

Original Article

Cite this article: Teresi JA, Ocepek-Welikson K, Ramirez M, Kleinman M, Ornstein K, Siu A, Luchsinger J (2020). Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care (FAMCARE): Tests of differential item functioning between Hispanic and non-Hispanic White caregivers. *Palliative and Supportive Care* **18**, 544–556. <https://doi.org/10.1017/S1478951520000152>

Received: 10 April 2019
Revised: 7 February 2020
Accepted: 8 February 2020


Key words:

Differential item functioning; Ethnic diversity; Family satisfaction with end-of-life care; Item response theory; Palliative care

Author for correspondence:

Mildred Ramirez, Research Division, Hebrew Home at Riverdale in RiverSpring Health, 5901 Palisade Avenue, Riverdale, New York, NY 10471, USA.
E-mail: milramirez@aol.com

Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care (FAMCARE): Tests of differential item functioning between Hispanic and non-Hispanic White caregivers

Jeanne A. Teresi, ED.D., PH.D.^{1,2,3,4}, Katja Ocepek-Welikson, M.PHIL.¹, Mildred Ramirez, PH.D.^{1,2,4} , Marjorie Kleinman, M.S.³, Katherine Ornstein, PH.D., M.P.H.⁵, Albert Siu, M.D.⁶ and Jose Luchsinger, M.D.⁷

¹Research Division, Hebrew Home at Riverdale, Riverdale, New York, NY; ²Measurement and Data Management Core, Mount Sinai Pepper Older Americans Independence Center, Mount Sinai Medical Center, and Analytic Core, Columbia University Alzheimer's Disease Resource Center for Minority Aging Research, New York, NY; ³Columbia University Stroud Center, New York State Psychiatric Institute, New York, NY; ⁴Division of Geriatrics and Palliative Medicine, Weill Cornell Medical Center, New York, NY; ⁵Department of Geriatrics and Palliative Medicine, Institute for Translational Epidemiology Mount Sinai School of Medicine, New York, NY; ⁶Department of Geriatrics and Palliative Medicine, General Internal Medicine, Health Evidence and Policy, Mount Sinai Medical Center, New York, NY and ⁷Department of Medicine, Columbia University Medical Center, PH9 Center, New York, NY 10032

Abstract

Objective. Although the psychometric properties of the Family Satisfaction with End-of-Life Care measure have been examined in diverse settings internationally; little evidence exists regarding measurement equivalence in Hispanic caregivers. The aim was to examine the psychometric properties of a short-form of the FAMCARE in Hispanics using latent variable models and place information on differential item functioning (DIF) in an existing family satisfaction item bank.

Method. The graded form of the item response theory model was used for the analyses of DIF; sensitivity analyses were performed using a latent variable logistic regression approach. Exploratory and confirmatory factor analyses to examine dimensionality were performed within each subgroup studied. The sample included 1,834 respondents: 317 Hispanic and 1,517 non-Hispanic White caregivers of patients with Alzheimer's disease and cancer, respectively.

Results. There was strong support for essential unidimensionality for both Hispanic and non-Hispanic White subgroups. Modest DIF of low magnitude and impact was observed; flagged items related to information sharing. Only 1 item was flagged with significant DIF by both a primary and sensitivity method after correction for multiple comparisons: "The way the family is included in treatment and care decisions." This item was more discriminating for the non-Hispanic, White responders than for the Hispanic subsample, and was also a more severe indicator at some levels of the trait; the Hispanic respondents located at higher satisfaction levels were more likely than White non-Hispanic respondents to report satisfaction.

Significance of results. The magnitude of DIF was below the salience threshold for all items. Evidence supported the measurement equivalence and use for cross-cultural comparisons of the short-form FAMCARE among Hispanic caregivers, including those interviewed in Spanish.

Introduction

The Family Satisfaction with End-of-Life Care (FAMCARE) scale (Kristjanson, 1986, 1989), although used most widely with cancer patients in palliative care, has also been applied to a range of serious illness (Hwang et al., 2003), including caregivers to patients with Alzheimer's disease (Teresi et al., 2019) and residents in long-term care (Rodriguez et al., 2010). The FAMCARE is used widely internationally as a quality measure of end-of-life care in clinical and research settings, and translations are available in many languages, including Italian (D'Angelo et al., 2017), Spanish (Teresi et al., 2019), and Swedish (Ljungberg et al., 2015). Although the psychometric properties of the scale have been examined in cancer patients in diverse settings internationally, little evidence exists regarding measurement equivalence across ethnically diverse groups. There is also little experience with the scale among individuals with different diseases such as Alzheimer's disease and related disorders (ADRD) or among ethnic subgroups, including Spanish speakers and caregivers. While several studies have examined the relationship of demographic characteristics to satisfaction with

end-of-life care (Kristjanson, 1993; Lo et al., 2009; Aoun et al., 2010), no studies have examined these characteristics in terms of measurement equivalence in Hispanic samples.

A study of measurement equivalence comparing Black with White non-Hispanic caregivers of patients with cancer found that 13 items evidenced differential item functioning (DIF), a type of item bias; however, none of high magnitude (Teresi et al., 2015). Moreover, the scale-level impact was negligible. One item related to pain relief evidenced DIF for race and education and was also hypothesized to show DIF. To our knowledge, no other studies have examined the FAMCARE for equivalence of item endorsement across different socio-demographic groups using item response theory (IRT) methods to detect DIF. Thus, the aim of this set of analyses was to examine the psychometric properties of the scale in a sample of Hispanics using latent variable models and to obtain information on DIF to place in an existing item bank on family satisfaction and care transitions.

Methods

Qualitative

Qualitative methods, including content analyses and cognitive interviews, were used to develop Spanish translations for use among Spanish speakers (Teresi et al., 2019). The first step in the evaluation of DIF is the generation of a priori hypotheses regarding potential group differences in item responses, conditional on the trait. Hypotheses regarding potential racial/ethnic group differences in item response were established qualitatively by a panel of content experts. The following instructions related to hypotheses generation were given.

Differential item functioning means that individuals in groups with the same underlying trait (state) level will have different probabilities of endorsing an item. Put another way, item endorsement should depend only on the level of the trait (state), e.g., satisfaction, and not on membership in a group, e.g., race/ethnicity. Very specifically, randomly selected persons from each of two groups (e.g., minority and non-minority) who are at the same (e.g., mild) level of satisfaction should have the same likelihood of reporting being very satisfied with the aspects of care provided. If it is hypothesized that this is not the case, it would be hypothesized that the item has DIF with respect to race/ethnicity.

The rationale for DIF hypotheses is that items may be posited to have a different meaning for some individuals and may measure a trait that is not expected. Thus, the item could perform differently for some groups, conditional on the trait.

Quantitative analyses and tests of DIF hypotheses

The graded (Samejima, 1969) form of the IRT model (Lord and Novick, 1968; Lord, 1980; Hambleton et al., 1991) was used for the analyses of DIF. An item shows DIF if people from different subgroups but at the same level of satisfaction have unequal probabilities of endorsement. The item characteristic curve (ICC) that relates the probability of an item response to the underlying state, e.g., satisfaction, measured by the item set can be characterized by two parameters: location (denoted b and also called threshold, difficulty, or severity) and a discrimination parameter (denoted a) that is proportional to the slope of the curve. DIF analyses approaches to assessment of patient and caregiver-reported outcomes using IRT are described in Orlando-Edelen et al. (2006). The Wald test was used for examination of group differences in IRT item parameters

(Lord, 1980; Teresi et al., 2000; Cai et al., 2011) accompanied by magnitude measures (Thissen et al., 1993; Raju et al., 1995; Kleinman and Teresi, 2016).

