates.

In sum, this study addressed the validity of comparisons of satisfaction with care across ethnic groups. We found that lower ratings of care among Hispanics relative to whites were not attributable to item or scale bias and therefore reflect actual differences in experiences with care between the 2 groups. These results support the findings of other researchers that Hispanics are not as well served by the current health care system as whites.[10, 29–34] More generally, our findings suggest that when disparities in patient ratings of care are detected across ethnic groups, they should not be attributed to biased measurement unless significant DIF (in the statistical and practical sense) can be demonstrated.

Acknowledgements

We express appreciation to Tamara Breuder for her assistance in preparing the manuscript and Gail Della Vedova and the staff members of The Medical Quality Commission for their cumulative input. The views expressed herein are those of the authors and do not necessarily reflect the views of The Medical Quality Commission, RAND, or UCLA.

References

1. Edgman-Levitan S, Cleary PD. What information do consumers want and need? Health Aff 1996 winter;:42.
2. Crofton C, Luliban JS, Darby C. Foreword. Med Care 1999;37:MS1–MS9. [PubMed: 10098554]
3. Weidmer B, Brown J, Garcia L. Translating the CAHPS 1.0 survey instrument into Spanish. Med Care 1999;37:MS89–MS97. [PubMed: 10098563]
4. Johnson TP, O'Rourke D, Chavez N, Sudman S, Warnecke RB, Lacey L, et al. Cultural variations in the interpretation of health questions. Health Survey Res Methods: Conference PrOC 1996:57–62.
5. Angel R, Thoits P. The impact of culture on the cognitive structure of illness. Med Psychiatry 1987;11:465–494.
6. Liang J, Van Tran T, Krause N, Markides KS. Generational differences in the structure of the CES-D scale in Mexican Americans. J Gerontol 1989;44:S110–S120. [PubMed: 2715592]
7. Dick RW, Beals J, Keane EM, Manson SM. Factorial structure of the CE S-D among American Indian adolescents. J Adolesc Health 1994;17:73–79.
8. Weissman MM, Sholomskas D, Pottenger M, Prusoff BA, Locke BZ. Assessing depressive symptoms in five psychiatric populations: A validation study. Am J Epidemiol 1977;106:203–214. [PubMed: 900119]
9. Hayes RP, Baker DW. Methodological problems in comparing English-speaking and Spanish-speaking patients' satisfaction with interpersonal aspects of care. Med Care 1998;36:230–236. [PubMed: 9475476]
10. Morales LS, Cunningham WE, Brown JA, Lui H, Hays RD. Are Latinos less satisfied with communication by health care providers? A study of 48 medical groups. J Gen Intern Med 1999;14:409–417. [PubMed: 10417598]
11. Hays RD, Brown JA, Spritzer KL, Dixon WJ, Brook RH. Member ratings of health care provided by 48 physician groups. Arch Intern Med 1998;158:785–790. [PubMed: 9554685]
12. Hambleton, RK.; Swaiminathan, H.; Rogers, HJ. Fundamentals of item response theory. Thousand Oaks, Calif: Sage; 1991.
13. McDonald R. The dimensionality of tests and items. Br J Math Stat Psychol 1967;34:100–117.
14. SAS/STAT user's guide, version 6. 4. Cary, NC: SAS Institute Inc; 1989. p. 773-823.
15. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika 1973;38:1–10.
16. Lord, FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum; 1980.

17. Muraki, E.; Block, RD. Parscale IRT item analysis and test scoring for rating scale data. Chicago, III: Scientific Software International, Inc; 1997.

18. Orlando M, Thissen D. Likelihood based item-fit indices for dichotomous item response theory models. Appl Psychological Measurement 2000;24:50–64.

19. Reise SP. A comparison of item- and person-fit methods of assessing model-data fit in IRT. Appl Psychological Measurement 1990;14:127–137.

