Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques

DIFdetect and difwithpar

Paul K. Crane, MD, MPH, Laura E. Gibbons, PhD,* Lance Jolley, MS,† and Gerald van Belle, PhD‡*

Introduction: We present an ordinal logistic regression model for identification of items with differential item functioning (DIF) and apply this model to a Mini-Mental State Examination (MMSE) dataset. We employ item response theory ability estimation in our models. Three nested ordinal logistic regression models are applied to each item. Model testing begins with examination of the statistical significance of the interaction term between ability and the group indicator, consistent with nonuniform DIF. Then we turn our attention to the coefficient of the ability term in models with and without the group term. If including the group term has a marked effect on that coefficient, we declare that it has uniform DIF. We examined DIF related to language of test administration in addition to self-reported race, Hispanic ethnicity, age, years of education, and sex.

Methods: We used PARSCALE for IRT analyses and STATA for ordinal logistic regression approaches. We used an iterative technique for adjusting IRT ability estimates on the basis of DIF findings.

Results: Five items were found to have DIF related to language. These same items also had DIF related to other covariates.

Discussion: The ordinal logistic regression approach to DIF detection, when combined with IRT ability estimates, provides a reasonable alternative for DIF detection. There appear to be several items with significant DIF related to language of test administration in the MMSE. More attention needs to be paid to the specific criteria used to determine whether an item has DIF, not just the technique used to identify DIF.

Key Words: differential item functioning, test bias, ordinal logistic regression, item response theory, cognitive functioning

(*Med Care* 2006;44: S115–S123)

From the Departments of *Internal Medicine, ‡Biostatistics, and Environmental Health, University of Washington, Seattle; and †Georgetown Law Center, Washington, DC.

Dr. Crane was supported by grant number NIH K08 AG022232-01 from the National Institute on Aging. Dr. Gibbons was supported by grant number NIH K08 AG022232-01 and NIH 5 P50 AG05136, both from the National Institute on Aging. Dr. van Belle and Lance Jolley were supported by grants NIH U01 AG16976 and NIH 5 P50 AG05136, both from the National Institute on Aging.

Reprints: Paul K. Crane, Box 359780, Harborview Medical Center, University of Washington, 325 Ninth Avenue, Seattle, WA 98104. E-mail: pcrane@u.washington.edu.

ISSN: 0025-7079/06/4400-0115

One approach to the detection of differential item functioning (DIF) is to use logistic regression (LR)-based techniques. This article presents an ordinal logistic model for the identification of items that exhibit DIF and demonstrates that technique on the common Mini-Mental State Examination (MMSE) dataset.

CONCEPTUAL DEFINITION

DIF can be defined as interference by some demographic characteristic or grouping of the tight relationship between trait level (in this case, the level of true cognitive functioning, referred to as "ability" from here on) and item responses (in this case, responses to individual MMSE items). DIF is divided into uniform and nonuniform DIF. Nonuniform DIF is conceptualized as a statistically significant interaction between the trait level and the demographic variable. Nonuniform DIF is analogous to effect modification. Uniform DIF is conceptualized as a marked difference in the strength of the relationship between ability and item responses in models with and without the demographic variable in question. Uniform DIF is analogous to confounding. These relationships were discussed in our first publication on our technique for detecting DIF.[1] The specific operational definitions used in our work are discussed below.

REVIEW OF APPROACHES RELATED TO THIS METHOD AND FORMAL DEFINITION OF OUR ORDINAL LR PROCEDURE FOR DETECTING DIF

Mantel-Haenszel approaches to analyzing 2 × 2 contingency tables used for the assessment of DIF led quite naturally to the use of LR based approaches. In the 1980s, Mellenbergh[2] first defined the concept of nonuniform DIF as a significant interaction term in a LR-based framework[3]; this framework for DIF detection was promulgated by Swaminathan and Rogers in 1990. Simulation studies have found that LR methods were superior to Mantel-Haenszel in identifying items that had nonuniform DIF.[2,4,5]

Our group has extended this work by focusing attention on the specific criteria used for the identification of both uniform and nonuniform DIF. Swaminathan and Rogers combined the statistical test of uniform and nonuniform DIF detection into a single step.[2] Swaminathan and Rogers fit 3 logistic models to the data, and compared the difference in the -2 log likelihoods of the first and third models to a χ^2

distribution with 2 degrees of freedom. The first model included terms for the ability of each respondent:

$$\text{Logit p (item response is correct)} = \beta_0 + \beta_1 \text{ (ability)} \quad (1)$$

