Published in final edited form as:

Palliat Med. 2015 January ; 29(1): 83–96. doi:10.1177/0269216314545802.

Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care in an ethnically diverse cohort: Tests of differential item functioning

Jeanne A Teresi[1,2,3], **Katja Ocepek-Welikson**[1], **Mildred Ramirez**[1,2,4], **Marjorie Kleinman**[3], **Katherine Ornstein**[5], and **Albert Siu**[6]

[1]Research Division, The Hebrew Home at Riverdale, Riverdale, NY, USA

[2]Measurement and Data Management Core, Mount Sinai Pepper Older Americans Independence Center, Mount Sinai Medical Center, New York, NY, USA

[3]Columbia University Stroud Center, New York State Psychiatric Institute, New York, NY, USA

[4]Division of Geriatrics and Palliative Care, Weill Cornell Medical Center, New York, NY, USA

[5]Department of Geriatrics and Palliative Medicine, Institute for Translational Epidemiology, Mount Sinai School of Medicine, New York, NY, USA

[6]Divisions of Geriatrics and Palliative Medicine, General Internal Medicine, Health Evidence and Policy, Mount Sinai Medical Center, New York, NY, USA

Abstract

Background—The Family Satisfaction with End-of-Life Care is an internationally used measure of satisfaction with cancer care. However, the Family Satisfaction with End-of-Life Care has not been studied for equivalence of item endorsement across different socio-demographic groups using differential item functioning.

Aims—The aims of this secondary data analysis were (1) to examine potential differential item functioning in the family satisfaction item set with respect to type of caregiver, race, and patient age, gender, and education and (2) to provide parameters and documentation of differential item functioning for an item bank.

Design—A mixed qualitative and quantitative analysis was conducted. A priori hypotheses regarding potential group differences in item response were established. Item response theory and Wald tests were used for the analyses of differential item functioning, accompanied by magnitude and impact measures.

Results—Very little significant differential item functioning was observed for patient's age and gender. For race, 13 items showed differential item functioning after multiple comparison

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Mildred Ramirez, Research Division, The Hebrew Home at Riverdale, 5901 Palisade Avenue, Riverdale, NY 10471, USA. MilRamirez@aol.com.

Declaration of conflict of interests
The authors declare that there is no conflict of interest.

adjustment, 10 with non-uniform differential item functioning. No items evidenced differential item functioning of high magnitude, and the impact was negligible. For education, 5 items evidenced uniform differential item functioning after adjustment, none of high magnitude. Differential item functioning impact was trivial. One item evidenced differential item functioning for the caregiver relationship variable.

Conclusion—Differential item functioning was observed primarily for race and education. No differential item functioning of high magnitude was observed for any item, and the overall impact of differential item functioning was negligible. One item, satisfaction with “the patient's pain relief,” might be singled out for further study, given that this item was both hypothesized and observed to show differential item functioning for race and education.

Keywords

Family Satisfaction with End-of-Life Care; differential item functioning; item response theory; ethnic diversity; palliative care; item bank

Conceptual and psychometric measurement equivalence of scales is a basic requirement for valid cross-cultural and demographic subgroup comparisons. The Family Satisfaction with End-of-Life Care (FAMCARE) scale, developed in Australia to measure satisfaction with cancer care,[1,2] has been used extensively to assess satisfaction with palliative care.[2,3] The psychometric properties of the scale have been examined with cancer patients in diverse settings internationally (e.g. North America, Australia, and Europe); however, little or no evidence is available about the performance of the measure across ethnically diverse groups. The psychometric analyses performed within samples have resulted in varied recommendations. Principal component and bifactor analyses have supported essential unidimensionality, with one strong factor reflecting a single underlying attribute.[4–7] However, the results of one study in Australia identified a four-factor structure.[8] Shortened versions of the scale based on psychometric analyses have been suggested, for example, a 19-item version for terminal cancer victims in Norway.[4] Adaptations of the scale have also been recommended or used in (1) inpatient settings in Australia (FAMCARE-2, 17-item version[8]), (2) outpatient oncology palliative care settings among family members in Australia (FAMCARE-6, 6 items[5]) or patients in Canada (FAMCARE-P16, 16-items[9]), and (3) long-term care settings in the United States (18-item version[6]). The different versions of FAMCARE have shown adequate estimates of internal consistency using Cronbach's alpha and other reliability statistics across diverse cancer patient samples and settings.[5,7,10]

Few studies have examined the relationship of demographic characteristics to satisfaction with care. Johnsen et al.,[11] using ordinal logistic regression analyses of individual items, found that age was the only variable associated consistently with dissatisfaction; that is, younger relatives in a Danish sample were more dissatisfied with care than older relatives with respect to 17 of the original 20 items. Another study conducted in Australia found that older, female caregivers and those with no strong ethnic identification reported higher average satisfaction scores than younger, male, ethnically identified individuals.[8] Similarly, Kristjanson,[10] using an Australian sample, reported race, education level, and patient's age as significant correlates of satisfaction; that is, White caregivers with higher education caring for older patients evidenced higher satisfaction. To our knowledge, no studies have examined the FAMCARE for equivalence of item endorsement across different socio-demographic groups using methods to detect differential item functioning (DIF). Without such studies, the validity of comparison of means across ethnic and sociodemographically different subgroups could be questioned. One goal of these analyses was to obtain information on DIF to place in an item bank on family satisfaction and care transitions that is under development.

Methods

Sample characteristics

The analytic sample was from a multisite study of patients whose family members were interviewed using the FAMCARE instrument, comprising 20 items. After omission of individuals who responded to less than 50% of items, the analytic sample comprised 1983 patients. Among them, 56.2% of patients were female; the mean age was 59.91 years (standard deviation (SD) = 11.8 years), and 35.1% were 65 years of age or older. The mean educational level was 13.6 years (SD = 3.2 years); 20.4% were non-Hispanic Black, and 79.6% were non-Hispanic White people.