Uniform DIF is detected when the b parameters differ because the direction of the DIF (more or less severe) for one group as contrasted with a comparison group is the same across the latent continuum. If the a parameters differ, this result is called non-uniform DIF because the ICC curves cross and the direction of DIF can differ across the latent continuum. Non-uniform DIF occurs when the probability of response is in a different direction for the reference and focal groups, at different levels of the latent trait (θ). For example, Hispanic persons may have a lower probability than White, non-Hispanic persons of endorsing a satisfaction item at low levels of the satisfaction trait and higher probabilities of an endorsement than White, non-Hispanic persons at higher levels. If non-uniform DIF is detected in the context of the IRT method, this finding assumes primacy over findings of uniform DIF because tests for group differences in the a parameters are followed by conditional tests of the b parameters (tests of b parameters are performed, constraining the a parameters to be equal).

An iterative process was used in the selection of the anchor items for theta estimation. There are several methods for selecting anchor items, assumed to be DIF-free (Orlando-Edelen et al., 2006; Woods, 2009; Wang et al., 2012). The approach that was used in these analyses was a modified “all-other” method in which initial DIF estimates were obtained by treating each item as a “studied” item while using the remainder as “anchor” items. The purification process was also iterative, and items identified as DIF-free were those included in the final anchor set. IRTPRO, version 3.1 option 3, which permits the all-other approach for the multiple group case was used. This (Wald-type) procedure is more robust than just relying on the all-other anchor procedure and may take several iterations.

The final P values testing for DIF were adjusted using the Bonferroni method (Bonferroni, 1936). Other methods such as Benjamini-Hochberg (B-H; Benjamini and Hochberg, 1995; Thissen et al., 2002) have been used in sensitivity analyses for many of our studies. Generally, the results are almost identical. Thus, the Bonferroni method was selected as the primary approach for adjustment for multiple comparisons.

Model assumptions and fit: Exploratory and confirmatory factor analyses (Asparouhov and Muthén, 2009) to examine dimensionality were performed within each subgroup studied, and fit indices (Bentler, 1990) examined. Additionally, the explained common variance (ECV) was used as an indicator of unidimensionality. The ECV (Sijtsma, 2009), estimated as the percent of observed variance explained (Reise, 2012), can be calculated as the ratio of the first eigenvalue to the sum of all eigenvalues extracted (see Reise et al., 2010).

Local independence requires that all pairs of item responses be independent, conditional on the latent trait. Local dependency (LD) was examined using the methods of Chen and Thissen (1997). A suggested cutoff indicative of potential LD is 10 (Chen and Thissen, 1997; Cai et al., 2011). This approach is based on a comparison of observed and expected frequencies derived from item-by-item two-way cross-tabulations; the likelihood ratio statistic resulting from this comparison is chi-square distributed. LD statistics are affected by sample size and increase in value with the increased sample size. Thus, to ensure comparability in sample sizes between the Hispanic and non-Hispanic White sample, a random sample of the White non-Hispanic group comparable in size

Table 1. Demographic characteristics of the caregivers and care recipients for the White and Hispanic samples

	Caregivers			Care recipients		
	Hispanic	Non-Hispanic White	Total	Hispanic	Non-Hispanic White	Total
Gender						
Female	262 (83%)	832 (55%)	1,094 (60%)	195 (62%)		195 (62%)
Male	55 (17%)	684 (45%)	739 (40%)	118 (38%)		118 (38%)
Missing	0	1	1	4		1,521
Age						
Age <65	234 (74%)	931 (62%)	1,165 (64%)	17 (6%)	931 (62%)	948 (52%)
Age 65 and over	83 (26%)	577 (38%)	660 (36%)	292 (95%)	577 (38%)	869 (47%)
Mean (SD)	57.9 (11.2)	60.9 (11.6)	60.2 (11.6)	79.9 (8.9)	60.7 (11.6)	63.9 (13.3)
Range	19–85	21–100	19–100	53–101	21–100	21–101
Missing	0	9	9	8	9	17
Education						
Less than high school (0–11)	73 (24%)	169 (11%)	242 (13%)		169 (11%)	169 (11%)
High school	95 (31%)	501 (33%)	596 (33%)		501 (33%)	501 (33%)
Some college and above (13+)	141 (45%)	842 (56%)	983 (54%)		842 (56%)	842 (56%)
Mean (SD)	12.7 (3.6)	14.1 (3.1)	13.8 (3.2)		14.1 (3.1)	14.1 (3.1)
Range	2–20	5–22	2–22		5–22	5–22
Missing	8	5	13		5	322
Caregiver and care recipient relationship						
Family member living with patient	243 (77%)	739 (54%)	982 (58%)	Not applicable		
Family member not living with patient	74 (23%)	504 (37%)	578 (34%)			
Friend	0	114 (8%)	114 (7%)			
Other	0	20 (1%)	20 (1%)			
Missing	0	140	140			

Sample size: Hispanic responders ($n = 317$); non-Hispanic White responders ($n = 1,517$); total ($n = 1,834$). Data were not available for care recipient education for the Hispanic sample.

to that of the Hispanic sample was selected. The root mean square error of approximation (RMSEA) was examined for both confirmatory factor analyses and IRT model fit.

The best methods and criteria for cutoff values for goodness of fit statistics have been debated (e.g., Cook et al., 2009), with recommendations to not be overly reliant on specific values, given the many factors that may affect these statistics. The following model fit statistics and criteria for goodness of fit (Bentler, 1990) provided general guidelines, and included the comparative fit index (CFI; Bentler, 1990; $CFI > 0.95$), Tucker–Lewis index (TLI; Tucker and Lewis, 1973; $TLI > 0.95$), and the root mean square error of approximation ($RMSEA < 0.06$).

Evaluation of DIF magnitude and impact: Expected item scores were measures of magnitude. A method for quantification of the difference in the average expected item scores is the non-compensatory DIF (NCDIF) index used by Raju et al. (1995). NCDIF is expressed as the average squared difference in expected scores for individuals as members of the focal group and as members of the reference group. The cutoff recommended as indicative of high DIF magnitude is 0.024 for polytomous items with three response options. An additional effect size measure (T1) proposed by Wainer (1993) and extended for polytomous data by Kim et al.

(2007) was also examined; however, primary reliance was on the NCDIF magnitude measure because little research has been conducted on the performance of T1. The use of these statistics is explicated in Kleinman and Teresi (2016) and Teresi et al. (2007).

Expected scale scores that provide information about the effect of DIF on the total score were calculated by summing the expected item scores. Group differences in these scale response functions provide overall aggregated measures of impact.

DIF sensitivity analyses: Sensitivity analyses using a different method was conducted using an ordinal logistic regression approach with a latent conditioning variable; lordif, version 0.3-3 (Choi et al., 2011) was used. This method was used to flag consistent DIF identified by both methods that might be salient based on magnitude and impact measures.

Additionally, sensitivity analyses were conducted comparing only Spanish speakers to White, non-Hispanic English speakers.

Reliability and information: Reliability was evaluated with McDonald's omega total (ω ; McDonald, 1999); this estimate is based on the proportion of total common variance explained. Reliability estimates were also calculated for various points along the latent continuum of family satisfaction using IRT. IRT also provides estimates of the information provided by

items and scales. This item information can be used to select items for short-form measures. Additionally, information function parameters stored in item banks are used to generate computerized adaptive tests that tailor item selection to target the respondent's level of the trait based on responses to a starting item and to other items administered.