20. Kok, F. Item bias and test multidimensionality. In: Langeheine, R.; Rost, J., editors. Latent trait and latent class models. New York, NY: Plenum Press; 1988. p. 263-275.

21. Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holand, PW.; Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Erlbaum; 1993.

22. Thissen, D. MULTILOG: Multiple categorical item analysis and test scoring using item response theory (version 6). Chicago, III: Scientific Software, Inc; 1991.

23. Aday LA, Chiu GY, Andersen R. Methodological issues in health care surveys of the Spanish heritage population. Am J Public Health 1980;70:367. [PubMed: 7361954]

24. Smith LL, Reise SP. Gender differences on negative affectivity: An IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. J PerS SOC Psychol 1998;75:1350–1362. [PubMed: 9866192]

25. McDonald, RP. Test theory: A unified treatment. Mawah, NJ: Lawrence Erlbaum Associates, Inc; 1999.

26. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychol Bull 1993;114:552–566. [PubMed: 8272470]

27. Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: applications in the substance abuse domain. In: Bryant, KJ.; Windle, M.; West, SG., editors. The science of prevention: Methodological advances from the alcohol and substance abuse research. Washington, DC: American Psychological Association; 1997.

28. Berry, JW. Acculturative stress. In: Organista, PB.; Chun, KM.; Marin, G., editors. Readings in ethnic psychology. London, England: Routledge; 1998.

29. Andersen R, Lewis SZ, Giachello AL, Aday LA, Chiu G. Access to medical care among the Hispanic population of the southwestern United States. J Health Soc Behav 1981;22:78–89. [PubMed: 7240708]

30. Baker DW, Parker RM, Williams MV, Coates WC, Pitkin K. Use and effectiveness of interpreters in an emergency department. JAMA 1996;275:783–788. [PubMed: 8598595]

31. Hu DJ, Covell RM. Health care usage by Hispanic outpatients as a function of primary language. West J Med 1986;144:490–493. [PubMed: 3716414]

32. Harpole LH, Orav EJ, Hickey M, Posther KE, Brennan TA. Patient satisfaction in the ambulatory setting: Influence of data collection methods and sociodemographic factors. J Gen Intern Med 1996;11:431–434. [PubMed: 8842937]

33. Molina, CW.; Zambrana, RE.; Aguirre-Molina, M. The influence of culture, class, and environment on health care. In: Molina, CW.; Aguirre-Molina, M., editors. Latino health in the US: A growing challenge. Washington, DC: American Public Health Association; 1997. p. 23-43.

34. Villa, ML.; Cuellar, J.; Gamel, N.; Yeo, G. Aging and health: Hispanic American elders. Stanford, Calif: Stanford School of Medicine, Stanford Geriatric Education Center; 1993.