They then added a term for the group assignment:

$$\text{Logit p (item response is correct)} = \beta_0 + \beta_1 \text{ (ability)} + \beta_2 \text{ (group)} \quad (2)$$

Finally, they added a term for the group-by-ability interaction:

$$\text{Logit p (item response is correct)} = \beta_0 + \beta_1 \text{ (ability)} + \beta_2 \text{ (group)} + \beta_3 \text{ (group} \times \text{ability)} \quad (3)$$

This model was compared with the model that included only the ability term (Model 1). The difference in the -2 log likelihoods of these 2 models was compared with a χ^2 distribution with 2 degrees of freedom. If the χ^2 (2 *df*) statistic was statistically significant at the $\alpha = 0.05$ level, that item was flagged as exhibiting DIF.

Various efforts were made to extend LR techniques to polytomous items.[6–8] These culminated in 1999 in Zumbo's work, which extends the Swaminathan and Rogers framework to polytomous items with an ordinal LR framework. Zumbo[9] recommended both a one-stage and a two-stage analysis, examining items for uniform and nonuniform DIF in 2 separate steps.

In ordinal LR, an underlying score is estimated as a linear function of the independent variables and a set of cut points. The probability of observing outcome *i* of *I* outcome possibilities corresponds to the probability that the estimated linear function, plus random error, is within the range of the cut-points estimated for the outcome

$$\Pr(\text{outcome}_j = i) = \Pr(\kappa_{i-1} < \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + u_j \leq \kappa_i) \quad (4)$$

In this formulation, u_j is assumed to be logistically distributed. The coefficients $\beta_1, \beta_2, \ldots, \beta_k$ together with the cut-points $\kappa_1, \kappa_2, \ldots, \kappa_{I-1}$, are estimated, where I is the number of possible outcomes. κ_0 is taken as $-\infty$, and κ_I is taken as $+\infty$. All of this is a direct generalization of the ordinary 2-outcome logit model.[10] For any particular item, the number of cutpoints κ (one less than the number of outcome possibilities I) is constant from model to model, so only one degree of freedom is required to move from the ordinal analogue to Model 1 to the ordinal analogue of Model 2, and an additional degree of freedom is lost in the move from the ordinal analogue of Model 2 to the ordinal analogue of Model 3.

Our approach to DIF detection is a further modification of Zumbo's techniques. We use different criteria for nonuniform and especially for uniform DIF, and we incorporate IRT-derived ability estimates rather than the observed sum score as the estimate of ability.

Our approach begins with an evaluation of nonuniform DIF for each item. Dichotomous data are first fit with Models 3 and 2 above (and polytomous data is fit with the analogous ordinal response models). The difference in the -2 log likelihoods of these models is compared with a χ^2 (1 *df*) distribution. Because in a previous study a majority of items had nonuniform DIF by this criterion,[1] ultimately it was decided to adjust the α level for multiple comparisons. Little attention has been paid to the issue of multiple hypothesis testing in DIF detection procedures. We discussed this in a previous publication,[1] and further discussion is available as the June 2002 Rule of the Month on the Statistical Rules of Thumb website (www.vanbelle.org). In the present work, we have adjusted α using the Bonferroni technique. Since there are 21 items in the MMSE, we used an α level of $0.05/21 = 0.0024$ as our indicator of statistical significance for the presence of nonuniform DIF. In previous work, we adjusted using a number of procedures (Bonferroni, Holm, Hochberg, and Šidák),[11,12] and the same items were flagged with DIF regardless of technique. We have thus used the Bonferroni technique in our subsequent work, as it is the simplest to explain.

After the evaluation of nonuniform DIF is an examination of uniform DIF. A similarity is noted in the language describing uniform DIF and the language describing confounding relationships in epidemiology. The authors of a simulation study of empirical confounder selection studies[13] did not recommend using the 0.05 statistical significance level to examine the β_2 term from Model 2 because that criterion led to a large proportion of true confounding relationships that were not detected. They recommended 2 other criteria as superior: the statistical significance of the β_2 term at the $\alpha = 0.20$ level, or, alternatively, a criterion that examines the change in point estimates for the β_1 parameter from models with and without the β_2 term.[13] In our approach, the proportional change of the β_1 parameter from Model 1 to Model 2 is compared. For ease of understanding, and so that the β coefficients from Models 1 and 2 can be distinguished, we reintroduce Model 1 with an asterisk after the β_1 term

$$\text{Logit p (item response is correct)} = \beta_0 + \beta_{1*} \text{ (ability)} \quad (1^*)$$

A 10% change in β_1 was chosen as the criterion for the presence of uniform DIF. Specifically, if the value of the following:

$$|(\beta_1 - \beta_{1*})/\beta_{1*}| \quad (5)$$

was greater than 0.10 for a particular item, it was flagged as having uniform DIF. Thus, in our procedure, the question is "does including the group term markedly impact the relationship between the ability and item response?" The approach used by Swaminathan, Rogers, and Zumbo asks instead, "when we control for ability, is there a statistically significant relationship between the group term and item response?"

