Qual Life Res (2007) 16:43–68
DOI 10.1007/s11136-007-9186-4

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

Jeanne A. Teresi · Katja Ocepek-Welikson · Marjorie Kleinman · Karon F. Cook · Paul K. Crane · Laura E. Gibbons · Leo S. Morales · Maria Orlando-Edelen · David Cella

Received: 29 August 2006 / Accepted: 29 January 2007 / Published online: 5 May 2007

Abstract *Background* Methods based on item response theory (IRT) that can be used to examine differential item functioning (DIF) are illustrated. An IRT-based approach to the detection of DIF was applied to physical function and general distress item sets. DIF was examined with respect to gender, age and race. The method used for DIF detection was the item response theory log-likelihood ratio (IRTLR) approach. DIF magnitude was measured using the differences in the expected item scores, expressed as the unsigned probability differences, and calculated using the non-compensatory DIF index (NCDIF). Finally, impact was assessed using expected scale scores, expressed as group differences in the total test (measure) response functions. *Methods* The example for the illustration of the methods came from a study of 1,714 patients with cancer or HIV/AIDS. The measure contained 23 items measuring physical functioning ability and 15 items addressing general distress, scored in the positive direction. *Results* The substantive findings were of relatively small magnitude DIF. In total, six items showed relatively larger magnitude (expected item score differences greater than the cutoff) of DIF with respect to physical function across the three comparisons: ''trouble with a long walk'' (race), ''vigorous activities'' (race, age), ''bending, kneeling stooping'' (age), ''lifting or carrying groceries'' (race), ''limited in hobbies, leisure'' (age), ''lack of energy'' (race). None of the general distress items evidenced high magnitude DIF; although ''worrying about dying'' showed some DIF with respect to both age and race, after adjustment. *Conclusions* The fact that many physical function items showed DIF with respect to age, even after adjustment for multiple comparisons, indicates that the instrument may be performing differently for these groups. While the magnitude and impact of DIF at the item and scale level was minimal, caution should be exercised in the use of subsets of these items, as might occur with selection for clinical decisions or computerized adaptive testing. The issues of selection of anchor items, and of criteria for DIF detection, including the integration of significance and magnitude measures remain as issues requiring investigation. Further research is needed regarding the criteria and guidelines appropriate for DIF detection in the context of health-related items.

J. A. Teresi (✉) · M. Kleinman
Faculty of Medicine, New York State Psychiatric Institute, Columbia University Stroud Center, New York, NY, USA
e-mail: Teresimeas@aol.com

J. A. Teresi · K. Ocepek-Welikson
Research Division, Hebrew Home for the Aged at Riverdale, Riverdale, NY, USA

K. F. Cook
Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA

P. K. Crane · L. E. Gibbons
Department of Internal Medicine, University of Washington, Seattle, WA, USA

L. S. Morales · M. Orlando-Edelen
RAND Corporation, Santa Monica, CA, USA

L. S. Morales
UCLA Department of Health Services and Divison of General Internal Medicine and Health Services Research, Los Angeles, CA, USA

M. Orlando-Edelen
Brown Medical School, Providence, RI, USA

D. Cella
Center on Outcomes, Research and Education, Evanstan Northwestern Healthcare, Evanston, IL, USA

D. Cella
Northwestern University, Evanston, IL, USA

Keywords Differential item functioning · Item response theory · Physical functioning · General distress

Introduction

The purpose of this paper is to illustrate methods based on item response theory (IRT) that can be used to examine differential item functioning (DIF). The companion paper by Crane and colleagues [1] illustrates the use of the ordinal logistic regression (OLR) approaches of Swaminathan and Rogers [2], Zumbo [3] and Crane et al. [4]; some of the analyses in that paper were based on a modified IRT approach. There are several other methods for DIF detection, which are reviewed in a recent special issue of Medical Care [5]. The advantages and disadvantages of different DIF detection methods are also reviewed in that issue [6]. While a discussion of these issues is beyond the scope of this paper; several simulation studies reviewed in that special issue support the use of the IRTLR approach to DIF detection.

This paper applies an IRT-based approach to the detection of DIF in physical function and general distress item sets. DIF was examined with respect to gender, age and race. These demographic variables were selected on theoretical grounds. DIF should be performed with respect to variables that are hypothesized to affect the relationship between the item response and the ability (disability) targeted for study. Previous studies have identified DIF in measures of affective disorder [7–9] and physical function [10–12] with respect to one or more of the three background variables examined.

The method used for DIF detection that is described in this paper was the IRT log-likelihood ratio (IRTLR) approach [13, 14]. DIF magnitude was assessed using the differences in expected item scores, expressed as the unsigned probability differences [15], and calculated using the non-compensatory DIF (NCDIF) index [16, 17]. Finally, impact was assessed using expected scale scores, expressed as group differences in the total test (measure) response functions. These latter functions show the extent to which DIF cancels at the scale level (DIF cancellation). The measures, sample and background are described in the paper by Crane and colleagues [1], and this information is briefly summarized in the Methods section.

Definition of DIF: DIF analysis in the context of health-related constructs involves three factors: item response, disability (ability) level and subgroup membership; the research question is how item response is related to disability for different subgroups. The relationship implied by this question is often defined in terms of item parameters so that DIF analysis frequently examines differences in these parameters. DIF analysis is concerned with the question of whether or not the likelihood of item (category) endorsement is equal across subgroups. A key issue is whether the method used is conditional or non-conditional; only conditional methods that take disability/ability into account are acceptable. The necessity of a conditioning variable has been illustrated by Dorans and Holland [18] and Dorans and Kulick [19], in the context of Simpson's [20] paradox. They provide examples showing that if two groups vary in the distribution of ability, overall, an item will appear to favor the group with more functional ability. However, examination of differences in proportions endorsing an item (claiming independence in function) at different ability levels or groupings can actually show a reverse pattern. Thus, as pointed out by Dorans and Kulick [19], it is important to compare the comparable, by controlling for disability/ability before examining differences in performance between groups on an item.

Another key issue is the nature of the conditioning variable. IRT disability/ability estimates are often used because observed scores (typically used in logistic regression) may not be adequate proxies for latent health status, and may result in false DIF detection, particularly with shorter scales (see Millsap and Everson [21]). However, logistic regression methods do not need to be limited to the use of observed scores. Latent conditioning variables can be used, as was done in the companion paper [1].

DIF in the context of IRT: A basic concept in IRT is that a set of items is being used to measure an underlying attribute (also called a trait or state, e.g., a health condition, state of emotional distress, functional ability, disability or disorder); the central concern is how the item responses are related to the trait. For the example presented in this paper, the underlying attributes are scored in the positive direction, and reflect physical functional ability and positive emotional state (lack of general distress).

Different models are used to model binary (dichotomous) items, as contrasted with ordered categorical (polytomous) items. A mixture of such items was used in the scales analyzed. The following discussion pertains to binary items; an explication of the model for polytomous items is discussed in the Appendix. The expectation is that respondents who are not disabled would be more likely than those who are disabled to respond asymptomatically (in a non-symptomatic direction) to an item measuring ability. Conversely, a person with disability is expected to have a lower probability of responding in a non-disabled direction to the item. The curve that relates the probability of an item response to the underlying health condition measured by the item set is known as an item characteristic curve (ICC). This curve can be characterized by two parameters in some forms of the model: a discrimination parameter (denoted as a) that is proportional to the slope of the curve, and a location (also called difficulty, or severity)

parameter (denoted as *b*) that is the point of inflection of the curve. (See also the Appendix.) According to the IRT model illustrated by the Figures contained within this paper, an item shows DIF if people from different subgroups but at the same functional ability level have unequal probabilities of endorsement. For example, in the absence of DIF, African-American people with mild disability should have the same chance of a given response to a particular physical functioning ability item as do white people with mild disability. Put another way, the absence of DIF is demonstrated by ICCs that are the same for each group of interest.

Description of the Model

The following analyses were conducted using the two parameter mixed logistic (for binary) and graded (for polytomous, ordered response category) item response models (see Hambleton et al. [22]; Lord [23]; Lord and Novick [24]; Samejima [25]). Important first steps (not presented here) in the analyses include examination of model assumptions (such as unidimensionality) and model fit. These analyses were conducted prior to release of these data sets, and provided evidence of essential unidimensionality.

Example of the Model: An example is shown in Fig. 1. The curves for two self-identified race groups for the item, ''trouble with a long walk'', represent the relationship between the probability of a positive (unimpaired) response and physical functioning ability. The fact that the curves are not identical and that there is space between the curves (area) indicates that some DIF is present. In this example, the curves are parallel and do not cross. This shows what is called ''Uniform DIF'' or ''Unidirectional DIF''. As an example of the meaning, locate the point on the solid curve corresponding to .5 along the *x* (ability) axis, also referred to as theta (θ), and draw a line to the *y* axis (probability of response). The intersection of these ICCs with a vertical line provides the probability of item endorsement for individuals, given selected levels of ability. For example, the probability of a randomly selected African-American person of above average physical function ($\theta = .5$) responding that s/he has no trouble ''with a long walk'' is higher (.62) than for a randomly selected White person (.44) at the same ability level. Specifically, at this ability level (*theta* = .5), the DIF results in a difference in response probabilities of .18. In fact, across much of the ability continuum, African-American respondents are more likely than White respondents of the same ability level to endorse the category, ''no trouble'', resulting in a difference in the areas under the curves for the two groups. It takes more ability for Whites than for African-Americans to claim that they have no trouble with a long walk. For example, a probability of .62 for Whites corresponds to a higher ability level (theta closer to 1.0) than for African-Americans. Thus, this item is not performing in the same manner for both groups, and model-based significance tests indicated that this item exhibited DIF: it maximally discriminates (separates ability levels) at higher levels of functional ability for Whites as contrasted with African-Americans. This also is demonstrated by the higher *b* (or severity) parameter estimate for White respondents (.61) than for African-American respondents (.35).

This difference is also apparent in the raw data, and can be illustrated by examining the crosstabulation between item response and race classification for a selected observed score level. For example, moving from the latent variable model just discussed to the more familiar sum score, it is observed that for raw sum score levels 26–32 on the Physical Function scale (reflecting above average physical function), 31.6% of African-American persons, as contrasted with 17.2% of White persons responded that

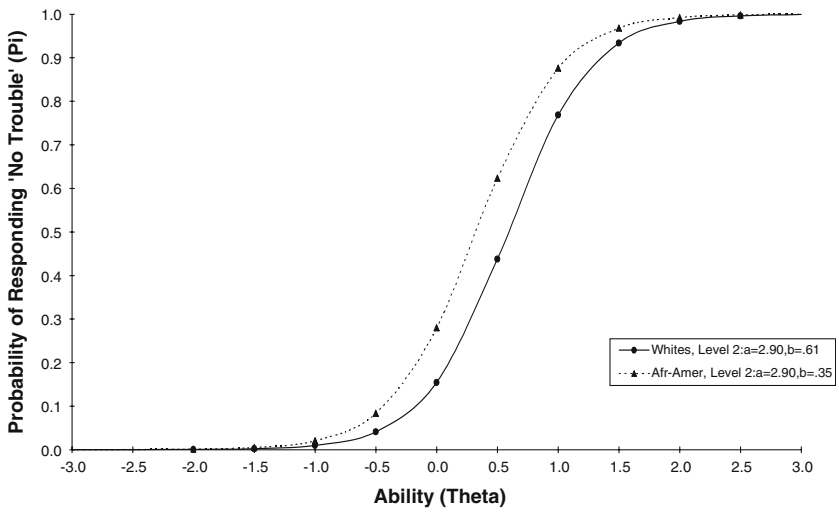

Fig. 1 Physical functioning item set: plot of boundary response functions item 5 – Trouble with a long walk 1.0

they had ''no trouble'' with a long walk. In the absence of DIF, it would be expected that these percentages would be roughly equal. In educational testing this would be regarded as an easier item for African-American people because more African-American people responded that they had ''no trouble'', or ''got it right''. However, this interpretation makes little sense in health and mental health assessment, in which speaking of item severity is more appropriate. (It is also noted that the practice of scoring symptom scales in the positive health direction, as was done in these analyses, might result in some confusion; however, because the scale had been used in this fashion, the decision was made to score the items to conform to past applications.) Shown in the Appendix are formulas and illustrations of calculations.

The item shown in Fig. 1 has equal discrimination (a) parameters because this graphic reflects the result from IRTLR where the discrimination parameters were found to be equivalent, and were constrained to be equal in the final analysis. Figure 2 is an example of non-uniform DIF for an item with three response categories, where the curves cross. The curve associated with level 3, ''not limited at all walking one block'', shows that the probability of response is higher for African-Americans than for Whites at lower levels of ability, but the reverse is observed for higher levels of ability.

A point of clarification is that when the a parameters are freely estimated, they most likely will not be identical; however, the test for non-uniform DIF is to determine whether or not, after constraining the a parameters to be equal, the likelihood of the model is statistically significantly worse, indicating DIF. For this example the a parameter estimates were not exactly equal for the two groups originally (a = 3.53 for Whites and 2.64 for African-Americans) and, in fact, were significantly different, indicating non-uniform DIF, prior to the Bonferroni [26] correction. Thus, in this particular case the actual a's were used in the plots in Fig. 2 in order to illustrate the basic points about non-uniform DIF.

Although the Bonferroni [26] method was used to adjust for multiple comparisons, other approaches, for example, Benjamini–Hochberg (B-H) [27, 28], have been recommended as more powerful for adjustment (see Steinberg [29]; Thissen et al. [30]; and Orlando et al. [31] for examples). For this example, there were few differences in results between the two approaches; thus the Bonferroni method was used for consistency with the approach used in the companion paper.