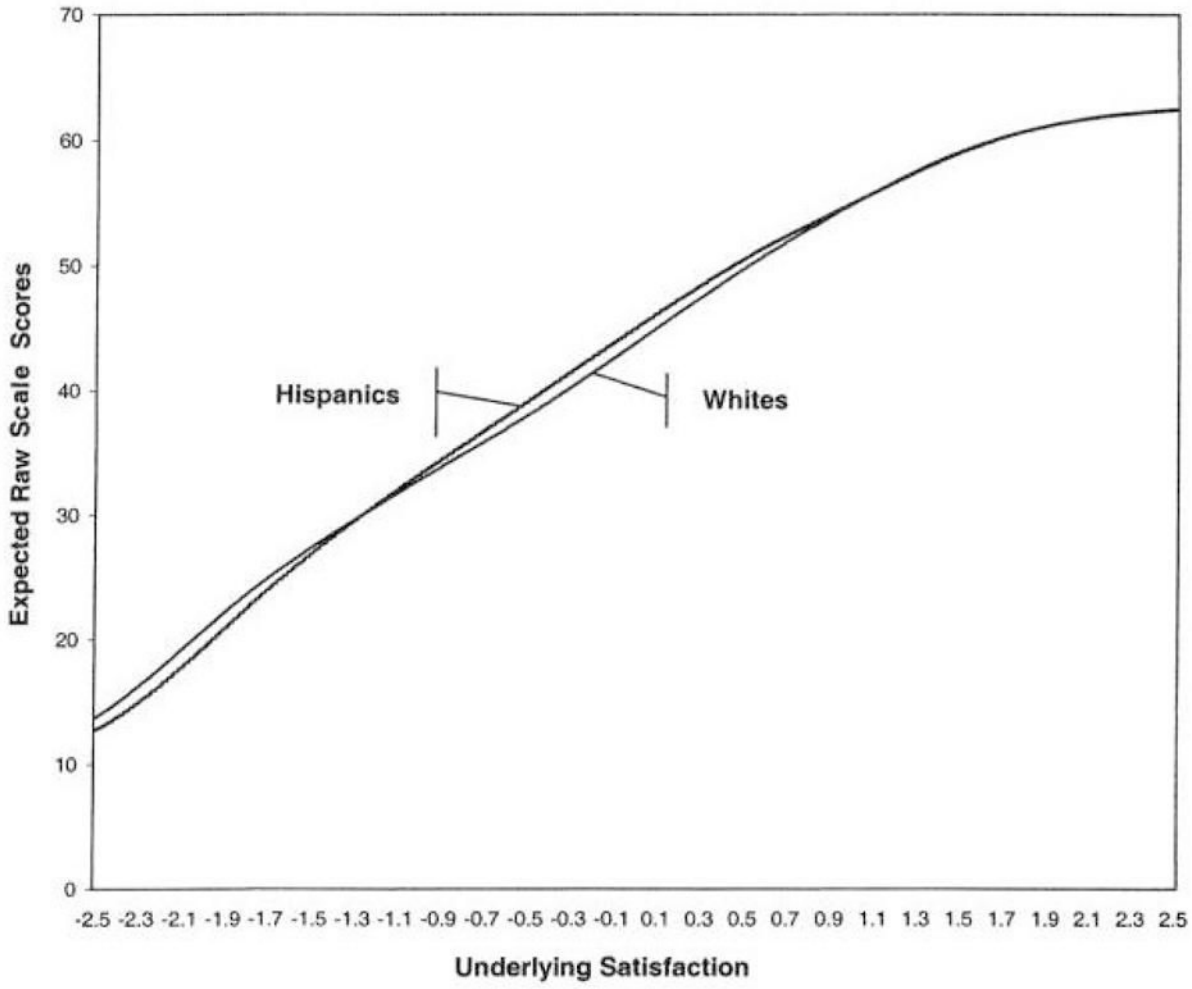

Fig. 1.

Expected raw scores for whites and Hispanics on 9-item satisfaction with care scale. Values >0 indicate Hispanic scores exceed white scores; values <0 indicate the converse. Each item is scored from 1 to 7, resulting in a 9–63 scale range.