MPlus, version 6.11 (Muthén and Muthén, 2011) was used for factor analyses and IRTPRO, version 3.12 (Cai et al., 2011) for IRT item parameter estimation and DIF analyses. Item level magnitude using NCDIF (Fleer, 1993; Raju et al., 1995; Flowers et al., 1999; Morales et al., 2006) was estimated using MAGNITS (Kleinman and Teresi, 2016). Scale-level impact was evaluated using lordif, version 0.3-3 (Choi et al., 2011) in the psych package in R. Reliability estimated with McDonald's omega was also calculated with R version 3.4.4 (R core team, 2018).

Measure

The short-form FAMCARE used in these analyses was based on earlier work (Teresi et al., 2014) with advanced psychometric methods. This work showed that lower categories were overlapping such that the probability of response was similar for the three categories: "very dissatisfied," "dissatisfied," and "undecided," indicating little if any unique information provided by these categories. Thus, items were coded as ordinal and collapsed as follows: "very satisfied" responses were coded as 2, "satisfied" as 1 and "not satisfied" (indecision or "dissatisfaction") as 0, with a resulting sum score from 0 to 20. The item analyses were thus performed with three ordinal response categories.

Sample

There were 1,834 respondents, 317 Hispanics, and 1,517 non-Hispanic Whites; among the Hispanic sample, 209 were interviewed in Spanish. For these analyses, the Hispanic Spanish and English speakers were combined because not enough respondents were interviewed in English to perform a separate DIF analysis by the language of administration. The Hispanic sample was comprised of caregivers to patients with Alzheimer's disease (study period June 1, 2013 through March 31, 2019), while the White non-Hispanic sample was comprised of caregivers to cancer patients (study period September 30, 2006 through July 31, 2013). A larger proportion of the Hispanic (83%) as contrasted with the non-Hispanic caregiver sample (55%) was female and younger (74% were below age 65 as contrasted with 62% of the non-Hispanic Whites; see Table 1). Among the Hispanic caregiver sample, 45% had some post-high school education and 24% had 0–11 years, as contrasted with the White non-Hispanic sample for which only 11% had less than high school education. More of the Hispanic sample of caregivers (77%) than the White non-Hispanic caregiver sample (54%) lived with the patient. The average age of the non-Hispanic White care recipients was 60.7 (11.6) as contrasted with the Hispanic care recipients with an average age of 79.9 (8.9).

This study was approved by the Institutional Review Board (IRB) at Mount Sinai Medical Center (study reported at https://projectreporter.nih.gov/project_info_description.cfm?aid=7892314) and at Columbia University Medical Center (protocols IRB-AAAL7251, IRB-AAAM5150), reported at https://projectreporter.nih.gov/project_info_description.cfm?aid=9251192&icde=43514731&ddparam=&ddvalue=&ddsub=&cr=10&csb=default&cs=ASC&MMOpt=.

Table 2. Eigenvalues from the exploratory factor analysis using principal components estimation and fit indices from confirmatory factor analyses^a (MPlus)

Statistic	Component 1	Component 2	Ratio component 1/ component 2
Total sample (<i>n</i> = 1,834) CFI (0.991); TLI (0.988); RMSEA (0.101); IRT (IRTPRO) RMSEA (0.09) ^b			
Eigenvalues	7.788	0.399	19.5
Explained variance	77.9%	4.0%	
Non-Hispanic Whites (<i>n</i> = 1,517) CFI (0.988); TLI (0.984); RMSEA (0.101); IRT (IRTPRO) RMSEA (0.09) ^b			
Eigenvalues	7.424	0.460	16.1
Explained variance	74.2%	4.6%	
Hispanics (<i>n</i> = 317) CFI (0.997); TLI (0.996); RMSEA (0.097); IRT (IRTPRO) RMSEA (0.08) ^b			
Eigenvalues	8.850	0.261	33.9
Explained variance	88.5%	2.6%	

Model fit statistics: comparative fit index (CFI); Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA) from MPlus and RMSEA from IRTPRO.

^aGeomin (oblique) rotation and fit statistics for one factor solutions.

^bBased on M_2 statistics which are based on full marginal tables.

Results

Qualitative

The DIF hypotheses were posited with respect to race/ethnicity and language. Although the majority (two-thirds) were interviewed in Spanish, the sample size was too small to examine language within the Hispanic subgroup. Thus, the hypotheses regarding ethnicity were relevant to these analyses. With respect to race/ethnicity, 5 items out of 10 were hypothesized to evidence DIF, however only 2 with a direction given: "The way the family is included in treatment and care decisions" and "Information given about the patient's tests." These were hypothesized to be more likely endorsed in the dissatisfied direction, conditional on the trait by minority than by White respondents.

Quantitative

Model assumptions: As shown by the eigenvalue ratios in Table 2, there was strong support for essential unidimensionality for the total sample and both subgroups, Hispanic and non-Hispanic White responders. All three ratios of component 1–2 were large (total sample — 19.5; non-Hispanic White responders — 16.1; Hispanic responders — 33.9). The first component accounted for between 74% and 89% of the variance for all groups, supporting the essential unidimensionality of the item set across comparison subgroups. The RMSEA index from the MPlus analysis was 0.10 for the total sample and for both demographic groups. The RMSEA indices from the IRTPRO estimation were slightly lower ranging from 0.08 to 0.09. The CFIs ranged from 0.988 to 0.997. The ECVs ranged from 92.66 to 96.77 (see Table 3).

In general, the LD statistics (Chen and Thissen, 1997) were in the acceptable range for Hispanics, and over the threshold for the non-Hispanic White sample. There were five instances of LDs above 10 for the White non-Hispanic sample (see Appendix

Table 3. Reliability and dimensionality estimates

	<i>n</i>	Ordinal alpha	McDonald's omega total	Explained common variance (ECV)
Total sample	1,834	0.968	0.968	93.687
Hispanic respondents	1,517	0.985	0.986	96.770
Non-Hispanic White respondents	317	0.961	0.962	92.662

All analyses based on polychoric correlations.

Table A1): items 2 (availability of doctors) and 8 (doctor assesses symptoms; 15.9); items 3 (coordination of care) and 4 (time required to make diagnosis; 13.2); items 5 (families included in treatment) and 8 (doctor assesses symptoms; 14.6); items 6 (information given about management of pain) and 10 (availability of the doctor; 14.5); and items 9 (tests and treatments followed up by doctor) and 10 (availability of the doctor; 12.2). These values did not appear to inflate the magnitude of the discrimination parameters, and the values were relatively low; thus, it was concluded that they did not require action.

The reliability estimates were high. The omega total values ranged from 0.962 to 0.986, and the ordinal alphas ranged from 0.961 to 0.985 (see Table 3). The classical test theory estimated Cronbach's alpha for the total sample was 0.95 for both non-standardized and standardized calculations. The corrected item-total correlations ranged from 0.72 to 0.83 (see Appendix Table A2). The internal consistency for those interviewed in English and Spanish were 0.96 and 0.97, respectively.

IRT-based reliability: The reliability estimates calculated along the satisfaction continuum were >0.90 in the range of theta from -2.0 to 0.8. Estimates were slightly lower at the dissatisfied tail (0.80, 0.83, 0.84 across the total, non-Hispanic White, and Hispanic subsamples) as well as the very satisfied range of the distribution. The overall reliability estimates were 0.90 for the total sample, 0.91 for the non-Hispanic White, and 0.93 for the Hispanic subgroup (see Table 4).

The information function for the items and overall scale for the total sample were bimodal with the highest peaks at theta levels -1.2 and 0.4. The most informative item was "The way tests and treatments are followed up by the doctor" (item 9), and the least informative item was "Coordination of care" (item 3; see Figure 1).