IRTLR modeling: IRTLR is based on a nested model comparison approach, used for identification of items exhibiting DIF. The concept is to test first a compact (or more parsimonious) model with all parameters constrained to be equal across groups for a studied item (together with the anchor items) (model 1), against an augmented model [2] with one or more parameters of the study item freed to be estimated distinctly for the two groups. The procedure involves comparison of differences in log-likelihoods (–2LL) (distributed as chi-square) associated with nested models; the resulting statistic is evaluated for significance with degrees of freedom equal to the difference in the number of parameter estimates in the two models. For example, the G^2 statistic would have 2 degrees of freedom for each tested item from a 2PL model (i.e., for binary items with difficulty (severity) and discrimination parameters constrained equal vs. estimated freely for the two groups). For the graded response model, the degrees of freedom increase with the number of b (difficulty or severity) parameters estimated. (There is one less b estimated than there are response categories.) It is noted that IRTLR is based on a hierarchical structure, such that b parameters are tested for uniform DIF only if the tests of the a parameters are not significant. Tests of b parameters

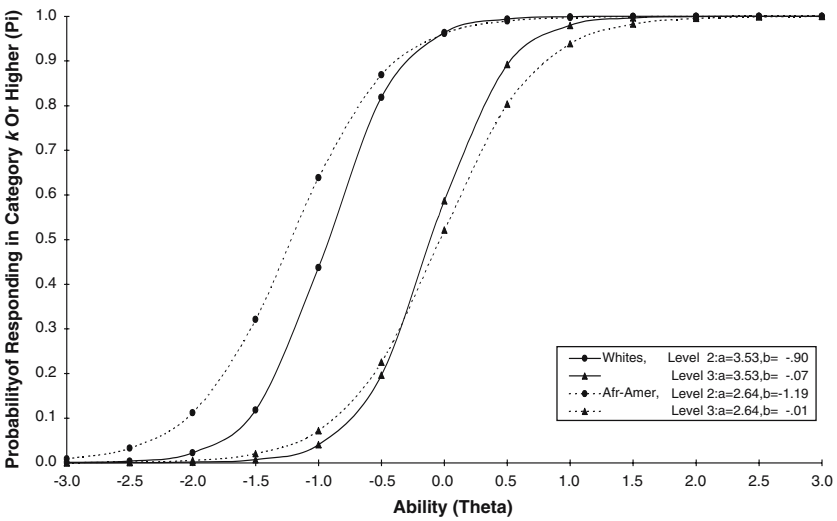

Fig. 2 Physical functioning item set: plot of boundary response functions item 22 – Walking one block

are performed, constraining the *a* parameters to be equal; in that context, if the *a* parameters are found to differ, further tests of the *b* parameters are not warranted. The rationale is that if the slopes are not equal, then the curves must cross, and the threshold parameter is useful only for testing whether the crossing point is near the threshold (in which case the test is not significant) or not (in which case the test is significant). This can be contrasted with one of the OLR approaches examined by Crane and colleagues [1], in which log-likelihood tests of both non-uniform and uniform DIF are examined in a two-step procedure.

Anchor Items: If no prior information about DIF in the item set is available, initial DIF estimates can be obtained by treating each item as a ''studied'' item, while using the remainder as ''anchor'' items. Anchor items are assumed to be without DIF, and are used to estimate theta (ability), and to link the two groups compared in terms of ability. This process of log-likelihood comparisons is performed iteratively for each item. (See the Steps in the analyses below for an illustration.)

While one recommendation (see Thissen [14]) is to reject as anchor items all items meeting the criteria in 1a below, this can result in the selection of a very small anchor set for some comparisons. As discussed below, our view was that a somewhat larger anchor set would be preferable for this example. Anchor item selection is an area that requires additional research. While as few as one anchor item could be used, in general more anchor items may be associated with less conceptual drift in terms of the construct measured, and one simulation study found that a larger number of anchor items (10 as contrasted with 4 or 1) resulted in greater power for DIF detection (Wang et al. [32]).

Steps in the analyses

Presented below is an example of the use of IRTLR. The following procedures for performing the analyses are adapted from Orlando et al. [31]). Examples of the use of IRTLR can be found in Orlando and Marshall [33] and Teresi et al. [34]).

DIF detection

A general description of the steps is provided below; comments refer to the physical function example shown in the tables and graphs.

Identification of anchor items

1a. The first comparison is between a model with all parameters constrained to be equal for any two comparison groups, including the studied item, and a model with separate estimation of all parameters for the studied item. IRTLRDIF is designed using stringent criteria for DIF detection, so that if any model comparison results in a chi-square value greater than 3.84 (d.f. = 1), indicating that at least one parameter differs between the two groups at the .05 level, the item is assumed to have DIF. The results are then reviewed so that the chi-square statistic is evaluated using the correct degrees of freedom, which are dependent on the number of response categories for an item. Non-DIF items are selected as anchor items.

As an example, the G^2 for the overall test of all parameters equal versus all free for one of the studied items was 3.9 (4 d.f.); the G^2 for the *a*'s was .1, (1 d.f.) and the G^2 for the *b*'s was 3.8 (3 d.f. corresponding to the three *b*'s estimated for a four category item). Note that the overall G^2 is the sum of those for the *a*'s and *b*'s because the models are nested. Note also that 3.84 (1 d.f.) is the threshold for testing whether any parameter evidences DIF, assuming a theoretical probability that all DIF is in one parameter.

1b. If there is any DIF, further model comparisons are performed. For the two-parameter model, the *a* parameter (referred to as the slope or discrimination) is constrained to be equal, and the *b* parameter (referred to as difficulty, location, threshold or severity) is estimated freely; this model is compared to that with both *a* and *b* parameters estimated freely (for all other items the parameters are constrained to be equal for both groups). This is a test of DIF in the *a* parameter.

The same procedure is followed with respect to the tests of DIF for the *b* parameters. For all models, all items are constrained to be equal within the anchor set, and the *a* parameter for the item tested is also constrained to be equal. Two models are compared, one in which the *b*'s are the same and one in which the *b*'s are different. The value of G^2 for the last model tests for DIF in the *b* parameters when the *a* parameters are constrained equal and the b parameters are free to be estimated as different. The G^2 for this last model is derived by subtraction of the G^2 for evaluation of the *a* parameters from the overall G^2 value evaluating any difference (G^2 all equal—G^2*a*'s equal).

For example, for item 5 (trouble with a long walk), the overall G^2 for all equal vs. all parameters free is 11.0, with the DIF observed for the *b* parameter (G^2 = 9.7), while the G^2 for the test of the *a* parameter was 1.3. In the current analyses of race groups, 13 items out of 23 physical function items were identified as anchor items.

Purification of the anchor set

2. Even if anchor items were identified prior to the analyses using IRTLRDIF, additional items with DIF may be identified. All of the candidate anchor items are again evaluated, following the procedures described in step 1 (but only for the anchor items), in order to exclude any addi-

tional items with DIF, and to finalize the anchor set. At each step of the purification process, ability estimates (θ) are based on the anchor set used at that stage. It is noted that the item studied is included in the theta estimate. As an example, for the gender comparisons, originally 10 anchor items were identified; at the stage two confirmation process, two additional items with DIF were removed from the anchor set; these are shown in Table 2.

Final DIF detection

3. After the anchor item set is defined, all of the remaining (non-anchor) items are evaluated for DIF against this anchor set. Some items that have been identified as having DIF in earlier stages of the analyses can convert to non-DIF with the use of a purified anchor. However, these items (that converted) are not added to the anchor pool for further iterative purification. At this point in the analyses of the general distress item set, one non-anchor item was no longer found to have DIF. Items with values of G^2 indicative of DIF in this last stage are subject to adjustment of *p* values for multiple comparisons, used in order to reduce over-identification of items with DIF. For this example, the Bonferroni method was used. The *p* value is divided by the number of items.

Final parameter estimation and adjustment for multiple comparisons

4. The final model for a studied scale was estimated using MULTILOG, and all items were included in this model. Parameter estimates for the anchor items as well as those items in which no DIF was observed are set to be equal across groups (using a command for all equal or fixed) in this final model specification; for the items exhibiting DIF in either the *a* or *b* parameters, item parameters are estimated as different (freed) for the two groups. Specifically, if the DIF is only in the *a* parameter, the *a* is estimated as different, together with *b*′s. (As explicated above, IRTLRDIF performs tests of the *b* parameter, constraining the *a* to be equal; thus once the *a* is found to be significant, no further test of the *b* parameter(s) is performed, in which case, the *b* parameter(s) would be set to be different.) If the DIF is in the *b* parameter, only the *b* parameter is estimated as different.

The final parameter estimates and their standard errors were obtained from applications of MULTILOG. Theta estimates at this point are based on the entire item set with parameters estimated as described above. These thetas can be used in the evaluation of DIF magnitude and impact, described below. An area for study is the identification of the best theta estimate for use when individual ability estimates are to be used, e.g., in computerized adaptive testing or for construction of a ''DIF-free'' theta estimate for use in analyses of relationships among variables.

Evaluation of DIF magnitude

5. Following these analyses, graphs of item response functions are useful in examining magnitude of DIF. 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. Examination of the magnitude of DIF has been based on evaluation of theoretically invariant parameters or statistics flowing from a model, such as the odds ratio.

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.

6. A method for quantification of the difference in the average expected item scores is the non-compensatory DIF (NCDIF) index (the average squared difference in expected item scores for a given individual as a member of the focal group, and as a member of the reference group) used by Raju and colleagues [16]. (See also Chang and Mazzeo [35]), who demonstrated that items with identical IRFs or expected scores have equivalent item category response functions under certain polytomous response models, including the graded response model used here. The implication of this work is a generalization from binary to some of the more commonly used polytomous response models of the IRF invariance assumptions that permit DIF detection.)

In essence this method provides an estimate of what expected score would obtain for an individual if s/he was scored based on the parameters and ability estimates for group X, and then based on the ability and parameter estimates for group Y. (See the Appendix.) The advantage of this magnitude measure is that NCDIF is based on the actual distribution of individual estimated thetas, rather than on an arbitrary range of ability. While chi-square tests of significance are available, these were found to be too stringent, over identifying DIF. Cutoff values established based on simulations [36, 37], provide an estimate of the magnitude of item-level DIF. For example, for dichotomous items the NCDIF cutoff is 0.006; for polytomous items with three response options the cutoff is .024; for four response options the cutoff is 0.054; for five it is .096; and for polytomous items with six response options the cutoff is 0.150. Use of this method requires that thetas be estimated separately for each group, and equated together with the item parameters prior to calculation of expected

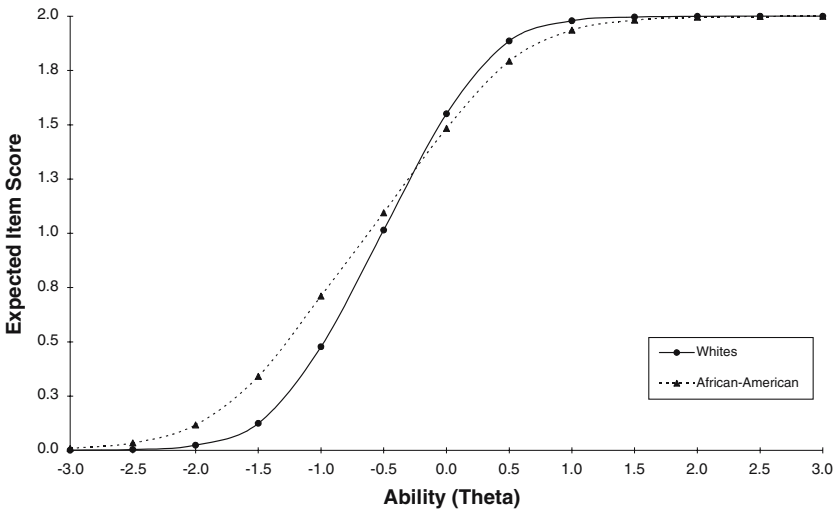

Fig. 3 Physical functioning item set: expected item score function by race groups item 22 – Walking one block

item scores. (Equating constants are purified iteratively, if DIF is detected.)

Evaluation of impact of DIF

7. Expected item scores (see Fig. 3) can be summed to produce an expected scale score, which provides evidence regarding the effect of the DIF on the total score (see Fig. 4). Group differences in these test response functions provide measures of impact.

Methods

Measures

Twenty three physical functioning ability items and 15 general distress items were analyzed. These items were selected from a larger item set taken from four measures described elsewhere in this special issue, and in the companion paper. The process by which the items were selected included exploratory and confirmatory factor analyses; these methods are described elsewhere in this special issue. The 23 items measuring physical function were scored in the positive direction, and positive physical function was measured. The 15 items measuring general distress were also scored in the positive direction, so that a high score was indicative of positive affect.