The caregivers were family members living with the patient (43.5%), family members not living with the relative (35.1%), friends (10.5%), home health aides (1.4%), and staff or certified nursing aides (0.1%); 1.6% refused to provide their relationship, and 7.9% were missing. The study was approved by the Institutional Review Board at Mount Sinai Medical Center.

Analyses

This article describes the caregiver respondent DIF analyses with respect to type of caregiver and patient race, age, gender, and education.

Qualitative—One of the initial steps in DIF analyses is the establishment of an a priori set of hypotheses regarding potential group differences in item response by combining information gathered via two methods: (1) qualitatively, from ratings by a panel of content experts, and (2) from a review of the literature documenting prior research-based findings.

Panel of experts

DIF Hypotheses—DIF hypotheses were generated by asking a set of clinicians and other content experts to indicate whether or not they expected DIF to be present, and the direction of the DIF with respect to several comparison groups: gender, age, race/ethnicity, language, and education. A definition of DIF was provided, and 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., male or female. Very specifically, randomly selected persons from each of two groups (e.g., black and white people) who are at the same (e.g., high) 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.

The FAMCARE items were reviewed qualitatively by 12 content experts regarding potential sources of DIF. All the members of the panel of experts were medical doctors, five were geriatricians, one specializes in palliative care, and another was a palliative care geriatrician. The experts were asked to rate individually each of the 20 items with respect to gender, age, race/ethnicity, language, and education. They provided the hypotheses in terms of presence and direction of DIF. The goal was to identify items that might have a different meaning or not be understood well and/or equivalently by individuals of any of the groups referenced. A grid containing a row for each of the 20 items and separate columns for each of the referenced groups was distributed to the experts for completion in order to facilitate the rating.

It was posited that gender, age, race/ethnicity, language, and education were variables that should be investigated because they have been examined in many studies of DIF in other contexts. In hypothesis generation, language was included, even though it was not in the data set, in the event that a data set with a translated version of the items could be obtained for future study. We did not include type of caregiver in the hypotheses generation because this is not a variable that had been examined in previous DIF analyses. However, we decided to include it in the analyses for completeness.

Literature review—A web-based academic library advanced search was conducted on 12 March 2013 via ProQuest (which includes 80 databases) using “FAMCARE” or “Family Satisfaction with the End-of-Life Care scale” and “DIF” or “Differential Item Functioning” or “Factorial Invariance” as key words. No time frame was specified for the search. No article was identified within the parameters specified.

Quantitative analyses and tests of DIF hypotheses

Item Response Theory (IRT)[12–14] applying the graded (polytomous, ordered response category) response model[15] was used for the analyses of DIF. The item characteristic curve (ICC) that relates the probability of an item response to the underlying state, for example, satisfaction, measured by the item set can be characterized by two parameters in some forms of the model: a discrimination parameter (denoted a) that is proportional to the slope of the curve and location (also called severity) parameters (denoted b). According to the IRT model, an item shows DIF if people from different subgroups but at the same level of satisfaction have unequal probabilities of endorsement. Put another way, the absence of DIF is demonstrated by ICCs that are the same for each group of interest.

DIF detection—The method used for DIF detection was the Wald test for examination of group differences in IRT item parameters[13–16] accompanied by magnitude measures.[17] Because there were three education groups, non-orthogonal contrasts were used. The final p values were adjusted using Bonferroni[18] methods; other methods such as Benjamini–Hochberg (B-H) have been used in sensitivity analyses.[19,20] The Bonferroni tests applied here were used to adjust for multiple modeling associated with testing DIF across the entire item set. In this case, the *p* value was adjusted for examination of 20 items (the adjusted *p* value was 0.0025).

The first step in the analyses is to link the two groups compared in terms of satisfaction and to estimate the mean and variance for the target groups studied (while setting the reference group mean to 0 and variance to 1). There are several methods for accomplishing this.[21–23] Typically, anchor items are specified. Anchor items are assumed to be without DIF (no significant differences in the *a* or *b* parameters), and are used to estimate theta (satisfaction), and this process is performed iteratively. The method that was used in these analyses is a modified “all-other” anchor method in which initial DIF estimates can be obtained by treating each item as a “studied” item, while using the remainder as “anchor” items. The procedure described below is performed iteratively in a purification procedure, such that the analyses are repeated using the final subset of items identified as free of DIF as the “purified” anchor set. This procedure is more robust than just relying on the all-other anchor procedure and may take several iterations.

For each studied item, a model is constructed with all parameters constrained to be equal across groups for the anchor items (in this case, all items except the studied item), with the item parameters of the studied item freed to be estimated distinctly for the comparison groups. An overall simultaneous joint test of differences in the *a* or *b* parameters is performed followed by step down tests for group differences in the *a* parameters, followed by conditional tests of the *b* parameters. Uniform DIF is detected when the *b* parameters differ and non-uniform DIF when the *a* parameters differ. Severity (*b*) parameters are interpreted as uniform DIF only if the tests of the *a* parameters are not significant because tests of *b* parameters are performed, constraining the *a* parameters to be equal.

Evaluation of DIF magnitude and impact—The magnitude of DIF refers to the degree of difference in item performance between or among groups, conditional on the trait or state being examined. Expected item scores can be examined as measures of magnitude. An expected item score is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible categories for the item. 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.[24] Cutoff values established based on simulations[25,26] were used in the estimation of the magnitude of item-level DIF. For polytomous items with three response options (after collapsing categories due to sparse data), the recommended cutoff is 0.024.[27]

Expected item scores were summed to produce an expected scale score (also referred to as the test or scale response function), which provides evidence regarding the effect of the DIF on the total score. Group differences in these expected scale score (test response) functions provide overall aggregated measures of impact. The expected scale score functions are shown in Figure 1.

If salient DIF above the magnitude threshold is observed, and the item was hypothesized to have DIF, actions are considered. These include removal, rewording of the item, based on further qualitative cognitive interviews or separate calibrations for the groups in the context of computerized adaptive tests. In analyses, the parameters would be freed to be estimated separately for the groups involved. As discussed, below, given that items did not evidence salient DIF, these considerations were not relevant. However, we do discuss the relationship of the hypotheses to the findings of significant DIF, even if not of high magnitude.