The analyses of DIF showed that three items evidenced DIF consistently by two methods: IRTPRO and *lordif* (see Table 5 and Appendix Table A3). However, only one item was flagged as significant by both methods. After the Bonferroni adjustment, non-uniform DIF was flagged with IRTPRO for the item, "The way the family is included in treatment and care decisions" (item 5). The item was more discriminating (more highly related to the satisfaction state) for the non-Hispanic, White responders than for the Hispanic subsample, and was also a more severe indicator (higher difficulty parameter) for this group at specific levels of the trait; the non-Hispanic White responders were less satisfied at higher levels of the satisfaction trait.

The items "Information given about how to manage the patient's pain" (item 6) and "Information given about the patient's tests" (item 7) were identified with uniform DIF by IRTPRO; however, the result was not significant after application of the Bonferroni adjustment for multiple comparisons. The item "Information given about patient's tests" was also flagged for uniform DIF by *lordif*. *Lordif* identified non-uniform DIF for both items, after the adjustment; the items were more discriminating for the Hispanic responders (see Appendix Figure A1). The

Table 4. IRT reliability estimates at varying levels of the attribute (theta) estimate based on results of the IRT analysis (IRTPRO)

IRT reliability			
Satisfaction (theta)	Total sample (<i>n</i> = 1,834)	Non-Hispanic White responders (<i>n</i> = 1,517)	Hispanic responders (<i>n</i> = 317)
-2.4	0.80	0.83	0.84
-2.0	0.92	0.92	0.93
-1.6	0.97	0.96	0.96
-1.2	0.97	0.96	0.96
-0.8	0.94	0.93	0.92
-0.4	0.94	0.92	0.90
0.0	0.97	0.95	0.94
0.4	0.97	0.96	0.96
0.8	0.94	0.95	0.95
1.2	0.85	0.90	0.91
1.6	0.70	0.79	N/A
Overall (average)	0.90	0.91	0.93

Note: Reliability estimates were calculated for theta levels for which there are respondents.

magnitude of DIF was small; all NCDIF and T1 statistics were below threshold (see Table 5). The impact of DIF was negligible, as shown by the overlapping curves (see Appendix Figure A2).

Language sensitivity analysis: Sensitivity DIF analysis was performed comparing the White non-Hispanic group to Spanish-speaking Hispanics alone (see Appendix Table A4). The results were similar to those of the main analyses. Three items showed DIF, two the same as in prior analysis. No DIF comparisons were significant after the Bonferroni correction.

Discussion

The FAMCARE scale, although extensively used to assess satisfaction with care for cancer patients, has also been applied to palliative care, including caregivers to patients with Alzheimer's disease. The psychometric properties of the FAMCARE have been examined in cancer patients in diverse settings internationally, including the relationship of demographic characteristics to satisfaction with end-of-life care. However, little evidence exists concerning measurement equivalence across ethnically diverse groups, particularly in Hispanic samples.

These analyses identified only one item with consistent DIF after Bonferroni correction: item 5, "The way the family is included in treatment decisions." No items evidenced salient DIF.

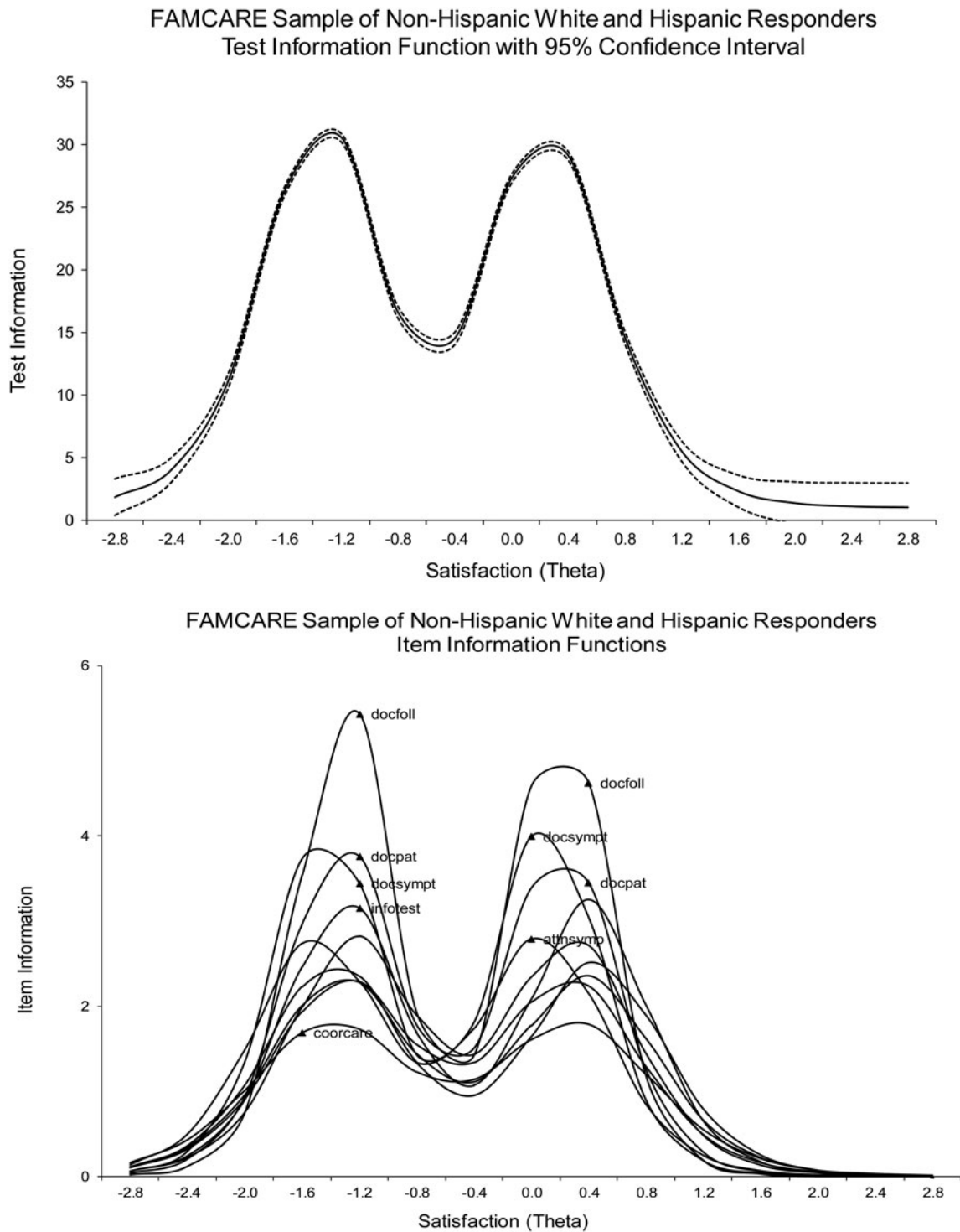


Fig. 1. FAMCARE: scale and item information functions.

Although the two groups examined in this study differ in disease type, we argue that the two groups have in common that they are caregivers to individuals with serious illness and poor prognosis. The diseases are different; however, it was not posited that the different diseases would result in DIF. It was posited that cultural and language differences can have an impact on item meaning and response. An advantage of IRT is that it produces arguably more invariant parameters that can be compared because they

are sample independent. Philosophically, DIF can be examined with IRT across many groups differing in socio-demographic characteristics; however, it is important to present a rationale for such analyses.