Sample

Data were collected as part of the Quality of Life Evaluation in Oncology Project funded by the National Cancer Institute (RO1 CA 60068, David Cella PI). This study was of patients with cancer or HIV/AIDS. Data were analyzed with respect to age, gender and race. The sample sizes used in the analyses shown in the Figures and Tables were 236 African-Americans and 1324 whites, 719 females and 914

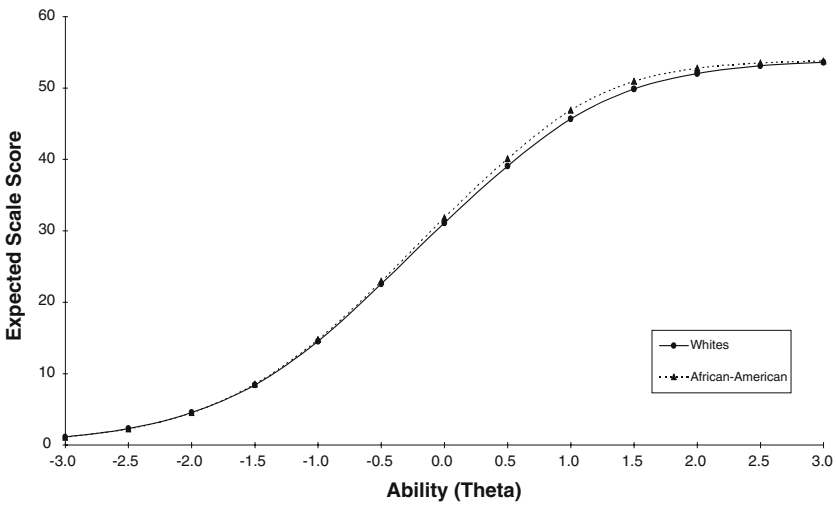

Fig. 4 Physical functioning item set: total expected response function comparing race groups

males and 1183 younger (less than 66 years of age) and 449 older subjects.

Software

Software used was IRTLRDIF developed by Thissen [14], and available on his website, and MULTILOG (Thissen [13]). The IRTLR approach to DIF detection is discussed in Thissen et al. [38]). IRTLRDIF can be used for the analyses performed in the first steps, followed by application of MULTILOG.

Follow-up examination of magnitude of item-level DIF was conducted using expected item scores and area statistics. These expected scores can be plotted for different values of theta using software such as EXCEL (see Fig. 3 and the Appendix).

Additionally the non-compensatory DIF index of Raju (Raju and colleagues [16]; Flowers and colleagues [17]) contained in DFIT (Raju [39]) was examined. (See also Collins et al. [40] and Morales and colleagues [41] for examples.) In order to assess DIF magnitude, Raju's program DFITP5 was used. To run this program, it is necessary to run MULTILOG separately for the two groups under study, and then to place the parameter estimates for the two groups on the same metric. (When thetas and item parameters are obtained separately for each group, they have to be equated in order to be on the same metric scale. Equating is performed iteratively; originally no DIF is assumed; however, if DIF is detected, the item showing DIF is excluded from the equating algorithm.) For this purpose, Baker's EQUATE program [42] was used in an iterative fashion. In the first run, all items in the scale were used as the anchor set. Next, the program DFITP5 was run, and those items with values above the recommended cut-off for NCDIF were excluded from the anchor set for the next run of the EQUATE program. The equating constants resulting from this second run of EQUATE were the ones used for the final run of DFITP5, to evaluate DIF magnitude.

Impact of DIF on the total score was examined using test response functions. The method for integration of magnitude and impact measures with significance testing is an area requiring further research.

Results

Example of IRTLR using items measuring physical functioning and general distress

Tables 1 through 6 show the final result for the physical function and general distress item sets. The tables show the anchor items without DIF, and the studied items with separately estimated parameters for the two groups. This result

Table 1 Item parameters and standard errors for the anchor items and studied items with DIF from the physical functioning item set (PF23): Comparison of race groups (White vs. African-American)

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF	*b*DIF*
Difficulty bending or lifting	White	2.09 (0.10)	–1.30 (0.08)	–0.72 (0.06)	–0.08 (0.05)	1.08 (0.05)	NS, Anchor Item	
	Afr. Amer.							
Difficulty doing household chores	White	3.01 (0.14)	–1.00 (0.06)	–0.47 (0.04)	0.03 (0.04)	0.86 (0.04)	NS, Anchor Item	
	Afr. Amer.							
Difficulty bathing, brushing teeth, or grooming myself	White	2.48 (0.17)	–2.28 (0.16)	–1.73 (0.11)	–1.23 (0.07)	–0.52 (0.05)	NS, Anchor Item	
	Afr. Amer.							
Trouble with strenuous activities (carrying)	White	2.66 (0.18)	0.70 (0.04)				NS, Anchor Item	
	Afr. Amer.							
Trouble with a long walk	White	2.90 (0.21)	0.61 (0.04)				1.3 (0.254)	**9.7 (0.002)**
	Afr. Amer.	2.90 (0.21)	0.35 (0.09)					
Trouble with a short walk	White	2.92 (0.23)	–0.84 (0.06)				2.9 (0.089)	4.3 (0.038)
	Afr. Amer.	2.92 (0.23)	–1.10 (0.11)					
Have to stay in bed or chair most of the day	White	2.15 (0.18)	–0.74 (0.07)				NS, Anchor Item	
	Afr. Amer.							

Table 1 continued

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF	*b*DIF*
Need help eating, dressing, washing, toileting	White	1.93 (0.21)	–1.67 (0.15)				NS, Anchor Item	
	Afr. Amer.							
Limited in work or other daily activities	White	2.00 (0.11)	–0.94 (0.08)	–0.29 (0.06)	0.83 (0.05)		8.5 (0.004)	2.5 (0.475)
	Afr. Amer.	2.89 (0.37)	–0.93 (0.12)	–0.34 (0.10)	0.58 (0.10)			
Limited in hobbies, leisure activities	White	1.89 (0.11)	–0.98 (0.09)	–0.30 (0.06)	0.72 (0.05)		6.0 (0.014)	1.3 (0.729)
	Afr. Amer.	2.56 (0.36)	–0.97 (0.15)	–0.37 (0.11)	0.54 (0.10)			
Have lack of energy	White	1.85 (0.07)	–1.32 (0.08)	–0.34 (0.06)	0.49 (0.05)	1.84 (0.07)	0.1 (0.752)	**25.1(<0.001)**
	Afr. Amer.	1.85 (0.07)	–1.50 (0.17)	–0.60 (0.14)	0.26 (0.12)	1.20 (0.14)		
Able to work	White	1.97 (0.07)	–1.07 (0.07)	–0.47 (0.06)	0.23 (0.05)	1.09 (0.05)	0.3 (0.584)	**17.3 (0.002)**
	Afr. Amer.	1.97 (0.07)	–0.98 (0.13)	–0.18 (0.12)	0.36 (0.13)	0.94 (0.13)		
Vigorous activities (running, lifting)	White	2.29 (0.10)	0.91 (0.05)	2.04 (0.06)			0.0 (1.000)	**45.1 (<0.001)**
	Afr. Amer.	2.29 (0.10)	0.35 (0.11)	1.47 (0.14)				
Moderate activities (moving table)	White	3.30 (0.18)	–0.20 (0.04)	0.79 (0.04)			NS, Anchor Item	
	Afr. Amer.							
Lifiting or carrying groceries	White	3.31 (0.17)	–0.61(0.05)	0.53 (0.04)			0.3 (0.584)	**30.1 (<0.001)**
	Afr. Amer.	3.31 (0.17)	–0.35 (0.09)	0.82 (0.10)				
Climb several fights of stairs	White	2.71 (0.15)	–0.28 (0.04)	0.89 (0.04)			NS, Anchor Item	
	Afr. Amer.							
Climb 1 flight of stairs	White	2.83 (0.16)	–0.94 (0.06)	0.16 (0.04)			NS, Anchor Item	
	Afr. Amer.							
Bending, kneeling, stooping	White	2.09 (0.12)	–0.83 (0.07)	0.56 (0.04)			NS, Anchor Item	
	Afr. Amer.							
Walk more than a mile	White	3.16 (0.16)	0.16 (0.04)	0.99 (0.04)			0.0 (1.000)	**13.3 (0.001)**
	Afr. Amer.	3.16 (0.16)	–0.10 (0.09)	0.87 (0.09)				
Walk several blocks	White	3.79 (0.23)	–0.30 (0.03)	0.47 (0.03)			NS, Anchor Item	
	Afr. Amer.							
Walk 1 block	White	3.53 (0.25)	–0.90 (0.06)	–0.07 (0.04)			3.8 (0.050)	5.1 (0.078)
	Afr. Amer.	2.64 (0.40)	–1.19 (0.17)	–0.01 (0.10)				
Bathing or dressing	White	2.34 (0.17)	–1.75 (0.12)	–0.74 (0.06)			NS, Anchor Item	
	Afr. Amer							

*DIF significant after Bonferroni adjustment is bolded

Table 2 Item parameters and standard errors for the anchor items and studied items with DIF from the physical functioning item set (PF23): Comparison of gender groups (Male vs. Female)

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF	*b*DIF*
Difficulty bending or lifting	Male	2.13 (0.08)	–1.68 (0.09)	–1.07 (0.07)	–0.43 (0.06)	0.74 (0.06)	2.5 (.114)	12.1 (.017)
	Female	2.13 (0.08)	–1.79 (0.10)	–1.26 (0.08)	–0.59 (0.07)	0.51 (0.07)		
Difficulty doing household chores	Male	3.08 (0.14)	–1.44 (0.06)	–0.92 (0.04)	–0.41 (0.04)	0.41 (0.04)	NS, Anchor Item	
	Female							
Difficulty bathing, brushing teeth, or grooming myself	Male	2.59 (0.13)	–2.54 (0.17)	–1.94 (0.10)	–1.42 (0.07)	–0.77 (0.06)	3.4 (.065)	**35.9 (<.001)**
	Female	2.59 (0.13)	–2.86 (0.20)	–2.39 (0.15)	–1.93 (0.10)	–1.14 (0.07)		
Trouble with strenuous activities (carrying)	Male	2.74 (0.18)	0.19 (0.05)				2.6 (.107)	6.8 (.009)
	Female	2.74 (0.18)	0.34 (0.06)					
Trouble with a long walk	Male	2.91 (0.21)	0.08 (0.05)				NS	NS
	Female	2.91 (0.21)	0.17 (0.05)					
Trouble with a short walk	Male	2.92 (0.28)	–1.31(0.06)				NS, Anchor Item	
	Female							
Have to stay in bed or chair most of the day	Male	2.22 (0.15)	–1.07 (0.07)				0.2 (.655)	5.2 (.023)
	Female	2.22 (0.15)	–1.26 (0.08)					
Limited in work or other daily activities	Male	2.13 (0.08)	–1.26 (0.07)	–0.63 (0.06)	0.40 (0.06)		0.1 (.752)	10.8 (.013)
	Female	2.13 (0.08)	–1.53 (0.08)	–0.87 (0.07)	0.32 (0.07)			
Need help eating, dressing, washing, toileting	Male	1.96 (0.21)	–2.09 (0.14)				NS, Anchor Item	
	Female							
Limited in hobbies, leisure activities	Male	2.00 (0.08)	–1.34 (0.08)	–0.61(0.07)	0.33 (0.06)		1.6 (.206)	12.0 (.007)
	Female	2.00 (0.08)	–1.51 (0.10)	–0.89 (0.08)	0.17 (0.06)			
Were you short of breath	Male	1.09 (0.05)	–2.80 (0.18)	–1.85 (0.13)	0.10 (0.09)		0.1 (.752)	**14.7 (.002)**
	Female	1.09 (0.05)	–2.99 (0.22)	–2.24 (0.16)	–0.27 (0.11)			
Have lack of energy	Male	1.85 (0.07)	–1.76 (0.10)	–0.70 (0.07)	0.13 (0.06)	1.37 (0.08)	1.9 (.168)	**20.4 (<.001)**
	Female	1.85 (0.07)	–1.83 (0.11)	–0.99 (0.08)	–0.13 (0.07)	1.20 (0.09)		
Able to work	Male	2.01 (0.07)	–1.33 (0.07)	–0.75 (0.06)	–0.09 (0.06)	0.70 (0.06)	0.5 (.480)	13.4 (.009)
	Female	2.01 (0.07)	–1.64 (0.09)	–0.95 (0.07)	–0.28 (0.07)	0.54 (0.07)		
Vigorous activities (running, lifting)	Male	2.21 (0.12)	0.37 (0.04)	1.51 (0.07)			NS, Anchor Item	
	Female							
Moderate activities (moving table)	Male	3.33 (0.18)	–0.63 (0.04)	0.35 (0.04)			NS, Anchor Item	
	Female							
Lifting or carrying groceries	Male	3.31 (0.17)	–1.06 (0.05)	0.07 (0.04)			0.4 (.527)	**13.9 (.001)**
	Female	3.31 (0.17)	–0.91 (0.05)	0.22 (0.05)				
Climb several flights of stairs	Male	2.76 (0.15)	–0.72 (0.04)	0.45 (0.04)			NS, Anchor Item	
	Female							

Table 2 continued

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF	*b*DIF*
Climb 1 flight of stairs	Male	2.86 (0.16)	–1.36 (0.06)	–0.26 (0.04)			NS	NS
	Female	2.86 (0.16)	–1.36 (0.06)	–0.26 (0.04)				
Bending, kneeling, stooping	Male	2.12 (0.11)	–1.24 (0.07)	0.20 (0.06)			0.8 (.371)	6.6 (.037)
	Female	2.12 (0.11)	-1.28 (0.08)	0.03 (0.06)				
Walk more than a mile	Male	3.17 (0.17)	–0.33 (0.03)	0.53 (0.04)			NS, Anchor Item	
	Female							
Walk several blocks	Male	3.80 (0.22)	–0.73 (0.03)	0.05 (0.03)			NS, Anchor Item	
	Female							
Walk 1 block	Male	3.23 (0.21)	–1.41 (0.05)	–0.51 (0.04)			NS	NS
	Female	3.23 (0.21)	–1.41 (0.05)	–0.51 (0.04)				
Bathing or dressing	Male	2.33 (0.11)	–2.13 (0.11)	–1.07 (0.07)			0.1 (.752)	9.0 (.011)
	Female	2.33 (0.11)	–2.23 (0.12)	–1.30 (0.08)				