Model assumptions and fit—IRT assumptions include unidimensionality and local independence. The latter implies that the items are independent, conditional on the trait level. Model assumptions and fit were tested. Traditional methods of examining essential unidimensionality were applied,[28] in which a merged exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed fitting a unidimensional model with polychoric correlations using MPlus.[29] The exploratory analyses used principal components estimation and examined tests of scree with cross-loadings permitted. This was followed by the confirmatory analyses of the unidimensional model. The root mean square error of approximation (RMSEA) was examined for model fit; however, we report the comparative fit index (CFI) in the table. Evidence suggests that the CFI may be more robust in the context of invariance testing.[30,31] CFI values > 0.95 generally indicate good model fit;[32,33] however, caution has been recommended in the use of such cutoffs.[34]

The explained common variance (ECV) provides information about whether the observed variance covariance matrix is close to unidimensionality.[35] The ECV can be estimated as the percent of observed variance explained. It is the ratio of the first eigenvalue to the sum of all eigenvalues extracted.[36] There are no firm guidelines for ECV magnitude;[37] however, values greater than 0.50 are desirable. Under the single common factor model, reliability can be evaluated by decomposing the scale score into the sum of the item scores, and the contribution of the common term (λF_j) or communality. Known as McDonald's[38] omega total (ω_t) this reliability estimate is based on the proportion of total common variance explained. As with most reliability estimates, it is desirable to achieve high values (0.80 or better) because unreliability attenuates estimates of relationships with other variables of interest.

We examined the generalized, standardized local dependency (LD) chi-square statistics[39] provided in IRTPRO.[40] Although it is desirable to have values of less than 10, these statistics are affected by sample size. Thus, we examined the smaller samples (the Black and the low education subsamples). We performed sensitivity analyses removing 1 item each from two pairs of items with higher LD values.

Model fit for the DIF models was examined using the RMSEA from IRTPRO. Although there are no set standards, it is generally desirable to achieve values of 0.06 or less.[40]

Software and procedures—The software used was MPlus[29] for factor analyses and IRTPRO Version 2.1[40] for IRT. Additionally, NCDIF[24,26] was evaluated using DFITP5.[27] Prior to application of the DFIT software, the estimates of the latent trait (theta) were calculated separately for each group and equated together with the item parameters. Baker's[41] EQUATE program was used in an iterative fashion in order to equate the theta and item parameter estimates for the two groups and place them on the same metric. If DIF was detected, the item showing DIF was excluded from the equating algorithm, and new DIF-free equating constants were computed and purified iteratively.

Results

Qualitative

The DIF hypotheses are summarized in Table 1. As shown, the majority of raters did not posit gender DIF for most items. Consensus was reached that conditional on satisfaction, women would be more likely to be satisfied than men regarding the patient's pain relief, information about prognosis, and family conferences. Some raters posited that women would be less satisfied with respect to the speed with which symptoms are tested and availability of nurses to the family. Age DIF was posited for 6 items: pain relief, answers from health professionals, speed with which symptoms are treated, availability of nurses, availability of doctors, and coordination of care. Most of the items were posited to be in the direction of younger subjects expressing less satisfaction than older subjects.

Most DIF was posited with respect to race/ethnicity, language, and education. With respect to race/ethnicity, 8 items were posited to evidence DIF, and a direction given: patient's pain relief, information provided about prognosis, referrals to specialists, availability of a hospital bed, family conferences, the way treatments are performed, inclusion of the family in treatment decisions, and information given about tests. White respondents were posited to be more satisfied than minority group members with respect to the above items, except for family conferences and availability of a hospital bed. Most items were posited to show DIF with respect to language; however, the direction was mixed, and our data did not permit examination of DIF by language. Finally, all items were posited to show DIF for education, 17 with a direction provided. Most were in the direction of those with more education being less satisfied. DIF hypothesis related to education (with a direction) were pain relief, information about prognosis, answers from a health professional, information about side effects, referrals to specialists, diagnosis speed, availability of a hospital bed, treatment speed, performance of tests and treatments, availability of doctors and nurses, care received, inclusion of family in decisions, pain management, information about patient's tests, the way tests and treatments are followed up by doctors, and availability of the doctor to the patient.

Quantitative

Our earlier work[7] showed that for the data set analyzed here, only 0.5%–2.3% responded “very dissatisfied,” and for most of the items, 1% or fewer of respondents reported being “very dissatisfied.” Moreover, the results of preliminary IRT analyses[7] using all response categories showed that for all items, the 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, due to sparse data in the very dissatisfied categories, equivocal classification in terms of the “undecided” category, and the results of preliminary IRT analyses, 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. The resulting sum score was from 0 to 40. The analyses were performed using these three collapsed response categories.

As shown in Table 2, there was strong support for essential unidimensionality across all comparison socio-demographic groups. The principal component analyses identified only one subgroup (non-Hispanic Black people) with a second eigenvalue greater than approximately 1. The ratio of component 1 to 2 was large (11.8 to 15.5) for all comparisons, including non-Hispanic Black people (9.7). The first component across comparison groups accounted for between 81% and 85% of the variance for all groups except living arrangement (63% to 66%), supporting the essential unidimensionality of the item set across comparison subgroups. The RMSEA indices (not shown) ranged from 0.05 to 0.09 for all groups except the non-Hispanic Black subsample (RMSEA = 0.11). The CFIs ranged from 0.952 to 0.974 for all groups except the living arrangement variables (0.910, 0.921). The ECVs ranged from 50.383 to 56.469.

In general, the LD statistics were in the acceptable range. However, in sensitivity analyses, we removed 2 items that evidenced higher LD values. Among the Black subsample, item 2 (information about prognosis) evidenced the highest LD values with other items, ranging from 20.6 to 35.5. Among the low education group, the highest LD value was observed for item 12 (availability of nurse to the family). Item 7 also evidenced poor fit ($p < 0.001$) using an additional chi-square diagnostic. The results of the DIF analyses after item removal varied only slightly in terms of the parameter estimates, and the DIF *p* values were very similar, resulting in no change in DIF designations.