Examination of the hypotheses for the qualitative analyses in conjunction with the quantitative analyses showed that two items were posited to evidence DIF for ethnic/race groups. In general, minority groups were hypothesized to express less

Table 5. Summary of DIF hypotheses and analyses

Item	DIF hypotheses ^a	IRTPRO	lordif	Magnitude (NCDIF)	Effect size (T1)
1. Doctor's attention to the patient's description of symptoms	3			0.0017	0.0270
2. Availability of doctors to the family				0.0001	-0.0013
3. Coordination of care	4		U	0.0017	-0.0146
4. Time required to make diagnosis				0.0010	0.0000
5. The way the family is included in treatment and care decisions	7 Minority group less satisfied	NU* U*	U*	0.0057	-0.0450
6. Information given about how to manage the patient's pain		U	NU*	0.0031	0.0074
7. Information given about the patient's tests	4 Minority group less satisfied	U	NU* U*	0.0047	-0.0407
8. How thoroughly the doctor assesses the patient's symptoms	4			0.0020	0.0213
9. The way tests and treatments are followed up by the doctor				0.0044	0.0342
10. Availability of the doctor to the patient				0.0005	0.0095

All NCDIF values were smaller than the threshold (0.0240); the range was from 0.0001 to 0.0057 and none of the T1 statistics were significant.

NU, non-uniform DIF involving the discrimination parameters; U, uniform DIF involving the location parameters.

^aThe numbers in bold are the number positing DIF. Not all provided a direction to the hypothesis; only those with a direction are presented.

*Significant after Bonferroni correction.

satisfaction than White groups, conditional on overall satisfaction. Content experts posited directional race/ethnicity hypotheses for the item that evidenced consistent DIF: "The way the family is included in treatment and care decisions" (item 5). It was posited that minority group members would be less satisfied, conditional on the trait. Contrary to the hypotheses, item 5 showed non-uniform DIF, and the uniform DIF observed was in the opposite direction of that hypothesized. As noted, this item did not reach the criteria for salient DIF. Because the experts used their clinical experience when establishing hypotheses, it is possible that they took into account the potential language barrier when suggesting lower satisfaction for Hispanics, in contrast to their White counterparts. It may be that they felt, even at the same levels of satisfaction, Hispanics might respond in a more dissatisfied direction because of general health disparities and health care disparities, both real and perceived. Although there is no literature on the FAMCARE in a sample of Hispanic caregivers to persons with ADRD, earlier work on ethnically diverse caregivers may have informed the hypotheses. In contrast to the findings reported here, in an earlier paper on DIF in the FAMCARE (Teresi et al., 2015), Black responders reported less satisfaction with their care, conditional on the trait.

The non-uniform DIF observed showed that conditional on overall satisfaction, the reported satisfaction for Hispanics was not constant (see the crossing item response curves); thus, supporting the dissatisfied direction posited by the experts for some satisfaction levels. This hypothesis is consistent with research evidence suggesting that Hispanics tend to endorse the extreme response categories in surveys (Clarke, 2000) potentially due to cultural values that relate such response style with demonstrating trustworthiness (McHorney and Fleishman, 2006).

A confirmatory directional hypothesis was not given for the item related to information about management of pain.

However, in an earlier study a similar item, "Satisfaction with the patient's pain relief" was found to show DIF for the comparison of Blacks and White non-Hispanics. In that study, it was found that conditional on the satisfaction level, caregivers of Black patients were less satisfied with pain relief (Teresi et al., 2015), a finding corresponding to findings of racial and ethnic disparities in pain treatment identified by Green et al. (2003). It is possible that the content experts posited the presence of an unmeasured secondary extraneous factor such as personal experiences that may have influenced responses to satisfaction items.

Strengths and limitations

Limitations of the study include the small number of Hispanics interviewed in English which did not permit systematic analyses of this group. The inability to perform other subgroup analyses due to sample size restrictions is also a limitation. As pointed out by a reviewer, the overlapping information curves and high corrected item-total correlations may be indicative of redundancy in the item set for this sample. IRT-based reliability estimates provided at varying points along the satisfaction trait continuum yielded somewhat lower reliability estimates, particularly at the tails of the distribution. Thus, while omnibus summary reliability estimates appear to show uniform item performance, the scale was not uniformly reliable across the trait; however, it is emphasized that estimates were above 0.80 for nearly all theta points for which reliability was estimated.

Strengths of the study include the provision of information for placement in an item bank on family satisfaction and care transitions. Such a bank was used to develop the short-form of the FAMCARE (Ornstein et al., 2015) used in these analyses. Additionally, the short-form version developed with IRT was used to develop a Japanese translation (Ito and Tadaka, 2018). This

study is the first to examine the measurement equivalence of the FAMCARE scale in a sample of Hispanic caregivers to patients with AD/DR using latent variable models. This paper provides information on DIF for inclusion in an existing item bank on family satisfaction with care and care transitions. Additionally, reliability estimates indicated that the scale was highly reliable (estimates ≥ 0.90). Most items provided adequate information, although the item related to care coordination was less informative.

In summary, the analyses showed modest DIF of low magnitude and impact for the Hispanic sample in comparison to a White non-Hispanic sample. The item flagged related to information sharing: the way the family is included in treatment and care decisions. No items rose to the level of salient DIF of high magnitude or impact. Evidence from this study supports the measurement equivalence of the FAMCARE among Hispanics interviewed in Spanish and English. Thus, the short-form FAMCARE can be recommended for use in cross-cultural assessments and research involving such groups.

Authorship. J.A.T. substantially contributed to the design of the work, oversaw analyses, and drafted the article. K.O.-W. performed analyses and participated in drafting the article. M.R. contributed to the design, qualitative analyses, and review of the article. M.K. performed analyses. K.O. contributed to the design of the work and reviewed the manuscript. A.S. and J.L. acquired the data and participated substantially in the work. All authors have approved the publication of the article.

Acknowledgments. The authors thank Stephanie Silver, MPH for her expert editing of the manuscript.

Funding. Support for these analyses was provided by a collaboration between the Claude Pepper Older Americans Independence Center: National Institute on Aging (grant number 1P30AG028741) and the National Institute on Aging Alzheimer's Disease Resource Center on Minority Aging Research (grant number 1P30AG059303). The studies from which data were supplied were funded by the Patient-Centered Outcomes Research Institute (PCORI) (contract number CE-1304-7160) and the National Institute of Nursing Research (NINR) (grant number 1R01NR0114430-01) and the National Cancer Institute (NCI) (grant number 5R01CA116227-059999).

Conflict of interest. The authors declare that there is no conflict of interest with respect to the research, authorship, and/or publication of this article.