*DIF significant after Bonferroni adjustment is bolded

Table 3 Item parameters and standard errors for the anchor and studied items with DIF from the Physical Functioning set (PF23): Comparison of age groups (age 66 and over, age 65 and under)

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF	*b*DIF*
Difficulty bending or lifting	Age ≤ 65	2.14 (.08)	–1.65 (.08)	–1.08 (.07)	–.47 (.06)	.63 (.05)	.0 (1.00)	**22.8 (<.001)**
	Age 66+	2.14 (.08)	–1.64 (.13)	–1.04 (.10)	–.31(.08)	.92 (.09)		
Difficulty doing household chores	Age ≤ 65	3.10 (.12)	–1.33 (.06)	–.78 (.05)	–.27 (.04)	.54 (.04)	3.0 (.083)	**19.7 (.001)**
	Age 66+	3.10 (.12)	–1.41 (.10)	–.95 (.08)	–.44 (.07)	.37 (.06)		
Difficulty bathing, brushing teeth, or grooming myself	Age ≤ 65	2.60 (.20)	–2.58 (.20)	–2.07 (.13)	–1.54 (.08)	–.83 (.06)	**9.3 (.002)**	16.0 (.003)
	Age 66+	2.35 (.29)	–2.71 (.27)	–2.04 (.20)	–1.61 (.15)	–.92 (.09)		
Trouble with strenuous activities (carrying)	Age ≤ 65	2.75 (.18)	.34 (.04)				2.0 (.157)	6.6 (.010)
	Age 66+	2.75 (.18)	.33 (.07)					
Trouble with a long walk	Age ≤ 65	2.99 (.21)	.18 (.04)				.1 (.752)	**13.0 (<.001)**
	Age 66+	2.99 (.21)	.25 (.07)					
Trouble with a short walk	Age ≤ 65	2.98 (.34)	–1.23 (.07)				**16.4 (<.001)**	4.0 (.046)
	Age 66+	2.67 (.43)	–1.23 (.11)					
Have to stay in bed or chair most of the day	Age ≤ 65	2.20 (.21)	–.99 (.07)				7.4 (.007)	**21.1 (<0.001)**
	Age 66+	2.23 (.35)	–1.28 (.14)					

Table 3 continued

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF	*b*DIF*
Need help eating, dressing, washing, toileting	Age ≤ 65	1.84 (.24)	–2.04 (.18)				**29.3 (<.001)**	**15.1 (<.001)**
	Age 66+	2.32 (.49)	–1.94 (.21)					
Limited in work or other daily activities	Age ≤ 65	2.19 (.08)	–1.18 (.06)	–.53 (.05)	.58 (.05)		.6 (.439)	**41.1 (<.001)**
	Age 66+	2.19 (.08)	–1.56(.12)	–.97 (.10)	.10 (.08)			
Limited in hobbies, leisure activities	Age ≤ 65	2.07 (.08)	–1.18 (.07)	–.50 (.06)	–.47 (.05)		3.7 (.054)	**54.2 (<.001)**
	Age 66+	2.07 (.08)	–1.69 (.13)	–1.00 (.10)	.00 (.08)			
Were you short of breath	Age ≤ 65	1.08 (.08)	–2.83 (.22)	–1.95 (.15)	.02 (.07)		2.0 (.157)	7.2 (.066)
	Age 66+	–	–	–	–			
Have lack of energy	Age ≤ 65	1.87 (.06)	–1.60(.08)	–.63 (.06)	.19 (.05)	1.41 (.07)	2.9 (.089)	**25.6 (<.001)**
	Age 66+	1.87 (.06)	–2.02(.17)	–1.01 (.10)	–.12 (.08)	1.25 (.11)		
Able to work	Age ≤ 65	2.04 (.07)	–1.25(.06)	–.62 (.05)	.03 (.06)	.75 (.06)	1.1 (.294)	**53.7 (<.001)**
	Age 66+	2.04 (.07)	-1.74 (.13)	–1.13 (.10)	–.39 (.09)	.61 (.09)		
Vigorous activities (running, lifting)	Age ≤ 65	2.24 (.10)	.43 (.05)	1.54 (.07)			2.8 (.094)	**12.9 (.002)**
	Age 66+	2.24 (.10)	.51 (.08)	1.73 (.12)				
Moderate activities (moving table)	Age ≤ 65	3.34 (.18)	–.55 (.04)	.43 (.04)			NS, Anchor Item	
	Age 66+							
Lifting or carrying groceries	Age ≤ 65	3.23 (.18)	–.91 (.04)	.22 (.03)			NS, Anchor Item	
	Age 66+							
Climb several flights of stairs	Age ≤ 65	2.76 (.15)	–.63 (.04)	.53 (.04)			1.9 (.168)	4.6 (.100)
	Age 66+	2.76 (.15)	–.63 (.04)	.53 (.04)				
Climb 1 flight of stairs	Age ≤ 65	3.00 (.20)	–1.26 (.07)	–.17 (.04)			5.1 (.024)	1.1 (.294)
	Age 66+	2.51 (.27)	–1.33 (.12)	–.19 (.07)				
Bending, kneeling, stooping	Age ≤ 65	33 (.16)	1.13 (.07)	12 (.05)			4.6 (.032)	**33.8 (<.001)**
	Age 66+	1.78 (.22)	–1.24 (.14)	.44 (.10)				
Walk more than a mile	Age ≤ 65	3.29 (.17)	–.29 (.04)	.59 (.04)			0.2 (.655)	**24.1 (<.001)**
	Age 66+	3.29 (.17)	–.10 (.06)	.64 (.06)				
Walk several blocks	Age ≤ 65	3.81 (.22)	–.64 (.03)	.13 (.03)			NS, Anchor Item	
	Age 66+							
Walk 1 block	Age ≤ 65	3.19 (.20)	–1.33 (.06)	–.42 (.04)			1.0 (.317)	3.8 (.150)
	Age 66+	3.19 (.20)	–1.33 (.06)	–.42 (.04)				
Bathing or dressing	Age ≤ 65	2.20 (.19)	–2.11 (.16)	–1.06 (.08)			4.2 (.040)	.7 (.705)
	Age 66+	2.60 (.38)	–2.10 (.20)	–1.16 (.11)				

*DIF significant after Bonferroni adjustment is bolded

Table 4 Item parameters and standard errors for the anchor and studied items with DIF from the General Distress set (GD15): Comparison of race groups (White vs. African-American)

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*b5*	*a*DIF	*b*DIF*
I feel sad	White	2.54 (.12)	–2.36 (.13)	–1.47 (.07)	–.66 (.04)	.43 (.04)		NS, Anchor Item	
	Afr .Amer.								
I feel nervous	White	2.34 (.11)	–2.24 (.13)	–1.56 (.08)	–.76 (.05)	.46 (.04)		NS, Anchor Item	
	Afr. Amer.								
I worry about dying	White	1.34 (.06)	–2.70 (.16)	–1.94 (.11)	–1.02 (.08)	.24 (.06)		2.3 (.129)	**27.8 (<.001)**
	Afr. Amer.	1.34 (.06)	–1.85 (.23)	–1.47 (.21)	–.95 (.18)	–.12(.15)			
Able to enjoy life	White	1.44 (.05)	–2.76 (.15)	–1.83 (.09)	–.74 (.07)	.51 (.07)		.6 (.439)	13.8 (.008)
	Afr. Amer.	1.44 (.05)	–2.49 (.32)	–1.43 (.19)	–.32 (.16)	.54 (.15)			
Content with my QOL right now	White	1.40 (.08)	–1.52 (.11)	–.93 (.08)	.00 (.06)	1.11 (.08)		NS, Anchor Item	
	Afr. Amer.								
Frequently feel anxious	White	2.48 (.13)	–1.85 (.11)	–1.22 (.07)	–.57 (.05)	.68 (.05)		7.4 (.007)	4.5 (0.343)
	Afr. Amer.	1.90 (.26)	–1.89 (.33)	–1.49 (.24)	–.63 (.15)	.72 (.15)			
Felt tense	White	2.64 (.12)	–2.10 (.11)	–1.28 (.06)	.30 (.04)			1.1 (.294)	8.9 (0.031)
	Afr. Amer.	2.64 (.12)	–1.78 (.21)	–1.16 (.13)	.08 (.09)				
Felt worried	White	2.83 (.12)	–1.78 (.08)	–.87(.05)	.73 (.04)			.7 (.403)	**23.0 (<.001)**
	Afr. Amer.	2.83 (.12)	–1.19 (.11)	–0.66 (.11)	.66 (.10)				
Felt irritable	White	2.35 (.12)	–2.18 (.12)	–1.29 (.07)	.32 (.04)			NS, Anchor Item	
	Afr. Amer.								
Felt depressed	White	3.96 (.20)	–1.87 (.09)	–1.11 (.04)	.24 (.03)			1.7 (.192)	12.1 (.006)
	Afr. Amer.	3.96 (.20)	–1.47(.14)	–1.03 (.09)	.08 (.08)				
Have you been a very nervous person	White	1.66 (.09)	–3.01 (.21)	–2.22 (.14)	–1.53 (.09)	–.55 (.06)	.66 (.06)		
	Afr. Amer.								
Felt down in the dumps	White	2.13 (.14)	–3.03 (.26)	–2.28 (.15)	–1.74 (.10)	–.98 (.07)	–.09 (.05)	4.1 (.043)	9.5 (.091)
	Afr. Amer.	1.80 (.25)	–2.71 (.53)	–2.17 (.35)	–1.54 (.24)	–.86 (.17)	.14 (.14)		
Felt calm and peaceful	White	1.85 (.09)	–2.19 (.14)	–1.31(.08)	–.45 (.05)	.22 (.05)	1.96 (.09)	3.1 (.078)	10.5 (.062)
	Afr. Amer.	1.85 (.09)	–2.19 (.14)	–1.31 (.08)	–.45 (.05)	.22 (.05)	1.96 (.09)		
Felt downhearted	White	2.15 (.11)	–2.62(.19)	–2.02(.13)	–1.50 (.09)	–.52 (.05)	.67 (.06)	4.8 (.028)	8.9 (.113)
	Afr. Amer.	1.86 (.24)	–2.70 (.48)	–2.09(.35)	–1.38 (.21)	–.55 (.14)	.48 (.14)		
Been a happy person	White	1.80 (.09)	–2.52(.17)	–1.69(.10)	–.75 (.06)	–.09 (.05)	1.74 (.08)	NS, Anchor Item	
	Afr. Amer.								

*DIF significant after Bonferroni adjustment is bolded

Table 5 Item parameters and standard errors for the anchor and studied items with DIF from the General Distress set (GD15): Comparison of gender groups (Male vs. Female)

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*b5*	*a*DIF	*b*DIF*
I feel sad	Male	2.51 (.09)	–2.39 (.14)	–1.62 (.09)	–.72 (.06)	.34 (.05)		.3 (.584)	11.5 (.021)
	Female	2.51 (.09)	–2.49 (.19)	–1.43 (.08)	–.69 (.06)	.46 (.06)			
I feel nervous	Male	2.22 (.10)	–2.33 (.13)	–1.62 (.08)	–.81(.05)	.43 (.04)		NS, Anchor Item	
	Female								
I worry about dying	Male	1.31 (.08)	–2.52 (.17)	–1.88 (.13)	–1.05 (.08)	.14 (.06)		3.6 (.058)	.9 (.343)
	Female	1.31 (.08)	–2.52 (.17)	–1.88 (.13)	–1.05 (.08)	.14 (.06)			
Able to enjoy life	Male	1.59 (.12)	–2.38 (.21)	–1.58 (.13)	–.56 (.08)	.65 (.08)		5.7 (.017)	**29.1 (<.0001)**
	Female	1.30 (.12)	–3.26 (.38)	–2.09 (.22)	–.94 (.12)	.27 (.10)			
Content with my QOL right now	Male	1.45 (.05)	–1.30 (.09)	–.78 (.08)	.14 (.08)	1.21 (.09)		3.3 (.069)	**39.8 (<.0001)**
	Female	1.45 (.05)	–1.91 (.12)	–1.19 (.09)	–.26 (.09)	.85 (.09)			
Frequently feel anxious	Male	2.33 (.09)	–1.88 (.13)	–1.32 (.08)	–.54 (.06)	.67 (.06)		.1 (.752)	10.1 (.039)
	Female	2.33 (.09)	–1.95 (.12)	–1.33 (.08)	–.73 (.07)	.59 (.06)			
Felt tense	Male	2.64 (.14)	–2.09 (.10)	–1.31 (.06)	.22 (.04)			.3 (.331)	4.2 (.241)
	Female	2.64 (.14)	–2.09 (.10)	–1.31 (.06)	.22 (.04)				
Felt worried	Male	2.76 (.13)	–1.72 (.08)	–.89 (.04)	.69 (.04)			NS, Anchor Item	
	Female								
Felt irritable	Male	2.34 (.12)	–2.25 (.12)	–1.35 (.07)	.27 (.04)			NS, Anchor Item	
	Female								
Felt depressed	Male	3.81 (.20)	–1.85 (.08)	–1.15 (.04)	.17 (.03)			NS, Anchor Item	
	Female								
Have you been a very nervous person	Male	1.49 (.11)	–3.31 (.35)	-2.50 (.22)	–1.73 (.15)	–.72 (.09)	.62 (.08)	8.0 (.005)	6.8 (.236)
	Female	1.89 (.14)	–2.84 (.25)	–2.05 (.15)	–1.39 (.11)	-.46 (.07)	.63 (.08)		
Felt down in the dumps	Male	2.05 (.07)	–2.88 (.19)	-2.30 (.13)	–1.69 (.09)	–.96 (.07)	.01 (.06)	.0 (1.000)	16.6 (.005)
	Female	2.05 (.07)	–3.00 (.24)	–2.32 (.16)	–1.84 (.11)	–1.06 (.08)	–.23 (.07)		
Felt calm and peaceful	Male	1.81 (.09)	-2.27 (.13)	–1.36 (.07)	–.50 (.05)	.18 (.05)	1.92 (.09)	NS, Anchor Item	
	Female								
Felt downhearted	Male	2.07 (.10)	–2.68 (.17)	–2.09 (.12)	–1.53 (.08)	–.58 (.05)	.60 (.05)	NS, Anchor Item	
	Female								
Been a happy person	Male	1.75 (.08)	–2.61 (.17)	–1.75 (.10)	–.80 (.06)	–.13 (.05)	1.71 (.08)	NS, Anchor Item	
	Female								

*DIF significant after Bonferroni adjustment is bolded

represents the final analyses, so that if no new DIF was observed in any of the prior iterative purification stages, the a's are estimated as the same. Tables 7 and 8 show the summary results, including the analyses of magnitude. Figures 5 through 9 show the expected item and scale scores for items that were significant after Bonferroni correction, depicting DIF magnitude and impact, respectively.