The fit statistics (RMSEAs) from IRTPRO for the IRT models (not shown) ranged from 0.04 to 0.05 across DIF subgroup comparisons models, indicating good fit.

The reliability estimates were high. The Omega total values (Table 2) ranged from 0.966 to 0.975, and the Cronbach's alphas (not shown) ranged from 0.951 to 0.959.

The analyses of DIF showed that there was very little DIF evident for patient's age and gender (see Appendix Tables 1 and 2, available online, and Table 3). After Bonferroni adjustment, non-uniform DIF was observed by age for 1 item, “Information given about patients’ tests.” However, the magnitude of DIF was small, and the NCDIF statistic was not significant or large. The impact of DIF was negligible, as shown by the overlapping curves (see Figure 1). For gender, no items showed DIF after Bonferroni adjustment.

For race, 13 items showed DIF after Bonferroni adjustment, most with non-uniform DIF (see Appendix Table 3, available online, and Table 3.) The items with uniform DIF were “The patient's pain relief,” “Doctors attention to patient's description of symptoms,” and “Availability of nurses to the family.” Conditional on satisfaction, these items were more likely to be endorsed in the satisfied direction by White than by Black people. The discrimination parameter estimates tended to be higher for the Black than for the White group for the 10 items with non-uniform DIF. No items evidenced DIF of high magnitude, and the impact was trivial (see Figure 1).

For education, 5 items evidenced DIF after Bonferroni adjustment, all uniform. Conditional on level of satisfaction, in contrast to caregivers of patients with lower education, caregivers of patients with higher education were likely to report less satisfaction with pain relief, coordination of care, and the way treatments are performed and more satisfaction with specialist referrals and availability of a hospital bed (see Appendix Table 4, available online, and Table 4). No items evidenced DIF of high magnitude or impact (see Figure 1).

Only 1 item evidenced DIF for the relationship variable: Family respondents living with the care recipients as contrasted with family members not living with the care recipients were more likely to be dissatisfied with the availability of a nurse, conditional on level of satisfaction (see Appendix Table 5, available online, and Table 4).

Discussion

Examination of the hypotheses for the qualitative analyses in conjunction with the quantitative analyses showed that most items were not hypothesized to show DIF for gender (5 items) or age (6 items) and little or no DIF was observed. For race, many items were posited to evidence DIF. In general, minority groups were hypothesized to express less satisfaction than White groups, conditional on overall satisfaction. For the 3 items with uniform DIF, a directional hypothesis was given for 1 item, and it was confirmatory. It was posited that conditional on satisfaction level, caregivers of Black patients would be less satisfied with pain relief, and this was the direction of the DIF. Two other items with uniform DIF were hypothesized to show DIF, but the direction was not specified ("doctor's attention to patient's description of symptoms" and "availability of nurses"). It is noted that while most items evidenced non-uniform DIF, the severity (b) parameters were also significantly different, all in the direction of lower conditional item satisfaction scores among Black as contrasted with White respondents. Except for 2 items, the hypotheses were confirmatory in that most items were hypothesized to show DIF in the direction of less satisfaction for Black people.

For education, 4 out of 5 items evidenced DIF in the direction hypothesized. The uniform DIF observed for "pain relief," "coordination of care," and "the way treatments are performed" were in the expected direction with those with higher education posited to be less satisfied, conditional on level of satisfaction. The DIF observed for "referrals to specialists" was also in the direction posited, with higher satisfaction expected for those with higher education. The finding related to hospital bed was in a direction opposite than hypothesized.

Limitations and strengths

Except for relationship of respondent to the patient and most likely race, the variables examined were with respect to the patient for whom the proxy was reporting. The findings for age and gender do not generalize to the population of caregivers; however, given the use of the measure as a proxy for patient response, the findings for gender and age may generalize to proxy reports of patient satisfaction. Although the lack of information about caregiver gender, age, and education is a limitation, to our knowledge, this is the first and only study of DIF in the FAMCARE using a relatively large, ethnically diverse sample.

Another limitation is that the results were mixed with respect to confirmatory evidence for the hypotheses. For example, while items with DIF were generally hypothesized to show DIF, 4 items observed to have non-uniform DIF for race were not hypothesized to show DIF. Additionally, 4 items posited to show DIF in the direction of minorities expressing less satisfaction, conditional on satisfaction, were not found to have DIF. However, most items with differences in the severity (location) parameters were in the posited direction.

Finally, while the fit statistics for the confirmatory factor model were generally acceptable, the fit was slightly poorer among the non-Hispanic Black subsample, which could result in over-identification of DIF in analyses involving that group.

Conclusion

In conclusion, within the limitations of the study, DIF was observed primarily for race and education. No DIF of high magnitude was observed for any item, and the total impact of DIF at the scale level was negligible. One item, satisfaction with “the patient's pain relief” might be singled out for further study, given that this item was both hypothesized and observed to show DIF for race and education. Racial and ethnic disparities on the overall experience of pain (i.e. perception, assessment management, and treatment) have been documented.[42] Black people have also been found to have less adequate pain care at referral, prior to specialty pain care as compared to non-Hispanic Whites.[43] The literature documents individual (patient and provider) and systemic (health care) factors as explanatory mechanisms for such disparities.[44] Given that Black and White people have been shown to differ in response to pain items[45,46] and that there are health disparities regarding pain recognition and treatment, this is an area that requires careful assessment.