References

- Aoun S, Bird S, Kristjanson LJ, et al. (2010) Reliability testing of the FAMCARE-2 scale: Measuring family care satisfaction with palliative care. *Palliative Medicine* 24(7), 674–681.
- Asparouhov T and Muthén B (2009) Exploratory structural equation modeling. *Structural Equation Modeling* 16, 397–438.
- Benjamini Y and Hochberg Y (1995) Controlling for the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57, 289–300. doi:10.2307/2346101
- Bentler PM (1990) Comparative fit indexes in structural models. *Psychological Bulletin* 107(2), 238–246. doi:10.1037/0033-2909.107.2.238
- Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8, 3–62.
- Cai L, Thissen D and du Toit SHC (2011) *IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling (Computer Software)*. Chicago, IL: Scientific Software International, Inc.
- Chen WH and Thissen D (1997) Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics* 22, 265–289.
- Choi SW, Gibbons LE and Crane PK (2011) lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software* 39, 1–30.
- Clarke I (2000) Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality* 15, 137–152.
- Cook KF, Kallen MA and Amtmann D (2009) Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research* 18, 447–460. doi:10.1007/s11136-009-9464-4
- D'Angelo D, Punziano AC, Mastroianni C, et al. (2017) Translation and testing of the Italian version of FAMCARE-2: Measuring family caregivers' satisfaction with palliative care. *Journal of Family Nursing* 23(2), 252–272.
- Fleer PF (1993) *A Monte Carlo Assessment of a New Measure of Item and Test Bias* (Dissertation, Dissertation Abstracts International, 54-04B, 2266). Illinois Institute of Technology, Chicago, IL.
- Flowers CP, Oshima TC and Raju NS (1999) A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement* 23, 309–332.
- Green CR, Anderson KO, Baker TA, et al. (2003) The unequal burden of pain: Confronting racial and ethnic disparities in pain. *Pain Medicine* 4(3), 277–294.
- Hambleton RK, Swaminathan H and Roger HJ (1991) *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.
- Hwang SS, Chang VT, Alejandro Y, et al. (2003) Caregiver unmet needs, burden, and satisfaction in symptomatic advanced cancer patients at a Veterans Affairs (VA) medical center. *Palliative & Supportive Care* 1, 319–329.
- Ito E and Tadaka E (2018) Development of a Japanese version of the short-form FAMCARE scale for family caregivers of terminal cancer patients at home in Japan. *Nippon Ronen Igakkai Zasshi. Japanese Journal of Geriatrics* 55(1), 81–89.
- Kim S, Cohen AS, Alagoz C, et al. (2007) DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement* 44, 93–116. doi:10.1111/j.1745-3984.2007.00029.x
- Kleinman M and Teresi JA (2016) Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling* 58(1), 79–98.
- Kristjanson LJ (1986) Indicators of quality of palliative care from a family perspective. *Journal of Palliative Care* 1(2), 8–17.
- Kristjanson LJ (1989) Quality of terminal care: Salient indicators identified by families. *Journal of Palliative Care* 5(1), 21–30.
- Kristjanson LJ (1993) Validity and reliability testing of the FAMCARE Scale: Measuring family satisfaction with advanced cancer care. *Social Science & Medicine* 36(5), 693–701.
- Ljungberg AK, Fossum B, First CJ, et al. (2015) Translation and cultural adaptation of research instruments – Guidelines and challenges: An example in FAMCARE-2 for use in Sweden. *Informatics for Health and Social Care* 40, 67–78. doi:10.3109/17538157.2013.87211
- Lo C, Burman D, Rodin G, et al. (2009) Measuring patient satisfaction in oncology palliative care: Psychometric properties of the FAMCARE-patient scale. *Quality of Life Research* 18, 747–752.
- Lord FM (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord FM and Novick MR (1968) *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Co.
- McDonald RP (1999) *Test Theory: A Unified Treatment*. Mahwah, NJ: L. Erlbaum Associates.
- McHorney C and Fleishman J (2006) Assessing and understanding measurement equivalence in health outcome measures: Issues for further quantitative and qualitative inquiry. *Medical Care* 44(Suppl 3), S205–S210. doi:10.1097/01.mlr.0000245451.67862.57
- Morales LS, Flowers C, Gutierrez P, et al. (2006) Item and scale differential functioning of the Mini-Mental State Exam assessed using the Differential Item and Test Functioning (DFIT) framework. *Medical Care* 44(11), S143–S151.
- Muthén LK and Muthén BO (2011) *M-PLUS Users Guide*, 6th ed. Los Angeles, CA: Muthén and Muthén, pp. 1998–2011.
- Orlando-Edelen M, Thissen D, Teresi JA, et al. (2006) Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. *Medical Care* 44, S134–S142.

- Ornstein KA, Teresi JA, Ocepek Welikson K, et al.** (2015) Use of an item bank to develop two short-form FAMCARE scales to measure family satisfaction with care in the setting of serious illness. *Journal of Pain and Symptom Management* **49**(5), 894–903.
- Raju NS, van der Linden WJ and Fleer PF** (1995) IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement* **19**, 353–368.
- R Core Team** (2018) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>
- Reise SP** (2012) The rediscovery of bifactor measurement models. *Multivariate Behavioral Research* **47**, 667–696. doi:10.1080/00273171.2012.715555
- Reise SP, Moore TM and Haviland MG** (2010) Bi-factor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment* **92**, 544–559. doi:10.1080/00223891.2010.496477
- Rodriguez KL, Bayliss NK, Jaffe E, et al.** (2010) Factor analysis and internal consistency evaluation of the FAMCARE Scale for use in the long-term care setting. *Palliative & Supportive Care* **8**(2), 169–176.
- Samejima F** (1969) *Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometrika Monograph; Supplement 17)*. Dordrecht: Springer.
- Sijtsma K** (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* **74**, 107–120. doi:10.1007/s11336-008-9101-0
- Teresi JA, Kleinman M and Ocepek-Welikson K** (2000) Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine* **19**, 1651–1683.
- Teresi J, Ocepek-Welikson K, Kleinman M, et al.** (2007) Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research* **16**, 43–68. doi:10.1007/s11136-007-9186-4
- Teresi JA, Ornstein K, Ramirez M, et al.** (2014) Performance of the Family Satisfaction with the End-of-Life Care (FAMCARE) measure in an ethnically diverse cohort: Psychometric analyses using item response theory. *Supportive Care in Cancer* **22**, 399–408.
- Teresi JA, Ocepek-Welikson K, Ramirez M, et al.** (2015) Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care in an ethnically diverse cohort: Tests of differential item functioning. *Palliative Medicine* **29**, 83–96.
- Teresi JA, Ocepek-Welikson K, Ramirez M, et al.** (2019) Psychometric properties of a Spanish-language version of a short-form FAMCARE: Applications to caregivers of patients with Alzheimer's disease and related dementias. *Journal of Family Nursing* **25**(4), 557–589.
- Thissen D, Steinberg L and Wainer H** (1993) Detection of differential item functioning using the parameters of item response models. In Holland PW and Wainer H (eds), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, Inc.
- Thissen D, Steinberg L and Kuang D** (2002) Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false discovery rate in multiple comparisons. *Journal of Educational and Behavioral Statistics* **27**, 77–83. doi:10.3102/10769986027001077
- Tucker LR and Lewis C** (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* **38**, 1–10. doi:10.1007/BF02291170
- Wainer H** (1993) Model-based standardization measurement of an item's differential impact. In Holland PW and Wainer H (eds), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, Inc., pp. 123–135.
- Wang W-C, Shih C-L and Sun G-W** (2012) The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement* **72**, 687–708. doi:10.1177/0013164411426157
- Woods CM** (2009) Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement* **33**, 42–57. doi:10.1177/0146621607314044

Appendix

Table A1. Local dependency statistics (bolded entries are slightly above the threshold for elevation).

Item	Label	Marginal χ^2	1	2	3	4	5	6	7	8	9
Marginal fit (χ^2) and Standardized LD χ^2 Statistics (Hispanics only, $n = 317$)											
1	attnsymp	1.8									
2	availdoc	0.4	3.4								
3	coorcare	0.8	2.3	3.5							
4	timediag	1.1	4.8	4.3	7.7						
5	famincl	1.6	7.6	5.4	7.8	3.0					
6	infomang	1.8	7.9	5.6	2.9	2.6	8.4				
7	infotest	2.8	5.5	4.9	8.6	9.9	10.3	6.7			
8	docsymp	2.6	6.0	3.4	5.5	5.9	9.2	5.9	9.7		
9	docfoll	3.0	2.5	3.0	4.6	2.9	3.2	3.4	6.8	9.8	
10	docpat	2.5	3.8	3.6	10.1	7.1	4.1	2.1	7.0	3.4	7.6
Marginal fit (χ^2) and Standardized LD χ^2 Statistics (Whites only, random sample of 300)											
1	attnsymp	1.3									
2	availdoc	0.7	6.0								
3	coorcare	1.1	2.7	4.9							
4	timediag	0.6	5.1	5.2	13.2						
5	famincl	0.7	8.7	3.8	2.0	9.8					
6	infomang	0.4	2.6	4.7	7.2	5.5	9.4				
7	infotest	0.5	3.4	2.3	4.8	5.5	5.3	8.4			
8	docsymp	2.6	4.9	15.9	2.3	3.0	14.6	8.7	7.2		
9	docfoll	2.5	2.3	3.5	4.2	2.8	5.4	8.2	3.0	5.2	
10	docpat	2.1	2.6	6.9	4.8	8.3	6.5	14.5	5.3	5.3	12.2