Physical function

As shown in Table 1, prior to adjustment for multiple comparisons, 13 anchor items were identified and 10 items were identified that showed DIF with respect to race (summarized in Table 7); three with non-uniform DIF. (One item, ''walk one block'' was borderline, $p = .051$). After adjustment, six items showed DIF (Table 1), four with relatively higher magnitude (NCDIF—expected item score difference values above cutoff) (Table 7). For example, after the adjustment, the six items that evidenced uniform DIF were: ''trouble with a long walk''; ''lack of energy'', ''able to work'', ''vigorous activities'', ''lifting or carrying groceries'', ''walk more than a mile''. The item: ''walk one block'', also showed significant non-uniform DIF using IRTLRDIF, prior to, but not after the Bonferroni correction for multiple comparisons. Four of these items evidenced a relatively large magnitude of DIF: ''long walk'', ''lack of energy'', ''vigorous activity'' and ''lifting or carrying groceries''. Most of these items were more severe indicators for White than for African-American respondents; the exception was ''lifting or carrying

Table 6 Item parameters and standard errors for the anchor and studied items with DIF from the General Distress set (GD15): Comparison of age groups (age 66 and over, age 65 and under)

Content	Group	*a*	*b1*	*b2*	*b3*	*b4*	*b5*	*a*DIF	*b*DIF*
I feel sad	Age < 65	2.47 (.09)	–2.78 (.13)	–1.85 (.07)	–1.00 (.05)	.14 (.05)		.6 (.439)	9.5 (0.050)
	Age 66+	2.47 (.09)	–2.72 (.28)	–1.86 (.15)	–1.12 (.09)	.07 (.07)			
I feel nervous	Age < 65	2.20 (.08)	–2.67 (.13)	–1.90 (.08)	–1.11 (.05)	.19 (.05)		0.0 (1.000)	14.6 (.006)
	Age 66+	2.20 (.08)	–2.64 (.25)	–2.18 (.21)	–1.20 (.10)	–.04 (.07)			
I worry about dying	Age < 65	1.27 (.05)	–2.84 (.14)	–2.16 (.11)	–1.31 (.08)	.04 (.07)		.2 (.129)	**26.4 (<.001)**
	Age 66+	1.27 (.05)	–3.09 (.31)	–2.53 (.24)	–1.66 (.17)	–.54 (.12)			
Able to enjoy life	Age < 65	1.43 (.09)	–3.07 (.19)	–2.12 (.12)	–1.04 (.07)	.18 (.06)		NS, Anchor Item	
	Age 66+								
Content with my QOL right now	Age < 65	1.33 (.09)	–1.96 (.12)	–1.31 (.09)	–.32 (.07)	.82 (.10)		5.3 (.021)	3.9 (.420)
	Age 66+	1.61 (.18)	–1.66 (.20)	–1.22 (.15)	–.41 (.10)	.64 (.11)			
Frequently feel anxious	Age < 65	2.31 (.11)	–2.24 (.11)	–1.64 (.07)	–.95 (.05)	.32 (.04)		1.8 (.180)	4.5 (.343)
	Age 66+	2.31 (.11)	–2.24 (.11)	–1.64 (.07)	–.95 (.05)	.32 (.04)			
Felt tense	Age < 65	2.62 (.14)	–2.42 (.10)	–1.63 (.06)	–.10 (.04)			NS, Anchor Item	
	Age 66+								
Felt worried	Age < 65	2.75 (.13)	–2.04 (.08)	–1.21 (.05)	.37 (.04)			NS, Anchor Item	
	Age 66+								
Felt irritable	Age < 65	2.34 (.12)	–2.57 (.12)	–1.67 (.07)	–.05 (.04)			NS, Anchor Item	
	Age 66+								
Felt depressed	Age < 65	3.85 (.20)	–2.17 (.08)	–1.47 (.04)	–.14 (.03)			NS, Anchor Item	
	Age 66+								
Have you been a very nervous person	Age < 65	1.64 (.09)	–3.42 (.21)	–2.61 (.13)	–1.89 (.09)	–.91 (.06)	.31 (.06)	NS, Anchor Item	
	Age 66+								
Felt down in the dumps	Age < 65	2.04 (.12)	–3.27 (.21)	–2.64 (.14)	–2.08 (.09)	–1.33 (.06)	–.41(.05)	NS, Anchor Item	
	Age 66+								
Felt calm and peaceful	Age < 65	1.84 (.05)	–2.73 (.12)	–1.70 (.07)	–.84 (.06)	–.12 (.06)	1.55 (.09)	.4 (.527)	**20.3 (.001)**
	Age 66+	1.84 (.05)	–2.06 (.16)	–1.55 (.11)	–.74 (.09)	–.18 (.09)	1.69 (.14)		
Felt downhearted	Age < 65	2.06 (.10)	–3.01 (.17)	–2.41 (.12)	–1.86 (.08)	–.90 (.05)	.28 (.05)	NS, Anchor Item	
	Age 66+								
Been a happy person	Age < 65	1.77 (.05)	–3.11 (.16)	–2.10 (.09)	–1.14 (.06)	–.44 (.06)	1.39 (.08)	0.00 (1.000)	12.4 (.030)
	Age 66+	1.77 (.05)	–2.37 (.19)	–1.90 (.15)	–1.04 (.10)	–.44 (.10)	1.39 (.12)		

*DIF significant after Bonferroni adjustment is bolded

Table 7 Summary of DIF analyses of the Physical Functioning items (PF23): Race, gender and age groups

Item	Item Name	Item Wording	Anchor Item			Type of DIF, if Present			DIF After Bonferroni Adjustment			Magnitude (Expected Item Score Difference: NCDIF)*		
			Race	Sex	Age	Race	Sex	Age	Race	Sex	Age	Race	Sex	Age
1	CARES1	Difficulty bending or lifting	√				U	U			√		.005	.040
2	CARES3	Difficulty doing household chores	√	√				U			√		.015	.007
3	CARES4	Difficulty bathing, brushing teeth, or grooming myself	√				U	NU		√	√		.024	.003
4	EORTC1	Trouble with strenuous activities (carrying)	√				U	U					.008*	.001
5	EORTC2	Trouble with a long walk				U		U	√		√	.010*	.004	.003
6	EORTC3	Trouble with a short walk		√		U		NU			√	.004	.002	.001
7	EORTC4	Have to stay in bed or chair most of the day	√				U	NU			√			.001
8	EORTC5	Need help eating, dressing, washing, toileting	√	√				NU			√		.001	.001
9	EORTC6	Limited in work or other daily activities				NU	U	U			√	.026	.003	.048
10	EORTC7	Limited in hobbies, leisure activities				NU	U	U			√	.018	.004	.059*
11	EORTC8	Were you short of breath	√				U			√			.008	.001
12	FACT 1	Have lack of energy				U	U	U	√	√	√	.008*	.009	.024
13	FACT 27	Able to work				U	U	U	√		√	.010	.010	.051
14	RAND3	Vigorous activities (running, lifting)		√		U		U	√		√	.097*	.001	.029*
15	RAND4	Moderate activities (moving table)	√	√	√								.004	.002
16	RAND5	Lifting or carrying groceries			√	U	U		√	√		.027*	.018	
17	RAND6	Climb several fights of stairs	√	√									.002	.002
18	RAND7	Climb 1 flight of stairs	√					NU					.002	.003
19	RAND8	Bending, kneeling, stooping	√				U	NU			√		.001	.026*
20	RAND9	Walk more than a mile		√		U		U	√		√	.016	.004	.020
21	RAND10	Walk several blocks	√	√	√								.003	.007
22	RAND11	Walk 1 block				NU						.007	.007	.006
23	RAND12	Bathing or dressing	√				U	NU						.001

*Difference greater than threshold; difference less than .001 is not shown

groceries'', which was a more severe indicator for African-Americans. For example, examination of the expected item scores (Fig. 5) show that for most items the solid curves for Whites is below the curve for African-Americans, indicating that conditional on functional status, on average White respondents are less likely to respond that they are capable of performing the task. The reverse pattern is observed with respect to the curve for ''lifting or carrying groceries''.

The analysis based on gender initially identified 8 anchor items and 15 with DIF (Table 2); however, after purification 12 items with DIF were identified, all with uniform DIF. After adjustment, four items with uniform DIF were identified: 3, 11, 12, 16, ''difficulty with personal care'', ''short of breath'', ''lack of energy'', ''problems lifting or carrying groceries'' (see Table 7). Among these, none evidenced DIF of high magnitude. (It is noted that one item (''strenuous activities''), identified before the Bonferroni correction as evidencing uniform DIF, was also identified as having higher magnitude DIF, however, the value was just over the threshold.) As shown in Fig. 6, most of these items were more severe indicators for males than females; the exception was ''lifting or carrying groceries'', which was a more severe indicator for females. For this latter item, on average, it takes somewhat more capability for females than for males to claim that they have little difficulty ''lifting or carrying groceries''. Numerous items (14 out of 23) evidenced DIF with respect to age, even after the Bonferroni adjustment (see Table 7.) However, three were of high magnitude: ''limited in hobbies, leisure activities'', ''vigorous activities'', and ''bending, kneeling and stooping''. There was a mixture in terms of whether the items were more severe for older or younger persons, with some (e.g., ''vigorous activities'') more severe for older persons, and some (e.g., ''able to work'') for younger persons. It is noted that ''bending, kneeling, stooping'' showed non-uniform DIF for age, and was a relatively poor discriminator for younger people

Table 8 Summary of DIF analyses of the General Distress items (GD15): Race, gender and age groups

Item	Item Name	Item Wording	Anchor Item			Type of DIF, if Present			DIF After Bonferroni Adjustment			Magnitude (Expected Item Score Difference: NCDIF)*		
			Race	Sex	Age	Race	Sex	Age	Race	Sex	Age	Race	Sex	Age
1	FACT20R	I feel sad	√				U	U				.004	.010	.003
2	FACT23R	I feel nervous	√	√				U						.002
3	FACT24R	I worry about dying				U		U	√		√		.003	.013
4	FACT29	Able to enjoy life			√	U	NU			√		.010	.011	.001
5	FACT33	Content with my QOL right now	√				U	NU		√		.012	.031	.016
6	CARES16R	Frequently feel anxious				NU	U					.004		.005
7	EORT21M	Felt tense			√	U						.006	.007	
8	EORT22M	Felt worried		√	√	U			√			.003	.004	.002
9	EORT23RM	Felt irritable	√	√	√							.004		
10	EORT24RM	Felt depressed		√	√	U						.007	.003	
11	RAND24M	Have you been a very nervous person	√		√		NU					.001	.007	.007
12	RAND25M	Felt down in the dumps			√	NU	U					.014	.003	.001
13	RAND26RM	Felt calm and peaceful		√				U			√	.008	.007	.004
14	RAND28M	Felt downhearted		√	√	NU						.002		.001
15	RAND30RM	Been a happy person	√	√				U				.002	.003	.004

*Difference greater than threshold; difference less than .001 is not shown

(a = .33), as contrasted with older persons (a = 1.78). This means that the item was not well-related to physical function for younger people. Similarly, ''shortness of breath'' was not a well-discriminating item, in general.

Based on prior experience examining DIF in health-related applications, the results indicate relatively low magnitude of DIF in the physical function item set for race and gender; however, somewhat more DIF was evidenced with respect to age. About 60% of the items showed DIF, even after Bonferroni correction, and it was difficult to obtain an anchor set. Originally, three items without DIF were identified; several iterations were necessary in order to obtain a final anchor set of items. Further testing indicated that all three of these items evidenced DIF, and a different three-item anchor set was produced. Because the DIF was in different directions, overall DIF cancellation was observed at the scale level; however, use of individual items out of context of the scale, for example in computer adaptive testing, could be problematic for individual assessment. Evaluation of the impact of DIF using the test response functions (shown in Fig. 4 for race, Fig. 6 for gender, and Fig. 7 for age) indicates that the impact of DIF on the test score is trivial.

General distress

Examination of the general distress item set for DIF based on race shows that six anchor items were initially identified (see Table 4). Eight out of 9 items originally identified with DIF evidenced DIF after purification, but before Bonferroni correction. After correction only two showed DIF, both uniform: ''worry abut dying'' and ''felt worried'' (see Table 8). Neither item demonstrated high magnitude DIF. While the direction was mixed, the indicators were somewhat more severe for African-Americans than for Whites, indicating that more positive mental health was required for endorsement of the item at most response levels. (However, inspection of Fig. 8 shows that the difference was small.) Seven anchor items were used in the analyses of gender DIF (see Table 5). After purification, six out of 15 items showed DIF for gender; however, only two were significant after Bonferroni adjustment (''able to enjoy life'' and ''content with my quality of life''), and none demonstrated high magnitude DIF. As shown in Fig. 8, the indicators were more severe for men. (This can also be seen in Table 5 where the b (severity) parameters are higher for males than for females.