Family satisfaction is frequently measured in patients receiving palliative care. As most societies experience greater ethnic and socio-demographic diversity, it is important to provide evidence regarding the performance of these measures among such groups of people. Although socioeconomic factors can impact access to quality palliative and end-of-life care for patients and caregivers, it was important to examine to what extent reported dissatisfaction with palliative care by minority groups might be due to a measurement artifact. Because palliative care populations are characterized by individuals who may have serious illness and may be frail, patients and caregivers may experience burden answering questions. One goal of the project was to develop an item bank of parameters that can be used to select items for shorter forms or targeted and tailored assessment. Additionally, a goal might be to use a computerized adaptive test in the future. Such efforts require well-calibrated parameters and evidence of DIF in the item bank. In that fashion, some items may be avoided for administration or deemed satisfactory for inclusion. It can be concluded that the DIF observed in the FAMCARE for this study was of low magnitude and impact. Based on these data, most items can be recommended for further use, with the caveat that more DIF testing may be needed to examine DIF with respect to selected caregiver characteristics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This project was funded by the United States National Cancer Institute (grant no. 5R01CA116227-05). Support for these analyses was provided by the Claude Pepper Older Americans Independence Center: United States National Institute on Aging, P30, AG028741.

References

1. Kristjanson LJ. Indicators of quality of palliative care from a family perspective. J Palliat Care. 1986; 1(2):8–17. [PubMed: 2453638]
2. Kristjanson LJ. Quality of terminal care: salient indicators identified by families. J Palliat Care. 1989; 5(1):21–30. [PubMed: 2715882]
3. Hwang SS, Chang VT, Alejandro Y, et al. Caregiver unmet needs, burden, and satisfaction in symptomatic advanced cancer patients at a Veterans Affairs (VA) medical center. Palliat Support Care. 2003; 1:319–329. [PubMed: 16594221]
4. Ringdal GI, Jordhoy MS, Kaasa S. Measuring quality of palliative care: psychometric properties of the FAMCARE Scale. Qual Life Res. 2003; 12(2):167–176. [PubMed: 12639063]
5. Carter GL, Lewin TJ, Gianacas L, et al. Caregiver satisfaction with out-patient oncology services: utility of the FAMCARE instrument and development of the FAMCARE-6. Support Care Cancer. 2011; 19(4):565–572. [PubMed: 20349317]
6. Rodriguez KL, Bayliss NK, Jaffe E, et al. Factor analysis and internal consistency evaluation of the FAMCARE scale for use in the long-term care setting. Palliat Support Care. 2010; 8(2):169–176. [PubMed: 20331914]
7. Teresi JA, Ornstein K, Ramirez M, et al. Performance of the Family Satisfaction with the End-of-Life Care (FAMCARE) measure in an ethnically diverse cohort: psychometric analyses using item response theory. Support Care Cancer. Epub ahead of print 5 October 2013. DOI:10.1007/s00520-013-1988-z.
8. Aoun S, Bird S, Kristjanson LJ, et al. Reliability testing of the FAMCARE-2 scale: measuring family care satisfaction with palliative care. Palliat Med. 2010; 24(7):674–681. [PubMed: 20621947]
9. Lo C, Burman D, Rodin G, et al. Measuring patient satisfaction in oncology palliative care: psychometric properties of the FAMCARE-patient scale. Qual Life Res. 2009; 18:747–752. [PubMed: 19513815]
10. Kristjanson LJ. Validity and reliability testing of the FAMCARE Scale: measuring family satisfaction with advanced cancer care. Soc Sci Med. 1993; 36(5):693–701. [PubMed: 8456339]
11. Johnsen AT, Ross L, Petersen MA, et al. The relatives' perspective on advanced cancer care in Denmark. A cross-sectional survey. Support Care Cancer. 2012; 12:3179–3188. [PubMed: 22526148]
12. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. SAGE; Newbury Park, CA: 1991.
13. Lord, FM. Applications of item response theory to practical testing problems. Lawrence Erlbaum; Hillsdale, NJ: 1980.
14. Lord, FM.; Novick, MR. Statistical theories of mental test scores. Addison-Wesley Publishing Company; Reading, MA: 1968.
15. Samejima, F. Estimation of latent ability using a response pattern of graded scores (Psychometrika Monograph; Supplement 17), 1969. Springer; Dordrecht:
16. Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models.. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Lawrence Erlbaum, Inc.; Hillsdale, NJ: 1993. p. 123-135.
17. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med. 2000; 19:1651–1683. [PubMed: 10844726]
18. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936; 8:3–62.
19. Benjamini Y, Hochberg Y. Controlling for the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B Met. 1995; 57:289–300.
20. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini Hochberg procedure for controlling the false discovery rate in multiple comparisons. J Educ Behav Stat. 2002; 27:77–83.

21. Orlando-Edelen M, Thissen D, Teresi JA, et al. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: applications to the Mini-Mental State Examination. Med Care. 2006; 44:S134–S142. [PubMed: 17060820]

22. Woods CM. Empirical selection of anchors for tests of differential item functioning. Appl Psych Meas. 2009; 33:42–57.

23. Wang W- C, Shih C-L, Sun G-W. The DIF-free-then-DIF strategy for the assessment of differential item functioning. Educ Psychol Meas. 2012; 72:687–708.

24. Raju NS, Van Der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Appl Psych Meas. 1995; 19:353–368.

25.

Fleer PF. A Monte Carlo assessment of a new measure of item and test bias. 1993Illinois Institute of TechnologyChicago, IL (Dissertation, Dissertation Abstracts International, 54–04B, 2266).

26. Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Appl Psych Meas. 1999; 23:309–332.

27. Raju, NS. DFITP5: a Fortran program for calculating dichotomous DIF/DTF (computer program). Illinois Institute of Technology; Chicago, IL: 1999.

28. Asparouhov T, Muthén B. Exploratory structural equation modeling. Struct Equ Modeling. 2009; 16:397–438.

29. Muthén, LK.; Muthén, BO. M-PLUS users guide (1998– 2011). 6th ed.. Muthén & Muthén; Los Angeles, CA: 2011.

30. Meade AW, Johnson EC, Bradley PW. Power and sensitivity of alternative fit indices in tests of measurement invariance. J Appl Psychol. 2008; 93:568–592. [PubMed: 18457487]

31. Cheung GW, Rensvold RB. Evaluating goodness-offit indexes for testing measurement invariance. Struct Equ Modeling. 2003; 9:233–255.

32. Muthén, BO. Latent variable modeling in heterogeneous populations.. Psychometrika; Meetings of Psychometric Society 1989 Los Angeles, California and Leuven; Belgium. 1989. p. 557-585.