Table A2. Classical test reliability estimates (SPSS): total sample ($n = 1,834$)

Item	Mean (SD)	Corrected item-total correlation	Cronbach's alpha if item deleted
Doctor's attention to patient's description of symptoms	1.36 (0.62)	0.76	0.94
Availability of doctors to the family	1.25 (0.66)	0.78	0.94
Coordination of care	1.26 (0.64)	0.72	0.94
Time required to make diagnosis	1.23 (0.63)	0.76	0.94
The way the family is included in treatment and care decisions	1.26 (0.65)	0.75	0.94
Information given about how to manage the patient's pain	1.22 (0.61)	0.76	0.94
Information given about the patient's tests	1.23 (0.62)	0.79	0.94
How thoroughly the doctor assesses the patient's symptoms	1.34 (0.61)	0.80	0.94
The way tests and treatments are followed up by the doctor	1.29 (0.62)	0.83	0.94
Availability of the doctor to the patient	1.29 (0.63)	0.80	0.94
Cronbach's alpha (standardized alpha)	0.946 (0.947)		

Table A3. IRT item parameters and DIF statistics for Hispanic compared to non-Hispanic White responders (reference group)

Item name	Group	<i>a</i>	<i>b</i> ₁	<i>b</i> ₂	<i>a</i> DIF ^a	<i>b</i> DIF ^a
Doctor's attention to the patient's description of symptoms	Non-Hispanic Whites	2.98	-1.56	0.21	NS, Anchor item	
	Hispanics	(0.15)	(0.06)	(0.04)		
Availability of doctors to the family	Non-Hispanic Whites	2.96	-1.27	0.44	NS, Anchor item	
	Hispanics	(0.14)	(0.05)	(0.05)		
Coordination of care	Non-Hispanic Whites	2.36	-1.46	0.47	NS, Anchor item	
	Hispanics	(0.11)	(0.06)	(0.05)		
Time required to make diagnosis	Non-Hispanic Whites	2.69	-1.39	0.56	NS, Anchor item	
	Hispanics	(0.13)	(0.05)	(0.05)		
The way the family is included in treatment and care decisions	Non-Hispanic Whites	2.84	-1.34	0.46	8.6 (0.0034)^b	24.1 (0.0001)^b
	Hispanics	2.06	-1.43	0.10		
Information given about how to manage the patient's pain	Non-Hispanic Whites	2.75	-1.47	0.67	0.1 (0.7259)	9.4 (0.0091)
	Hispanics	2.89	-1.23	0.44		
Information given about the patient's tests	Non-Hispanic Whites	3.16	-1.34	0.62	0.2 (0.6588)	8.9 (0.0117)
	Hispanics	3.38	-1.41	0.29		
How thoroughly the doctor assesses the patient's symptoms	Non-Hispanic Whites	3.65	-1.48	0.25	NS, Anchor item	
	Hispanics	(0.19)	(0.05)	(0.04)		
The way tests and treatments are followed up by the doctor	Non-Hispanic Whites	4.27	-1.34	0.35	DIF not significant	
	Hispanics	(0.24)	(0.05)	(0.04)		
Availability of the doctor to the patient	Non-Hispanic Whites	3.54	-1.36	0.36	NS, Anchor item	
	Hispanics	(0.18)	(0.05)	(0.04)		

"NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

^aStatistical test for differences in parameters is Wald test using 1 df for the test of differences in the *a* parameters for the comparison groups and 2 df for the test of differences in the *b* parameters.

^bBolded entries indicate items that evidence DIF after correction for multiple comparisons.

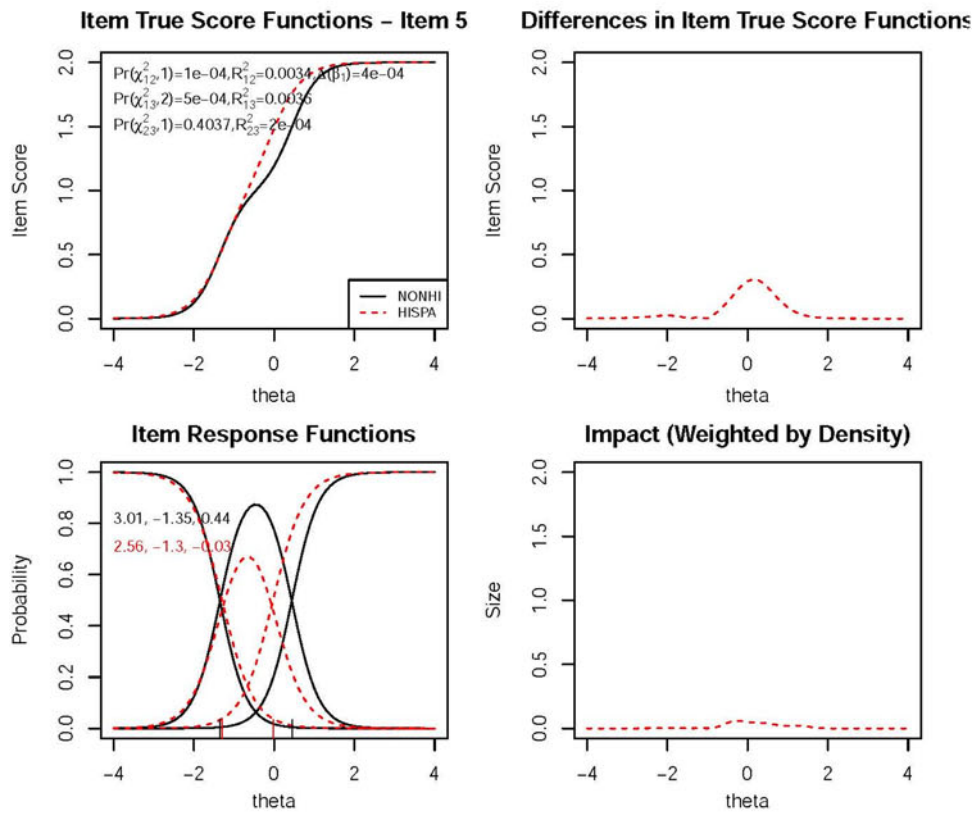
Table A4. Sensitivity analyses: summary of DIF analyses comparing White non-Hispanic subsample with Spanish-speaking Hispanics only

Item	IRTPRO
Doctor's attention to the patient's description of symptoms	
Availability of doctors to the family	
Coordination of care	
Time required to make diagnosis	NU (<i>P</i> = 0.0221)
The way the family is included in treatment and care decisions	U (<i>P</i> = 0.0089)
Information given about how to manage the patient's pain	U (<i>P</i> = 0.0091)
Information given about the patient's tests	
How thoroughly the doctor assesses the patient's symptoms	
The way tests and treatments are followed up by the doctor	
Availability of the doctor to the patient	

NU, non-uniform DIF involving the discrimination parameters; U, uniform DIF involving the location parameters.

Note: No items were significant after correction for multiple comparisons.