Age comparisons demonstrated seven items with DIF in the first iteration (see Table 6), and six after purification (Table 8), but before correction (all with uniform DIF, except for ''content with my quality of life''). After correction, two items showed uniform DIF: ''worry about dying'' and ''felt calm and peaceful''. ''Worry about dying'' was a more severe indicator for the younger cohort. Items that did not discriminate as well as others for most groups were ''content with my quality of life'', ''worry about dying'' and for women, ''able to enjoy

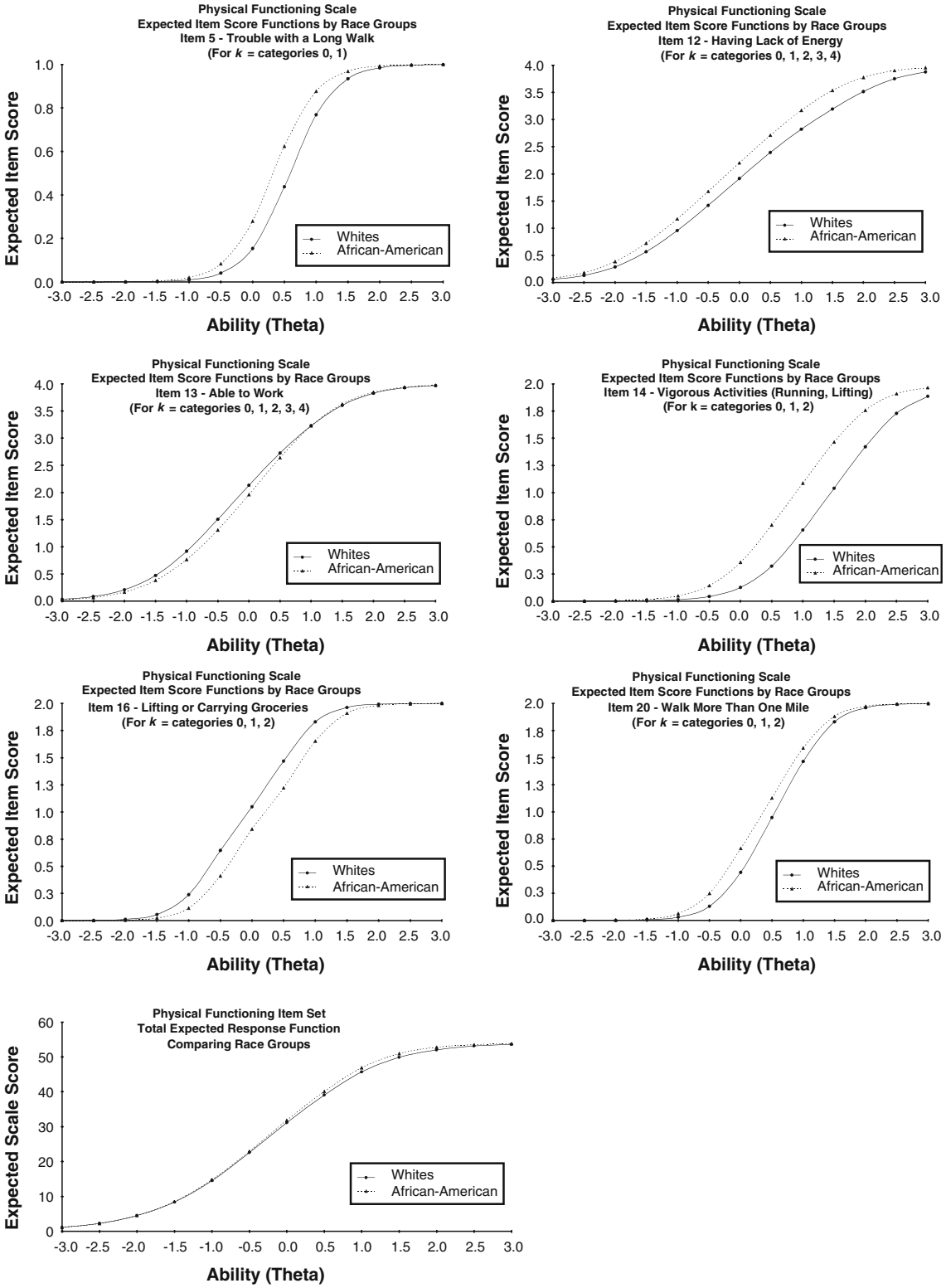

Fig. 5 Expected item and scale score functions: physical functioning item set, race groups

life''. The magnitude of DIF was not large, and the impact trivial (Figs. 8, 9).

Summary of findings: The analyses presented above were intended to illustrate the IRTLRDIF procedures and the calculation of magnitude and impact measures. The substantive findings indicated that there was a relatively small magnitude of DIF in the item sets. Examination of the expected item scores, and calculation of NCDIF for the

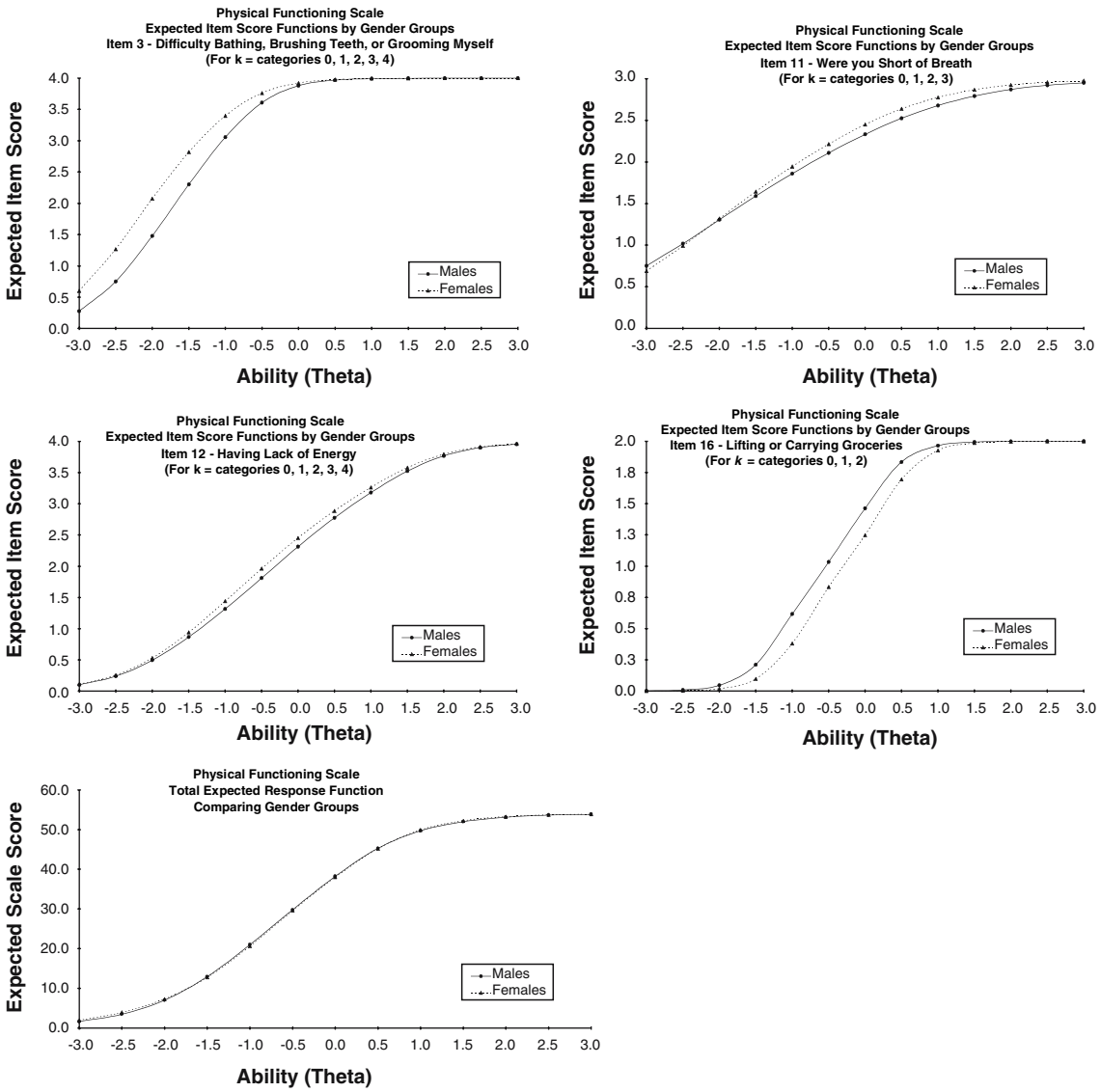

Fig. 6 Expected item and scale score functions physical functioning item set, gender groups

race group comparison identified four items of higher magnitude; these included three items related to mobility and physical functioning: ''trouble with a long walk''; ''vigorous activities'' and ''lifting or carrying''. ''Lack of energy'' also evidenced a relatively greater magnitude of DIF. Four items were identified with gender DIF after adjustment for multiple comparisons, none with high magnitude. Three items were of higher magnitude of DIF for age group comparisons: ''limited in hobbies, leisure activities, ''vigorous activities'', ''bending, kneeling, stooping''. In total, six items showed relatively larger magnitude of DIF with respect to physical function across the three comparisons: ''trouble with a long walk'' (race), ''vigorous activities'' (race, age), ''bending, kneeling stooping'' (age), ''lifting or carrying groceries'' (race), ''limited in hobbies, leisure'' (age), ''lack of energy'' (race). None of the general distress items evidenced high magnitude DIF, although ''worrying about dying'' showed some DIF with respect to both age and race, after adjustment.

Discussion

The fact that many physical function items showed DIF with respect to age, even after Bonferroni adjustment, indicates that the instrument may be performing differently for these groups. While the magnitude and impact of DIF at the item and scale level was minimal, caution should be exercised in the use of subsets of these items, as might occur with selection for clinical decisions or for computerized adaptive testing. In the companion paper, Crane and

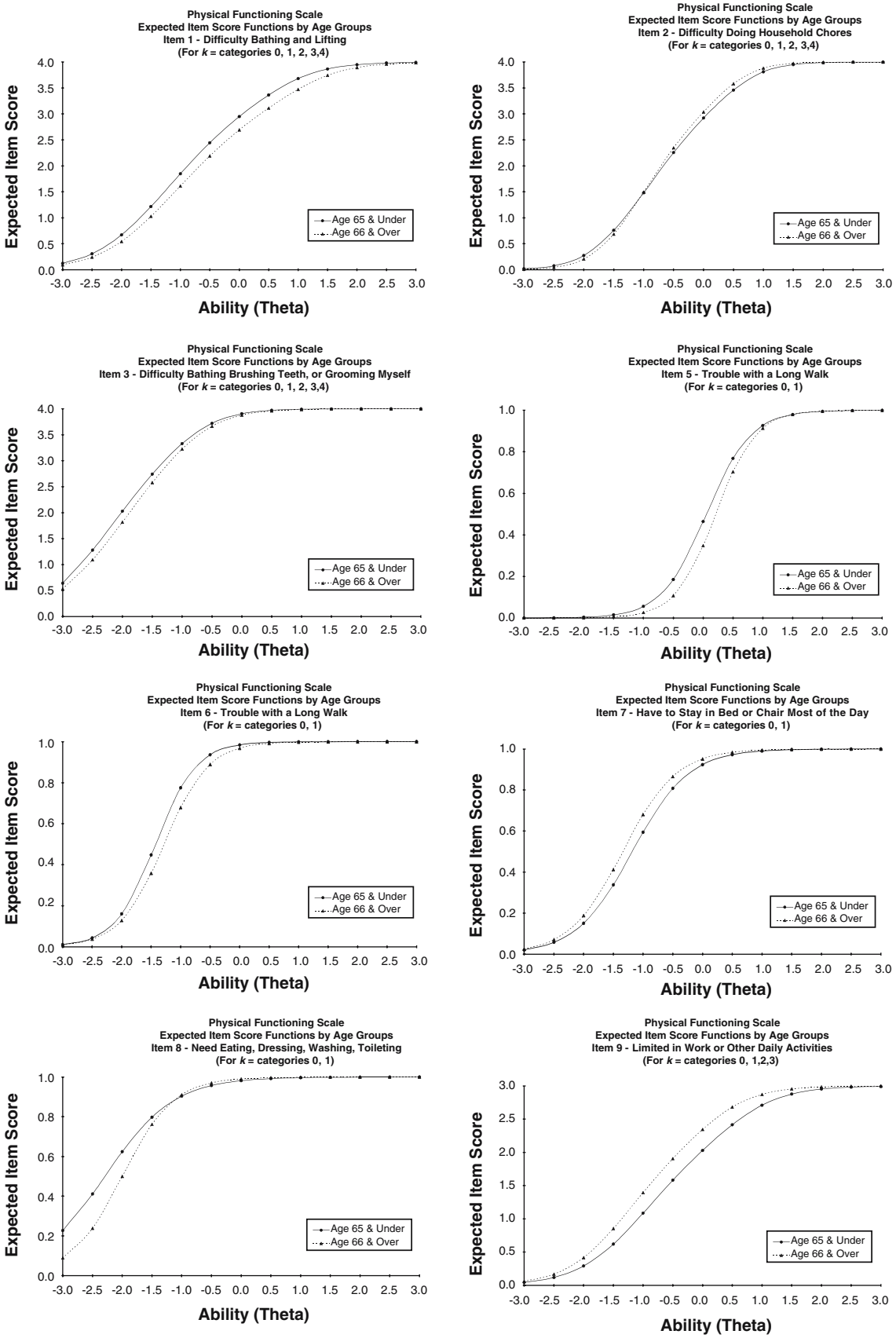

Fig. 7 Expected item and scale score functions: physical functioning item set, age groups