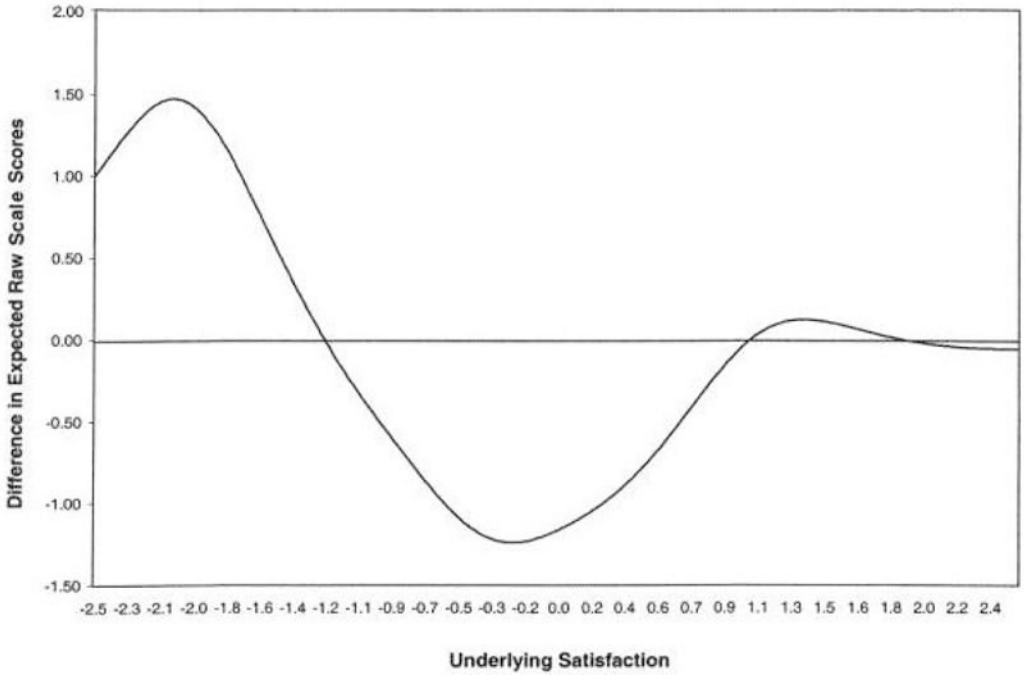

Fig. 2.

Difference between white and Hispanic expected raw scores on 9-item satisfaction with care scale. Values >0 indicate Hispanic scores exceed white scores; values <0 indicate the converse.

Table 1

Sample Description

	Whites (n = 5,508)	Hispanics (n = 713)	*P* for Difference
Age (mean ± SD), y	51.9 ± 17.5	41.7 ± 15.2	<0.01
Male, %	34.9	37.5	0.12
Married, %	73.7	78.1	<0.01
Graduated high school, %	69.3	46.4	<0.01
General health status (mean ± SD score on 0–10 scale; 10 = best)	7.3 ± 1.73	7.2 ± 1.79	0.24

Two-sided *t* tests were applied to continuous variables (age, health status) and χ^2 tests to proportions (% male, % married, and % graduated high school).

Table 2

Raw Score Descriptive Statistics and Inter-Item Correlations

	Whites (n = 5,508)		Hispanics (n = 713)			Inter-Item Correlations								
Item	**Mean**	**SD**	**Mean**	**SD**	**Effect Size**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	5.01	1.34	4.72	1.35	0.21		0.83	0.70	0.72	0.77	0.83	0.79	0.77	0.69
2	5.00	1.29	4.70	1.30	0.23	0.83		0.75	0.76	0.77	0.80	0.80	0.78	0.69
3	4.97	1.35	4.65	1.37	0.23	0.68	0.72		0.76	0.73	0.71	0.74	0.73	0.69
4	4.76	1.36	4.47	1.38	0.21	0.70	0.74	0.77		0.77	0.73	0.76	0.76	0.72
5	4.76	1.38	4.43	1.35	0.24	0.74	0.74	0.74	0.76		0.76	0.76	0.76	0.76
6	5.09	1.23	4.70	1.34	0.31	0.83	0.79	0.67	0.73	0.75		0.83	0.81	0.72
7	5.05	1.28	4.70	1.35	0.27	0.77	0.81	0.73	0.76	0.77	0.82		0.84	0.76
8	4.85	1.35	4.47	1.38	0.28	0.74	0.77	0.73	0.74	0.74	0.79	0.82		0.74
9	4.59	1.45	4.26	1.44	0.23	0.66	0.66	0.67	0.68	0.77	0.68	0.74	0.71	
Total	67.86	16.11	63.54	16.34	0.27									

Individual item scores range from 1 to 7 (7= the best). Total score was computed by summing, across the 9 item scores and then transforming to a 0 to 100 scale, where 100 is the highest possible rating. (All mean differences between whites and Hispanics were statistically significantly [$P < 0.05$].) Effect size was computed as the difference in means divided by the pooled SD. Inter-item correlation coefficients for Hispanics are shown above the diagonal; those for whites, below the diagonal.