33. Bentler PM. Comparative fit indexes in structural models. Psychol Bull. 1990; 107(2):238–246. [PubMed: 2320703]

34. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res. 2009; 18:447–460. [PubMed: 19294529]

35. Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 2009; 74:107–120. [PubMed: 20037639]

36. Reise SP, Moore TM, Haviland MG. Bi-factor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. J Pers Assess. 2010; 92:544–559. [PubMed: 20954056]

37. Reise SP. The rediscovery of bifactor measurement models. Multivariate Behav Res. 2012; 47:667–696. [PubMed: 24049214]

38. McDonald, RP. Test theory: a unified treatment. Lawrence Erlbaum Associates; Mahwah, NJ: 1999.

39. Chen WH, Thissen D. Local dependence indices for item pairs using item response theory. J Educ Behav Stat. 1997; 22:265–289.

40. Cai, L.; Thissen, D.; du Toit, SHC. IRTPRO: flexible, multidimensional, multiple categorical IRT Modeling (Computer software). Scientific Software International, Inc; Chicago, IL: 2011.

41. Baker, FB. EQUATE 2.1: computer program for equating two metrics in item response theory (Computer program). Laboratory of Experimental Design, University of Wisconsin; Madison, WI: 1995.

42. Shavers VL, Bakos A, Sheppar VB. Race, ethnicity, and pain among the U.S. adult population. J Health Care Poor Underserved. 2010; 21(1):177–220. [PubMed: 20173263]

43. Green CR, Hart-Johnson T. The adequacy of chronic pain management prior to presenting at a tertiary care pain center: the role of patient socio-demographic characteristics. J Pain. 2010; 11(8): 746–754. [PubMed: 20399710]

44. Green CR, Anderson KO, Baker TA, et al. The unequal burden of pain: confronting racial and ethnic disparities in pain. Pain Med. 2003; 4(3):277–294. [PubMed: 12974827]

45. Cykert S, Joines JD, Kissling G, et al. Racial differences in patients' perceptions of debilitated health states. J Gen Intern Med. 1999; 14:217–222. [PubMed: 10203633]

46. Rahim-Williams FB, Riley JL, Herrera D, et al. Ethnic identity predicts experimental pain sensitivity in African Americans and Hispanics. Pain. 2007; 129:177–184. [PubMed: 17296267]

What is already known about the topic?

- The Family Satisfaction with End-of-Life Care (FAMCARE) is a widely used measure of satisfaction with cancer care.
- The psychometric properties of the scale have been examined with cancer patients in diverse samples and settings, internationally; adequate estimates of internal consistency and other reliability statistics were observed in these studies.
- However, the FAMCARE has not been studied for equivalence of item endorsement across different socio-demographic groups using differential item functioning (DIF).

What this paper adds?

- This study is the first to examine the FAMCARE for equivalence of item endorsement across different socio-demographic groups.
- Examination of DIF using item response theory is important in finalizing item banks and developing short-form measures. These analyses provide information about DIF to place in an item bank on family satisfaction and care transitions that is under development.

Implications for practice, theory, or policy

- DIF was observed primarily for race and education.
- No DIF of high magnitude was observed for any item, and the total impact of DIF at the scale level was trivial.
- It is recommended that the item, satisfaction with “the patient's pain relief,” be studied further, given that racial and ethnic disparities on the overall experience of pain have been documented, and that this item was both hypothesized and observed to show DIF for race and education. Clinicians should be alert to potential response bias in reports of satisfaction with pain relief.