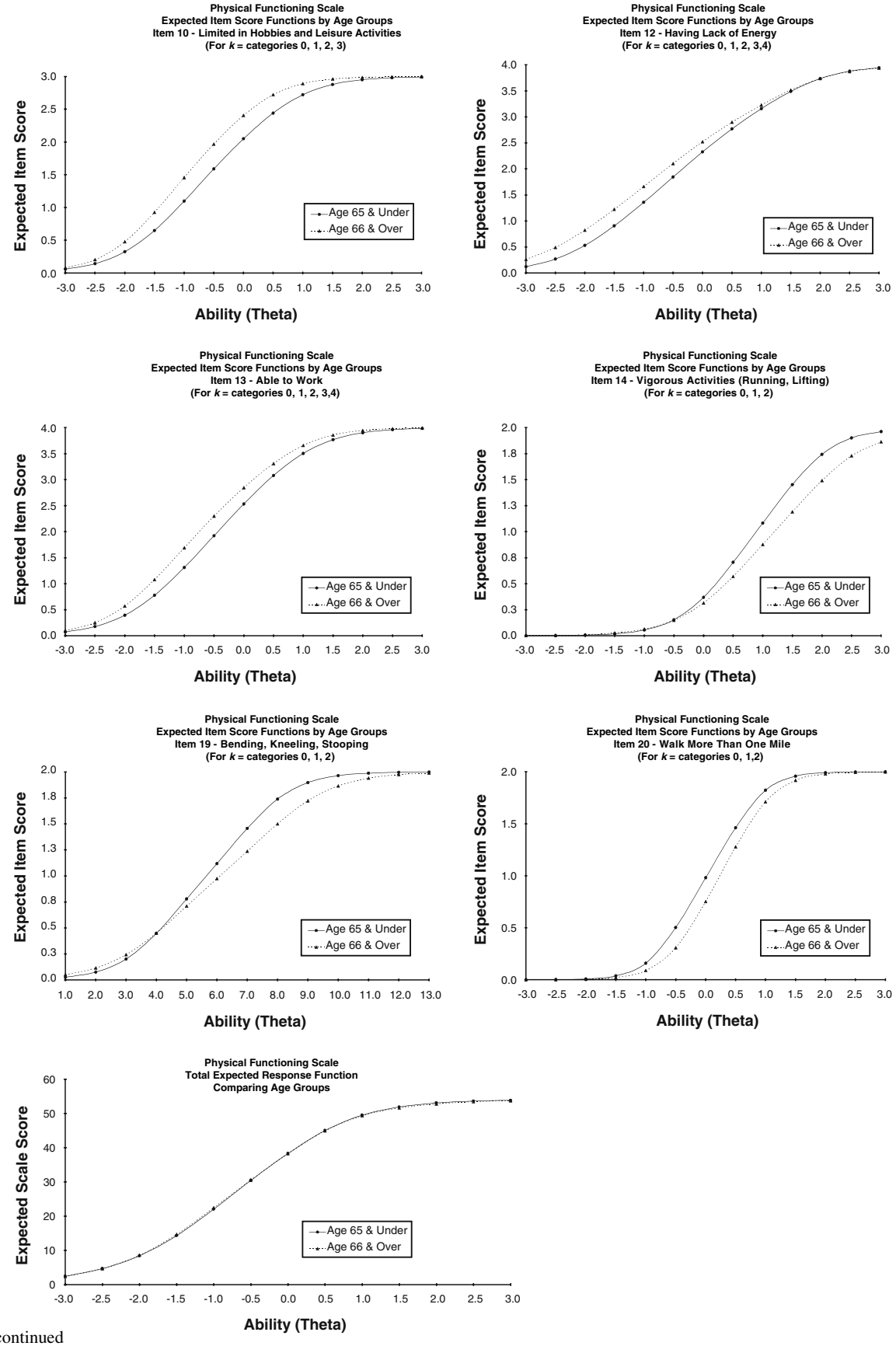

Fig. 7 continued

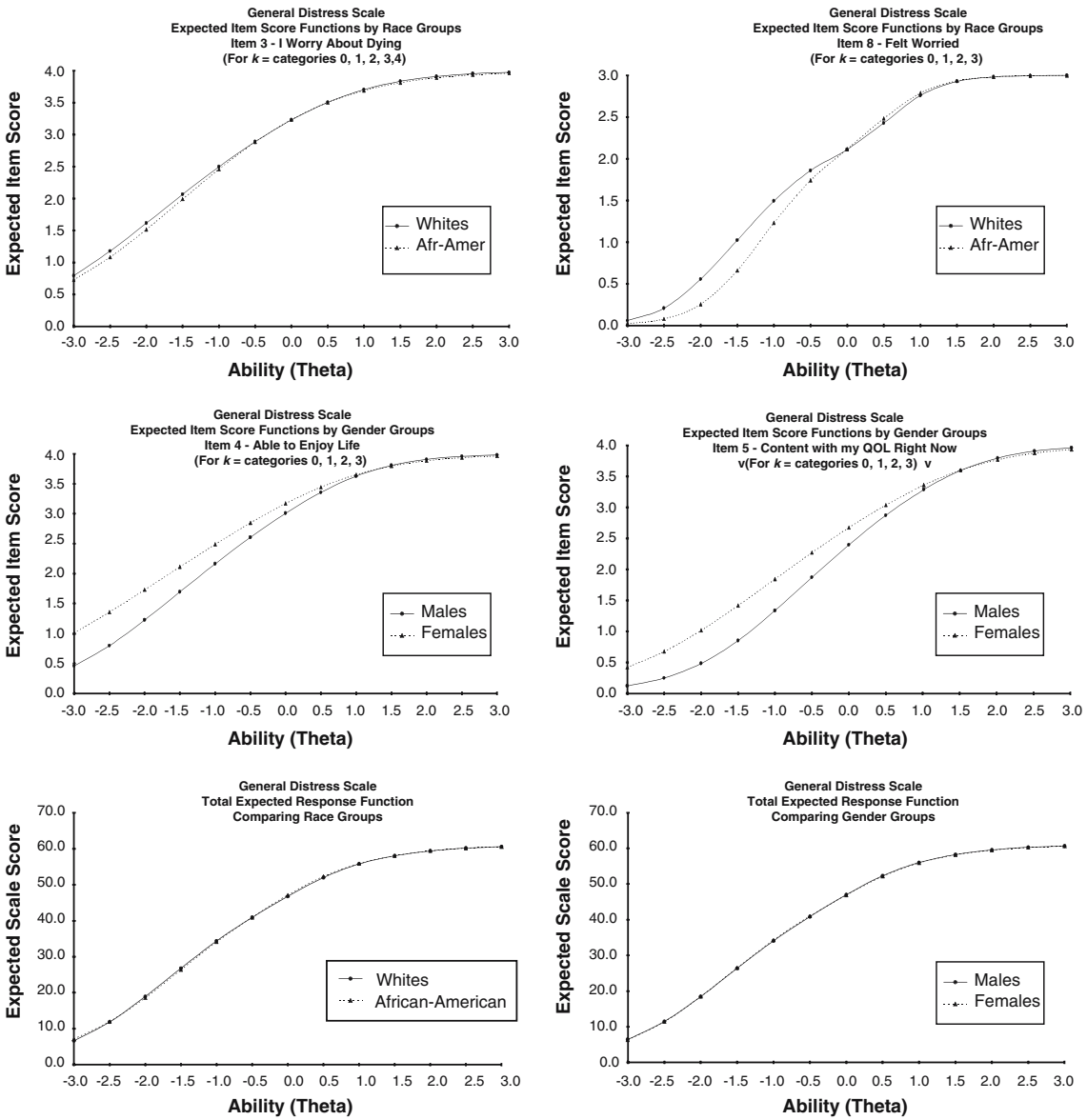

Fig. 8 Expected item and scale score functions: general distress item set, race and gender groups

colleagues [1] also found that the impact of DIF was small; however, these authors concluded that the impact of DIF related to race on the General Distress scale could affect some individuals.

Comparison of OLR and IRTLR results

The findings from the two analyses (OLR and IRTLR) of the General Distress scale agree in terms of the number of items identified pre- and post-adjustments, and with respect to DIF magnitude; however, there is some disagreement in terms of the individual items identified. Using the OLR method, based on significance tests after Bonferroni correction, six items were identified as having DIF. Using the IRTLR approach, and after Bonferroni correction, five items were observed to have DIF. After considering DIF magnitude, the OLR method identified two items with DIF: worry about dying, and content with quality of life. These two items were also identified with DIF using the IRTLR approach; however, none of the items evidenced high magnitude DIF using DFIT NCDIF criteria.

Crane and colleagues [1] identified 14 items with DIF in the Physical Functioning scale, using significance tests with Bonferroni adjustments. The IRTLR approach identified 16. Incorporation of a DIF magnitude measure into the OLR modeling procedure resulted in the identification of five items. Use of the DFIT magnitude adjustment, in the context of the IRT approach, reduced the number of identified items to seven; however, only two of the items were in common across the methods: limited in hobbies

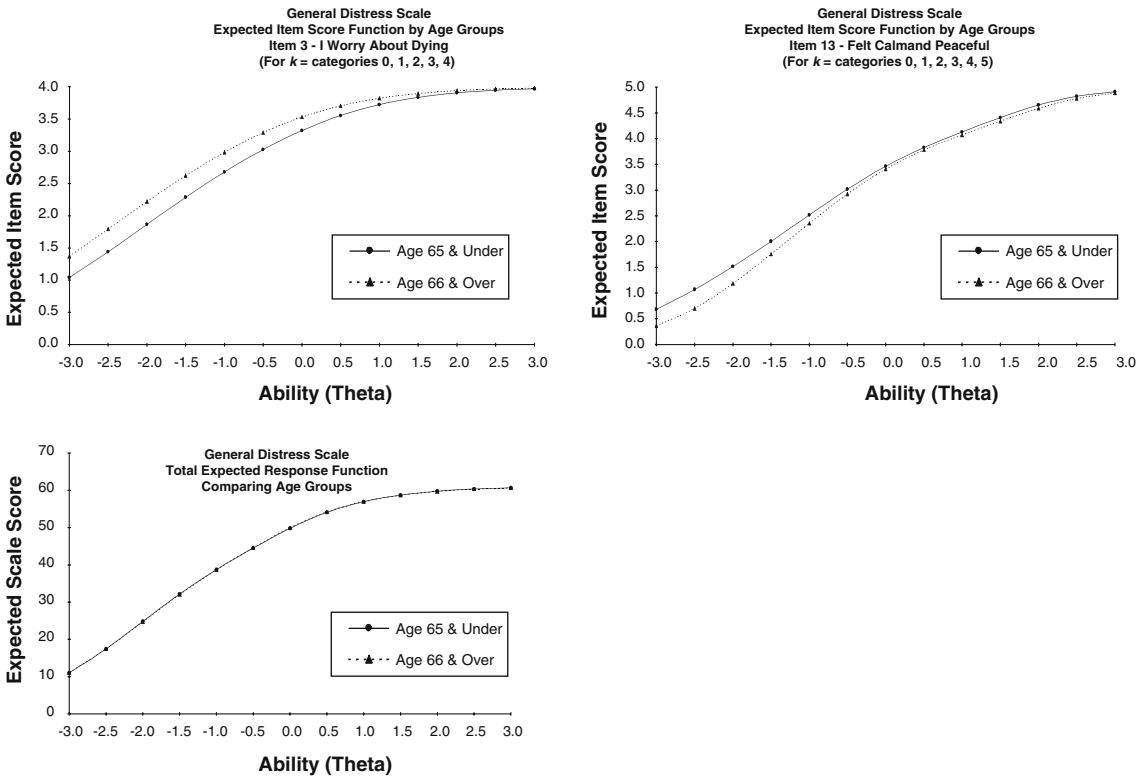

Fig. 9 Expected item and scale score functions: general distress item set, age groups

and bending, kneeling, stooping. Both sets of analyses demonstrated minimal impact of DIF; however, Crane and colleagues [1] observed that there could be DIF impact associated with race on the General Distress scale for some individuals.

Caveats: As with all parametric models, lack of model fit can result in errors in DIF detection, as can lack of proper purification through selection of anchor items. Finally, although not discussed, a first step is examination of dimensionality. If the assumption (of most IRT models used in DIF detection) of unidimensionality is not met, DIF detection will be inaccurate. While extensive tests of dimensionality were conducted, and the item sets were selected to be essentially unidimensional, an unanswered question is what constitutes being unidimensional enough for IRT DIF methods? Typically, violations of assumptions or model fit will lead to false DIF detection. Therefore, it is important to select the correct model prior to application of DIF methods. As mentioned above, the issues of selection of anchor items, and of criteria for DIF detection, including the integration of significance and magnitude measures remain as issues requiring investigation. Additionally, the issue of which purified theta is best to use requires study. Further research is needed regarding the criteria and guidelines appropriate for DIF detection in the context of health-related items. Different magnitude measures and procedures for flagging salient DIF may have contributed to the discrepancies in DIF detection between the two methods. Further simulation studies are needed. Despite these possible caveats, the IRTLR method has been used frequently to detect DIF in educational and psychological assessment measures, and as such is a relatively mature method. While many DIF detection methods exist, both of the methods presented in these two companion papers can be recommended for use in the evaluation of health and quality-of-life measures because both allow the identification of non-uniform DIF, which may be of concern in such measures.