Table 3

Principal Component Loadings and Item-Scale Correlations for Whites and Hispanics

Item	Principal Component Loadings		Item-Scale Correlations	
	White	Hispanic	White	Hispanic
1	0.89	0.88	0.86	0.84
2	0.90	0.89	0.87	0.86
3	0.85	0.85	0.82	0.81
4	0.88	0.87	0.84	0.83
5	0.89	0.89	0.86	0.86
6	0.90	0.89	0.87	0.86
7	0.92	0.91	0.89	0.89
8	0.90	0.89	0.87	0.86
9	0.85	0.83	0.81	0.79

Loadings were derived from a single-factor principal components model. The ratio of the first and second eigenvalues was 7.1/0.4 = 17.8 for whites and 6.9/0.5 = 17.3 for Hispanics. Item-scale correlations were corrected for overlap. Cronbach's α is 0.96 for both whites and Hispanics.

Table 4

Difference Between Observed and Expected Response Frequencies (Absolute Values) by Item and Response Category for Whites and Hispanics

	Principal Component Loadings		Item-Scale Correlations	
Item	**White**	**Hispanic**	**White**	**Hispanic**
1	0.89	0.88	0.86	0.84
2	0.90	0.89	0.87	0.86
3	0.85	0.85	0.82	0.81
4	0.88	0.87	0.84	0.83
5	0.89	0.89	0.86	0.86
6	0.90	0.89	0.87	0.86
7	0.92	0.91	0.89	0.89
8	0.90	0.89	0.87	0.86
9	0.85	0.83	0.81	0.79

Loadings were derived from a single-factor principal components model. The ratio of the first and second eigenvalues was 7.1/0.4 = 17.8 for whites and 6.9/0.5 = 17.3 for Hispanics. Item-scale correlations were corrected for overlap. Cronbach's α is 0.96 for both whites and Hispanics.

Table 5

Slope Parameters and DIF Tests for Whites and Hispanics

Item	White		Hispanic		SDIF	χ^2 (*df* = 1)	*P*
	Slope	SE	Slope	SE			
1	2.99	0.15	2.90	0.05	0.55	0.30	0.59
2	3.52	0.15	3.32	0.06	1.28	1.64	0.20
3	2.09	0.10	2.00	0.03	0.81	0.65	0.43
4	2.39	0.09	2.34	0.04	0.46	0.21	0.65
5	2.84	0.14	2.53	0.04	2.03	4.11	0.04*
6	3.09	0.16	3.70	0.07	−3.46	11.94	<0.01*
7	3.97	0.23	4.08	0.08	−0.47	0.23	0.64
8	3.11	0.15	3.28	0.05	−1.07	1.13	0.29
9	1.77	0.08	1.78	0.03	−0.10	0.01	0.88

SDIF indicates standardized DIF.

* *P* <0.05.

Table 6

Location Parameters and DIF Tests for Whites and Hispanics

Item	White		Hispanic		SDIF	χ^2 (*df* = 1)	*p*
	Slope	SE	Slope	SE			
1	2.99	0.15	2.90	0.05	0.55	0.30	0.59
2	3.52	0.15	3.32	0.06	1.28	1.64	0.20
3	2.09	0.10	2.00	0.03	0.81	0.65	0.43
4	2.39	0.09	2.34	0.04	0.46	0.21	0.65
5	2.84	0.14	2.53	0.04	2.03	4.11	0.04*
6	3.09	0.16	3.70	0.07	−3.46	11.94	<0.01*
7	3.97	0.23	4.08	0.08	−0.47	0.23	0.64
8	3.11	0.15	3.28	0.05	−1.07	1.13	0.29
9	1.77	0.08	1.78	0.03	−0.10	0.01	0.88

SDIF indicates standardized DIF.

* $p < 0.05$.