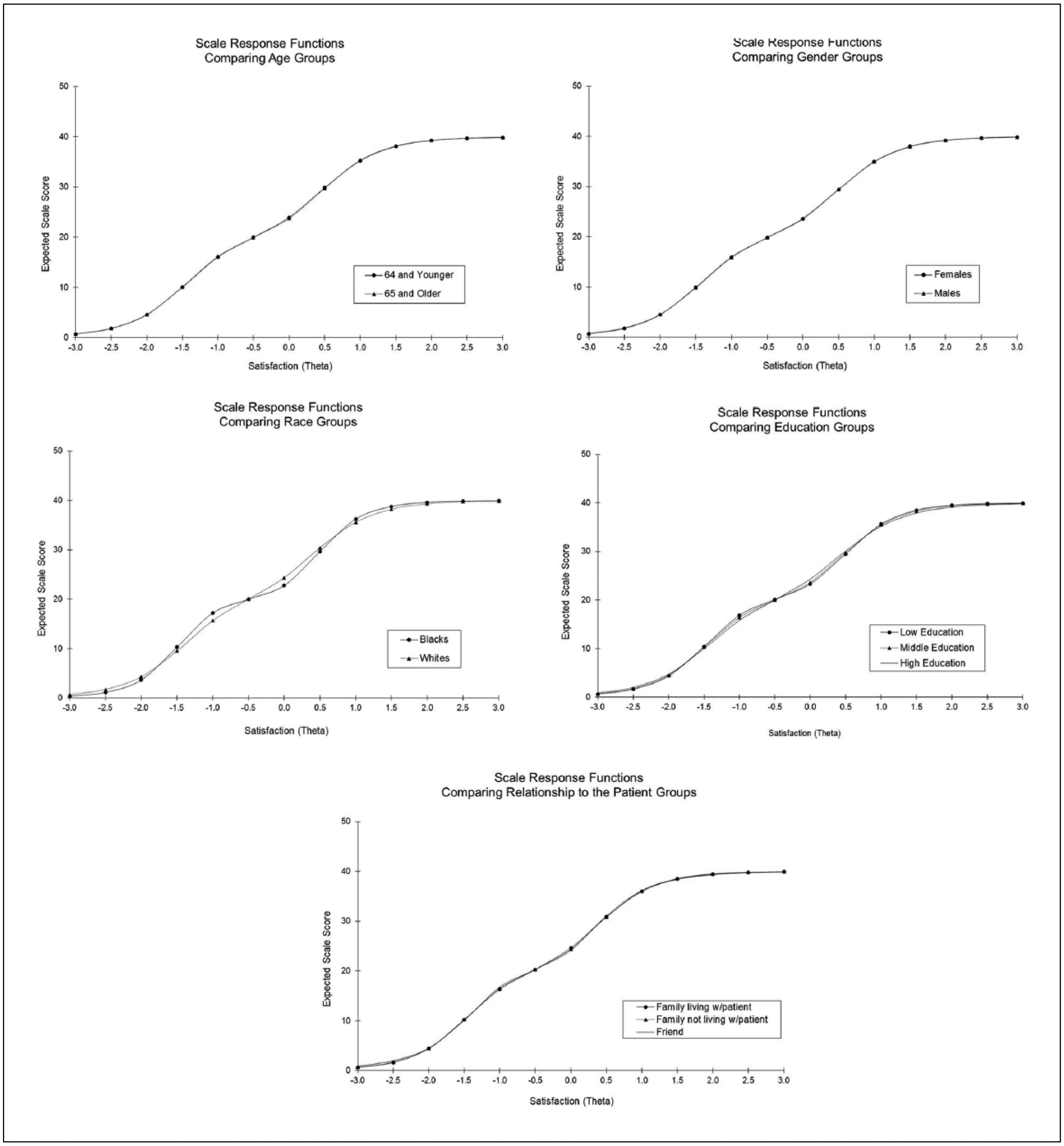

Figure 1.

FAMCARE item set: scale response functions by comparison groups.

FAMCARE: Family Satisfaction with End-of-Life Care.

Table 1

Summary of DIF hypotheses generated by 12 content experts.

No.	Item stem	Content expert hypotheses				
		Gender	Age	Race/ethnicity	Language	Education
1	The patient's pain relief	**7** (Women more satisfied)	**7** (Older more satisfied)	**5** (Spanish/Hispanic group more satisfied)/(White group more satisfied)	**5** (Spanish/Hispanic generally more satisfied)/(Non-English less satisfied)	**4** (More educated less satisfied)
2	Information provided about the patient's prognosis	**4** (Women more satisfied)		**5** (Minorities less satisfied)	**8** (Non-English; language barrier less satisfied)	**6** (More educated expect more details)/(Education higher)[a]
3	Answers from health professionals		**6** (Younger less satisfied—require more details)	**4** (Asians ask more questions)	**9** (Language barrier less satisfied)/(non-English higher)	**7** (Higher education less satisfied)/(Education higher)
4	Information given about side effects				**6** (Language barrier)/(Non-English less satisfied)	**5** (Higher education less satisfied)
5	Referrals to specialists			**5** (White people higher/ Black people less)		**5** (More education higher satisfaction)
6	Availability of a hospital bed			**2** (White people less satisfied)		**4** (Higher education less satisfied)
7	Family conferences held to discuss the patient's illness	**6** (Women more satisfied)	**4**	**6** (White people less satisfied)	**5** (Language barrier less satisfied)/(non-English higher)	**3**
8	Speed with which symptoms are treated	**5** (Female less satisfied)	**6** (Younger less satisfied)	**2**		**2** (Higher education less satisfied)
9	Doctor's attention to patient's description of symptoms			**3**	**4**	**3**
10	The way test and treatments are performed			**5** (Hispanic groups more trusting)/(Blacks less satisfied)	**5** (Language barrier less satisfied)	**6** (Higher education higher satisfaction)
11	Availability of doctors to the family				**4** (Non-English higher)	**4** (Higher education less satisfied)
12	Availability of nurses to the family	**4** (Female less satisfied)	**4** (Younger less satisfied)	**2**	**4** (Non-English higher)	**4** (Higher education less satisfied)
13	Coordination of care		**6** Older less satisfied)	**4**	**5** (Language barrier less satisfied)	**3** (Higher education less satisfied)
14	Time required to make a diagnosis					**3** (Higher education less satisfied)
15	The way the family is included in treatment and care decisions			**7** (Minority groups less satisfied)	**8** (Language barrier/Non-English less satisfied)/(Non-English higher)	**5** (Higher education more expectations)

No.	Item stem	Content expert hypotheses				
		Gender	Age	Race/ethnicity	Language	Education
16	Information given about how to manage the patient's pain					**3** (Higher education higher satisfaction)
17	Information given about the patient's tests			**4** (Caucasian groups higher satisfaction)	**4** (Non-English less satisfied)	**4** (Higher education higher satisfaction)
18	How thoroughly the doctor assesses the patient's symptoms			**4**	**6** (Non-English higher)	**4**
19	The way tests and treatments are followed up by the doctors				4	**6** (Higher education less satisfied)
20	Availability of the doctor to the patient		**4** (Younger expect more availability)		5 (Non-English higher)	**4** (Higher education less satisfied)

DIF: differential item functioning.

The numbers in bold are the number positing DIF; not all provided a direction to the hypothesis.

[a] Higher is indicative of more agreement or higher satisfaction.

Table 2

Eigenvalues from the exploratory factor analysis using principal component estimation and the model fit statistics for the total sample and subsamples.