Acknowledgements The authors thank Douglas Holmes, Ph.D. for his review of several versions of this manuscript. The authors also thank three anonymous reviewers and the editor for their helpful comments related to an earlier version of this manuscript. These analyses were conducted on behalf of the Statistical Coordinating Center to the Patient Reported Outcomes Information System (PROMIS) (AR052177). Funding for analyses was provided in part by the National Institute on Aging, Resource Center for Minority Aging Research at Columbia University (AG15294), and by the National Cancer Institute through the Veteran's Administration Measurement Excellence and Training Resource Information Center (METRIC). An earlier version of this paper was presented at the National Institutes of Health Conference on Patient Reported Outcomes, Bethesda, June, 2004.

Appendix

Illustration: Calculation of Boundary and Category Response Functions, Expected Item and Scale Scores and Non-compensatory Differential Item Functioning (DIF) Indices: Polytomous items with ordinal response categories

This appendix is an illustration of the calculation of several indices used in determining the presence, magnitude and impact of DIF using item response theory. Illustrations can also be found in Collins et al. [40]; Orlando-Edelen et al. [31]; Thissen et al. [38]; and Thissen [13, 14].

Boundary Response Functions: The Samejima [25] graded response model, which assumes ordinal categories can be used to model polytomous items (see also Cohen et al. [43]. The model is based on calculation of a series of cumulative dichotomies resulting in cumulative probabilities of responding in a category or higher. One models the probability that a randomly selected individual with a specific level of physical functioning will respond in category k or higher. The boundary response function defines the cumulative probability of scoring in category k or higher:

$$P_{ik}(\theta) = 1/\{1 + \exp[-a_i(\theta - b_{ik})]\}.$$

There are k–1 such dichotomies. For a three category item, scored 0,1,2: the first cumulative dichotomy is between people who have a zero response vs those who selected response 1 or 2. The second cumulative dichotomy is between those who selected 0 or 1 vs 2 and higher. To illustrate using the example shown in Fig. 2, calculations are presented for θ = –1.0.

The probability of category 0 or higher = 1 (because every one scores either 0 or higher).

For category 1: P(x = 1 or higher) = $1/\{1 + \exp[-a(\theta - b_{k1})]\}$

For Whites: P(x = 1 or higher) = $1/\{1 + \exp[-3.53((-1) -(-.90))]\}$ = .4127

For African-Americans: P(x = 1 or higher) = $1/\{1 + \exp[-2.64((-1)-(-1.19))]\}$ = .6228

For category 2:

P(x = 2 or higher (there is no higher)) = $1/1 + \exp[-a(\theta)-b_{k2})]$

For Whites: P(x = 2 or higher) = $1/\{1 + \exp[-3.53((-1) -(-.07))]\}$ = .0360

For African-Americans: P(x = 2 or higher) = $1/\{1 + \exp[-2.64((-1)-(-.01))]\}$ = .0683

Category response functions: The above formula does not give the probability of responding in a specific category; to get this probability, the adjacent probability is subtracted out. Note that boundary response functions are usually used because they have a consistent form across levels; category response functions do not, and are more difficult to compare and interpret. Note also that when an item is binary, the category response function for the second category (for an item coded 0,1 this is 1) is the same as the item response (boundary response) function.

To obtain the category response P(x = k), subtract out the probability that P is in a higher category: P(k)–P(k+1).

For k = 0 $P_{i0}(\theta) = [P_{i0}(\theta)–P_{i1}(\theta)] = [1–P_{i1}(\theta)]$

For k = 1 $P_{i1}(\theta) = [P_{i1}(\theta)–P_{i2}(\theta)]$

For k = 2 $P_{i2}(\theta) = [P_{i2}(\theta)– 0]$

For $\boldsymbol{\theta = -1.0}$ **:**

For Whites: P_{i0} = 1 – .4127 = .5873

For African-Americans: P_{i0} = 1 – .6228 = .3772

For Whites: $P_{i1}(\theta = -1.0)$ = .4127 – .0360 = .3767

For African-Americans: $P_{i1}(\theta = -1.0)$ = .6228 – .0683 = .5545

For Whites: $P_{i2}(\theta = -1.0)$ = .0360

For African-Americans: $P_{i2}(\theta = -1.0)$ = .0683

The shapes of the CRFs are not the same, because they are no longer cumulative. A person with a specific theta level will have a separate probability of response for each response category. This category response function provides the probability that a randomly selected individual at say θ = 0 (average ability) will respond in category k.

Computing expected item and test scores: Expected item and test (scale) scores can be computed for both dichotomous and polytomous items. Lord and Novick and Birnbaum [24, p 386] introduce the notion of true scores in the context of IRT. A person's true score is their expected score, expressed in terms of probabilities for binary items and in terms of weighted probabilities for polytomous items. The test characteristic curve described in Lord and Novick related true score or averaged expected test score to theta.

For a dichotomous item scored 0 and 1, the expected or true score is simply the probability of scoring in the '1' category, given an individual's estimated ability or $P_i(\theta_s)$, where $P_i(\theta_s)$ is the probability of scoring a '1' on item i for subject s.

$$P_i(\theta_s) = 1/\{1 + \exp[-a_i(\theta_s - b_i)]\}.$$

(Here it is assumed that c (guessing) parameters are estimated at 0).

For a polytomous item, taking a graded response form, the expected score is the sum of the weighted probabilities of scoring in each of the possible categories for the item. For an item with 5 response categories, coded 0 to 4, for example, this sum would be:

$$0 * [P_{i0}(\theta_s)] + 1 * P_{i1}(\theta_s) + 2 * P_{i2}(\theta_s) + 3 * P_{i3}(\theta_s) + 4 * P_{i4}(\theta_s)$$

For the graded response item with k categories, there are k–1 estimated *b*s and so in the example above, there will be four probabilities computed.

Recall that in the graded response model, the boundary response function defines the cumulative probability of scoring *in category k or higher*:

$$P_{ik}(\theta_s) = 1/\{1 + \exp[-a_i(\theta_s - b_{ik})]\}.$$

Thus, the probability of scoring above category k must be subtracted out in order to obtain the *probability of scoring in the category*. (This was illustrated in the previous section.)

Note that $P_i(\theta_s) = 1/\{1 + \exp[-a_i(\theta_s - b_{i1})]\} - 1/\{1 + \exp[-a_i(\theta_s - b_{i2})]\}$.

As an example, the expected or true score as a member of the African-American group is the sum of the weighted probabilities of scoring in each of the possible categories for each θ, coded 0,1,2.

$$0 + 1 * P_{i1}(\theta) + 2 * P_{i1}(\theta)$$

A true or expected score for an individual of mild disability (θ = –1.0) as a member of the white group would be:

$$0(.59) + 1(.3767) + 2(.0360) = .4487$$

A true or expected score for an individual of mild disability (θ = –1.0) as a member of the African-American group would be:

$$0 + 1(.5545) + 2(.0683) = .6905$$

The expected test score for a subject with estimated ability θ is simply the sum of the expected item scores for that individual. Plots of expected scores against theta can then be constructed for given values of theta. Individual expected scores are used in the calculation of magnitude and impact indices discussed below.

Computing Non-Compensatory Differential Item Functioning (NCDIF): Two measures developed by Raju and colleagues [16] are based on IRT (see also Flowers et al. [17]). These measures are compensatory and non-compensatory DIF, or CDIF and NCDIF. NCDIF is more like indices of DIF such as the area statistics and Lord's chi-square; the assumption is that all other items in the test are unbiased, except for the studied item. CDIF does not make this assumption. The advantage of these measures over Lord's chi-square and the area statistics, such as Raju's signed and unsigned area statistic is that they are based on the actual distribution of the ability estimates within the group for which it is desired to estimate bias, rather than the entire theoretical range of theta. If, for instance, most members of the focal group fall within the range of theta from –1 to 0, rather than between –1 to +1 on the continuum, the area statistics will give an inaccurate estimate of DIF.

NCDIF is computed exactly like the unsigned probability difference of Camilli and Shepard [15]. For each subject in the focal group, two estimated scores are computed. One is based on the subject's ability estimate and the estimated a, b and c parameters for the focal group, and the other based on the ability estimate and the estimated a, b and c parameters for the reference group. Each subject's difference score (d) is squared, and these squared difference scores are added for all subjects (j = 1, n) to obtain NCDIF.

$$NCDIF_i = [\sum_{j=1,n}(ES_{siF} - ES_{siR})^2]$$

As an example, NCDIF for item i is the average difference squared between the true or expected scores for an individual (s) as a member of the focal group (F) and as a member of the reference group (R). Using the example shown above for a person at theta=-1.0,

$$d = (.6905 - .4487)^2 = .0585$$

This quantity is then summed across people and averaged $\sum d$ / n. This value is the NCDIF, which (as mentioned above) is also the unsigned probability difference (UPD) illustrated by Camilli and Shepard [15]. NCDIF is the average difference between the true (expected) scores for groups, and provides a measure of DIF magnitude. New methods for determining NCDIF cutoffs for binary items have recently been described [44].

References

1. Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R., & Teresi, J. (2007). A comparison of two sets of criteria for determining the presence of differential item functioning using ordinal logistic regression (this issue).
2. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.
3. Zumbo, B. D. (1999) A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type(ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.
4. Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256.
5. Teresi, J. A., Stewart, A. L., Morales, L., & Stahl, S. (2006). Measurement in a multi-ethnic society. *Special Issue of Medical Care, 44*(Suppl. 3), S1–S210.
6. Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*, S152–S170.

7. Cole, S. R., Kawachi, I., Maller, S. J., Munoz, R. F., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: Experiences from the New Haven EPESE study. *Journal of Clinical Epidemiology, 53*, 285–289.
8. Gallo, J. J., Cooper-Patrick, L., & Lesikar, S. (1998). Depressive symptoms of Whites and African-Americans aged 60 years and older. *Journal of Gerontology, 53B*, 277–285.
9. Mui, A. C., Burnette, D., & Chen, L. M. (2001). Cross-cultural assessment of geriatric depression: A review of the CES-D and GDS. *Journal of Mental Health and Aging, 7*, 137–164.
10. Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: True differences or differential item functioning? *Medical Care, 41*(Suppl), 75–86.
11. Gelin, M. N., Carleton, B. C., Smith, M. A., & Zumbo, B. D. (2004). The dimensionality and gender differential item functioning of the Mini-Asthma Quality-of-Life Questionnaire (MINIAQLQ). *Social Indicators Research Dordrecht, 68*, 81.
12. Roorda, L. D., Roebroeck, M. E., van Tilburg, T., Lankhorst, G. J., Bouter L. M., Measuring Mobility Study Group (2004). Measuring activity limitations in climbing stairs: Development of a hierarchical scale for patients with lower-extremity disorders living at home. *Archives of Physical Medicine and Rehabilitation, 85*, 967–971.
13. Thissen, D. (1991). *MULTILOG™ User's Guide. Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software Inc.
14. Thissen, D. (2001). IRTLRDIF v2.0b; Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Available on Dave Thissen's web page.
15. Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications: Thousand Oaks, California.
16. Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368.
17. Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous DFITframework. *Applied Psychological Measurement, 23*, 309–326.
18. Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates: Hillsdale, NJ.
19. Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the MMSE: An application of the Mantel Haenszel and Standardization procedures. *Medical Care, 44* (Suppl. 3), S107–S114.
20. Simpson, E. H. (1951). The interpretation of interaction contingency tables. *Journal of the Royal Statistical Society (Series B), 13*, 238–241.
21. Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.
22. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications Inc.
23. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum.
24. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley Publishing Co.
25. Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika Monograph Supplement 17: Richmond, VA: William Byrd Press.
26. Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3–62.
27. Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*, 42–69.
28. Benjamini, Y., & 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.
29. Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*, 332–342.
30. Thissen, D., Steinberg, L., & 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.
31. Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-mental status examination. *Medical Care, 44*, S134–S142.
32. Wang, W. C., Yeh, Y. L., & Yi, C. (2003). Effects of anchor item methods on differential item functioning detection with likelihood ratio test. *Applied Psychological Measurement, 27*, 479–498.
33. Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish Translation of the PTSD Checklist: Detection and evaluation of impact. *Psychological Assessment, 14*, 50–59.
34. Teresi, J., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651–1683.
35. Chang, H. -H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 39*, 391–404.
36. Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. [dissertation] Illinois Institute of Technology. Dissertation Abstracts International 54-04B, 2266.
37. Flowers, C. P., Oshima, T. C., & Raju, N. S. (1995). A Monte Carlo assessment of DFIT with dichotomously-scored unidimensional tests. [dissertation] Atlanta, GA: Georgia State University.
38. Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Lawrence Erlbaum Inc.
39. Raju, N. S. (1999). DFITP5: A Fortran program for calculating dichotomous DIF/DTF [computer program]. Chicago: Illinois Institute of Technology.
40. Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential item functioning in a satisfaction scale. *Journal of Applied Psychology, 85*, 451–461.
41. Morales, L. S., Flowers, C., Gutiérrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental Status Exam assessed using the DFIT methodology. *Medical Care, 44*, S143–S151.
42. Baker, F. B. (1995). *EQUATE 2.1: Computer program for equating two metrics in item response theory [Computer program]*. Madison: University of Wisconsin, Laboratory of Experimental Design.
43. Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335–350.
44. Oshima, T.C., Raju, N.S., Nanda, A.O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1–17.