The way the family is included in treatment and care decisions



Information given about how to manage the patient's pain

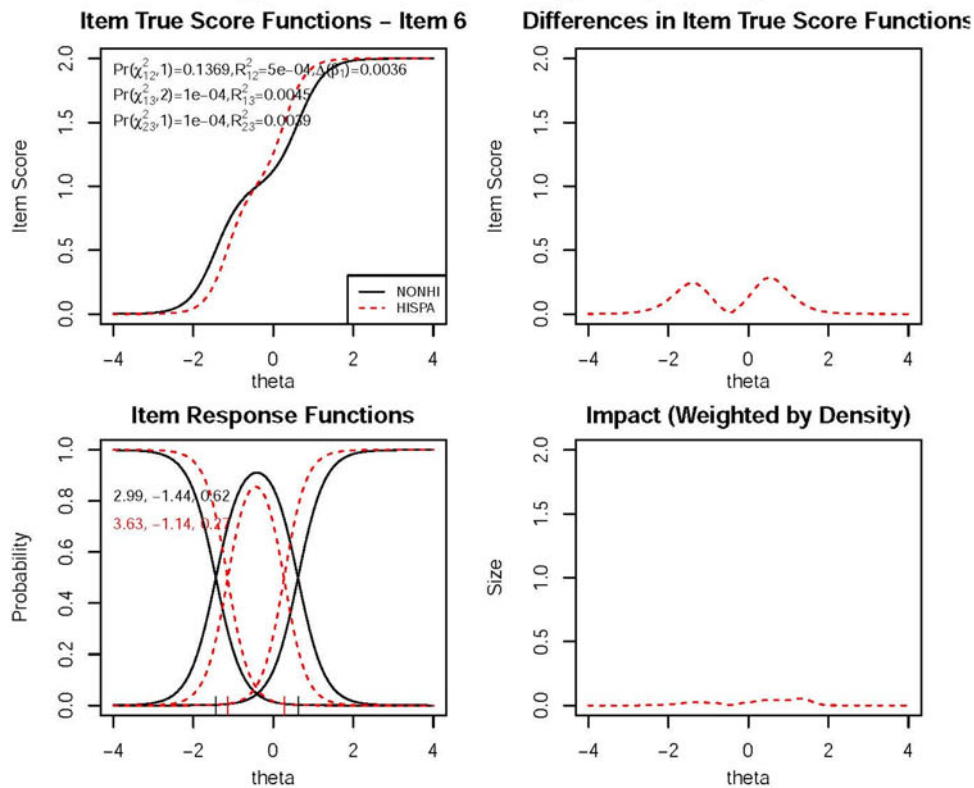


Fig. A1. Item response functions and magnitude of DIF.

Information given about the patient's tests

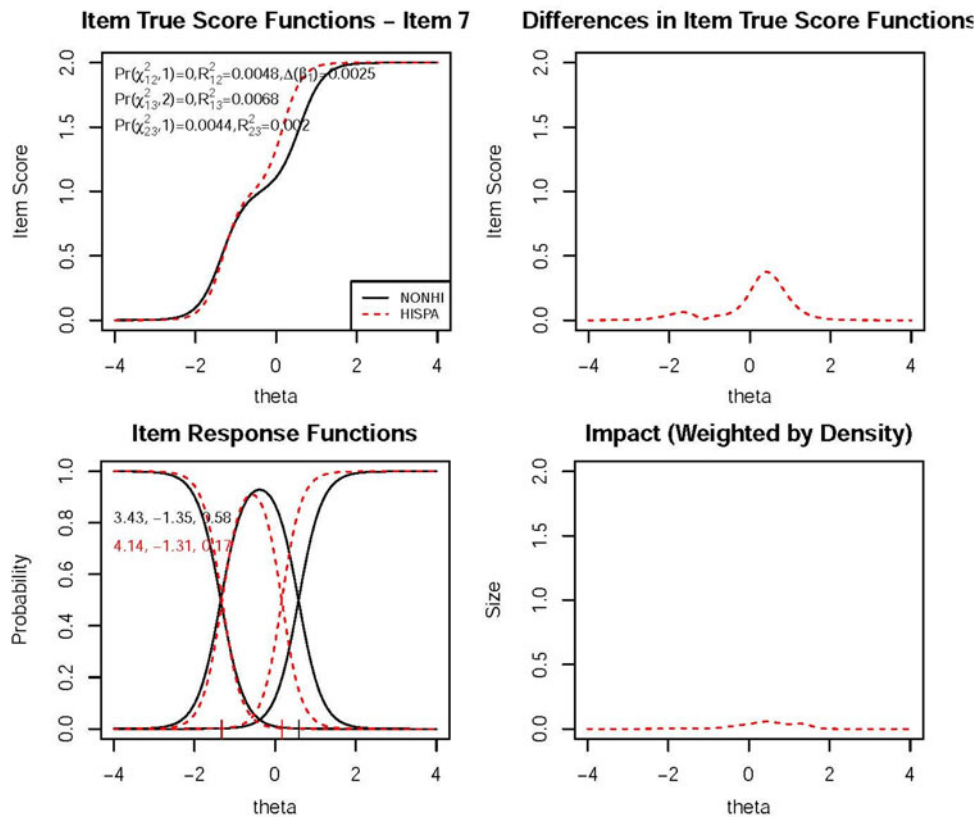


Fig. A1. Continued.

Note: Results are from lordif software. For each item, the upper left panel shows the expected item score plots (denoted item true score functions) for Hispanics and non-Hispanic Whites. The lower left panel shows the item characteristic curves (category response functions). The upper right panel displays the absolute group differences in expected item scores. The lower right panel shows the differences weighted by density and is indicative of the magnitude (impact) of DIF at the item level. This measure is related to the non-compensatory DIF statistic (NCDIF) described in the text.

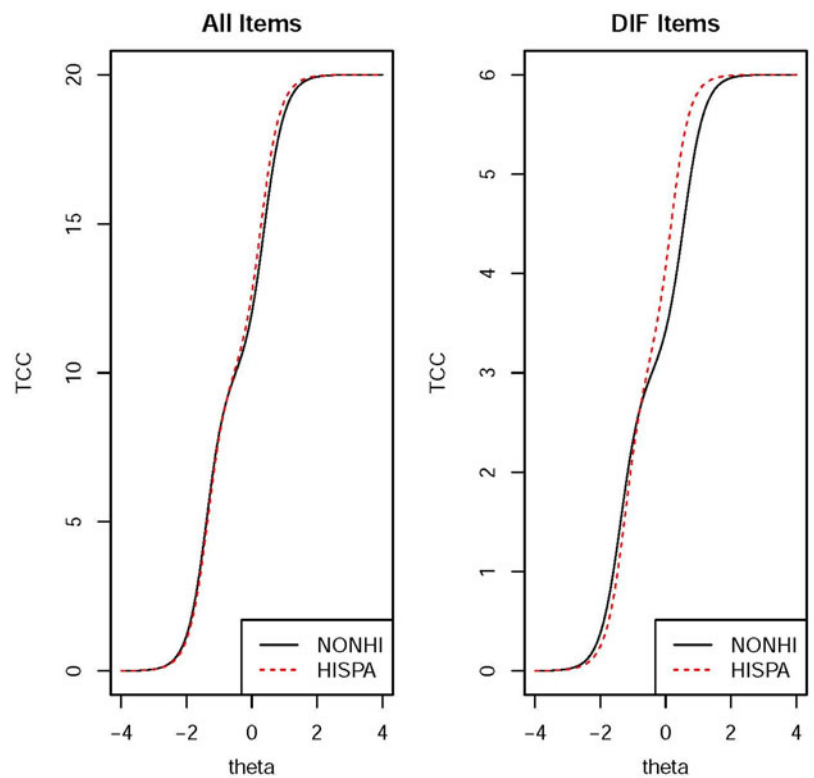


Fig. A2. Impact of DIF at the scale level: expected scale scores.