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio component 1/component 2	CFI	ECV	Omega total (ω_t)
Total sample (*n* = 1983)								
Eigenvalues	12.723	0.915	0.824	0.611	13.9	0.965	53.125	0.970
Explained variance	84.4%	6.1%	5.5%	4.1%				
Random first half of the total sample (*n* = 991)								
Eigenvalues	12.775	0.929	0.779	0.603	13.8	0.964	N/A	N/A
Explained variance	84.7%	6.2%	5.2%	4.0%				
Age 64 years and under (*n* = 1274)								
Eigenvalues	12.512	0.972	0.821	0.629	12.9	0.963	52.308	0.969
Explained variance	83.8%	6.5%	5.5%	4.2%				
Age 65 years and over (*n* = 696)								
Eigenvalues	13.142	0.875	0.852	0.656	15.0	0.973	55.005	0.972
Explained variance	84.7%	5.6%	5.5%	4.2%				
Females (*n* = 1115)								
Eigenvalues	12.852	0.919	0.816	0.647	14.0	0.964	50.383	0.968
Explained variance	84.4%	6.0%	5.4%	4.2%				
Males (*n* = 865)								
Eigenvalues	12.589	0.955	0.89	0.636	13.2	0.967	53.269	0.970
Explained variance	83.5%	6.3%	5.9%	4.2%				
Non-Hispanic Black (*n* = 388)								
Eigenvalues	12.424	1.279	0.891	0.738	9.7	0.952	53.739	0.971
Explained variance	81.0%	8.3%	5.8%	4.8%				
Non-Hispanic White (*n* = 1517)								
Eigenvalues	12.778	0.887	0.847	0.609	14.4	0.968	52.426	0.969
Explained variance	84.5%	5.9%	5.6%	4.0%				
Less than high school (*n* = 317)								
Eigenvalues	13.106	1.011	0.817	0.701	13.0	0.974	53.874	0.972
Explained variance	83.8%	6.5%	5.2%	4.5%				
High school (*n* = 666)								

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio component 1/component 2	CFI	ECV	Omega total (ω_t)
Eigenvalues	13.53	0.985	0.741	0.583	13.7	0.968	56.469	0.975
Explained variance	85.4%	6.2%	4.7%	3.7%				
Some college and above (*n* = 992)								
Eigenvalues	12.242	0.973	0.887	0.668	12.6	0.964	50.903	0.966
Explained variance	82.9%	6.6%	6.0%	4.5%				
Relative living with patient (*n* = 862)								
Eigenvalues	13.102	0.843	0.773	0.650	15.5	0.921	54.635	0.972
Explained variance	65.52%	4.22%	3.87%	3.25%				
Relative not living with patient (*n* = 696)								
Eigenvalues	12.591	1.067	0.78	0.668	11.8	0.910	52.477	0.969
Explained variance	62.96%	5.34%	3.90%	3.34%				

CFI: comparative fit index; ECV: explained common variance.

Table 3

Summary of DIF analyses: Age, gender, and race groups.

Item		Anchor item			Type of DIF, if present			DIF after Bonferroni adjustment		
		Age	Sex	Race	Age	Sex	Race	Age	Sex	Race
1	The patient's pain relief	√	√				NU, U			U
2	Information provided about prognosis	√	√				NU, U			
3	Answers from health professionals	√	√				NU, U			
4	Information given about side effects	√	√				NU			NU
5	Referrals to specialists	√	√	√						
6	Availability of hospital bed	√	√				NU, U			NU, U
7	Family conferences held to discuss the patient's illness	√		√		U				
8	Speed with which symptoms were treated	√	√				NU, U			NU, U
9	Doctor's attention to patient's description of symptoms	√	√				NU, U			U
10	The way tests and treatments are performed		√		NU		NU, U			NU, U
11	Availability of doctors to the family	√	√	√						
12	Availability of nurses to the family	√	√				NU, U			U
13	Coordination of care	√	√				NU, U			NU
14	Time required to make diagnosis	√	√				NU, U			NU
15	The way the family is included in treatment and care decisions		√		NU		NU, U			
16	Information given about how to manage the patient's pain		√		NU		NU			NU
17	Information given about the patient's tests		√		NU		NU	NU		
18	How thoroughly the doctor assesses the patient's symptoms	√				NU	NU, U			NU, U
19	The way tests and treatments are followed up by the doctor	√	√				NU, U			NU, U
20	Availability of the doctor to the patient	√	√				NU, U			NU, U

DIF: differential item functioning; NCDIF: non-compensatory DIF; NU: non-uniform DIF involving the discrimination parameters; U: uniform DIF involving the location parameters.

All NCDIF values were smaller than the threshold (0.0240); the range was from 0.0001 to 0.0018 for the age groups, from 0.0001 to 0.0044 for the gender groups, and from 0.0015 to 0.0105 for the race groups.

Table 4

Summary of DIF analyses: education and relationship to the patient.

Item		Anchor item				Type of DIF, if present				DIF after Bonferroni adjustment			
		Education		Caregiver relationship		Education		Caregiver relationship		Education		Caregiver relationship	
		Low and high	Middle and high	Living with/not[a]	Living with[b]/ friend	Low and high	Middle and high	Living with/not	Living with/ friend	Low and high	Middle and high	Living with/not	Living with/ friend
1	The patient's pain relief				√	U	U				U		
2	Information about prognosis			√	√	NU, U	NU, U						
3	Answers from health professionals				√	NU, U	U	NU					
4	Information about side effects			√	√	NU, U							
5	Referrals to specialists			√		U	U		NU		U		
6	Availability of hospital bed			√		NU, U	NU, U			U	U		
7	Family conferences about illness	√	√	√					U				
8	Speed symptoms were treated		√	√	√	NU, U							
9	Doctor's attention to patient's description of symptoms		√	√	√								
10	The way tests and treatments are performed			√		U	NU, U		U		U		
11	Availability of doctors to the family	√	√	√	√								
12	Availability of nurses to the family	√			√		U	NU, U				U	
13	Coordination of care			√	√	U	U			U	U		
14	Time required to make diagnosis			√		U	NU		NU				
15	The way the family is included in treatment and care decisions	√	√	√	√								
16	Information given about how to manage the patient's pain	√		√	√								
17	Information about tests		√	√		NU			NU				
18	How thoroughly the doctor assesses the patient's symptoms				√		U						
19	The way tests and treatments are followed up by the doctor		√	√		NU							
20	Availability of the doctor					NU	U	NU					

DIF: differential item functioning; NCDIF: non-compensatory DIF.

All NCDIF values were smaller than the threshold (0.024). The range was from 0.0004 to 0.0087 for the low versus high education groups, from 0.0004 to 0.0037 for the middle versus high education groups, from 0.0001 to 0.0129 for the relatives living with the patient versus relative not living with the patient groups, and from 0.0005 to 0.0107 for the comparison of friends versus family living with the patients.

[a] Relative living with the patient versus relative not living with the patient.

[b] Relative living with the patient versus friend.