The extension of these procedures from the dichotomous to the polytomous case was accomplished using the ordinal LR formulation suggested by Zumbo,[9] but with our new criteria for uniform and nonuniform DIF.

STATA code was developed for implementing this technique for DIF detection. This freeware program is available from the National Alzheimer's Coordinating Center (NACC) website (www.alz.washington.edu/DIFDETECT/welcome.html).

CHOICE OF ABILITY ESTIMATE FOR DIF DETECTION

A major disadvantage of LR-based DIF detection as described by Swaminathan and Rogers[3] was the reliance on the sum score as the ability estimate. An important review of DIF detection procedures was somewhat dismissive of LR-based DIF detection techniques because previous investigators using these techniques used standard sum scores as the ability estimates in their assessments of DIF.[14] As Millsap and Everson[14] point out, unless rather unlikely statistical properties hold (ie, unless all of the rather restrictive assumptions of the Rasch model are satisfied), the sum score is not a very good ability estimate. Millsap and Everson emphasized the importance of better ability measures, and advocate what they term unobserved conditional invariance models of measurement bias such as item response theory (IRT) approaches.[14]

As noted by Camilli and Shepard,[15] however, LR-based approaches are not tied into any particular ability estimate. Thus, ability estimates external to the test in question could be used, or ability estimates from some alternative method of scoring the test in question (including IRT-based ability estimates). It can be shown that LR is a reparameterization of the 2PL model, and that ordinal LR is a reparameterization of Samejima's graded response model.[16,17] If IRT-based ability estimates are used in ordinal LR, and a dichotomous or categorical demographic characteristic is being examined for DIF, IRT and ordinal LR are nearly equivalent procedures.

APPROACHES TO DEALING WITH DETECTED DIF

Several authors recommend that finding DIF in an item should prompt careful consideration by content experts as to the cause of the DIF, and whether the item should be retained.[15,18] For example, Linn states "the burden should be on those who want to retain an item with high DIF to provide a justification in terms of the intended purposes of the test" (p. 353).[18] This view posits that unless there are compelling reasons to retain the item with DIF, it should be removed from the test.

A related issue is that of false-negative or false-positive findings of DIF because of DIF in other items. This occurs because DIF may lead to systematic biases in the ability estimates for some individuals. DIF findings for any particular item depend to a great extent on the choice of ability metric, and whether that ability metric is influenced by DIF in other items.[15,19]

We have developed a procedure for iterative DIF detection and updated ability estimation, retaining items that have DIF and determining whether false-positive or false-negative identification of DIF causes initial findings. That technique is based on the insight of Reise and colleagues,[20] who pointed out that although items with DIF measure differently in different groups, they are still in fact measuring the same underlying construct. We sought to develop appropriate scoring techniques for individuals in both groups that accounts for DIF when it is present.

IRT ability estimates from an external program (in our case, PARSCALE) are used, and the initial run through DIFdetect is as described above. Findings from the initial DIF detection are used to obtain new IRT ability estimates. The data are handled as shown in Table 1. Items free of DIF have parameters estimated from the entire sample. Items with DIF have parameters estimated separately in the 2 groups, resulting in 2 sets of parameters for each item with DIF. The common DIF-free items thus serve to anchor the metric for both groups, and the resulting ability estimates take into account the different relationships in the 2 groups between item responses for those items that have DIF and the underlying ability measured by the test.

To address the issue of false-positive or false-negative DIF, we then rerun the DIFdetect program using this updated ability estimate. If the items found with DIF are different between the 2 runs, we ascribe those differences to false negative or false positive DIF. In that case we adjust the IRT ability estimation procedure again, by reassigning items as having or not having DIF, as found in the most recent run of DIFdetect, and obtaining new adjusted IRT ability estimates. We repeat these processes until items found with DIF are the same in 2 successive runs.

We have written STATA code to automate several steps of this process, including writing PARSCALE code and datasets and importing updated ability estimates back into STATA. This code, called "difwithpar", is available from the Statistical Components Archive at Boston College (http://ideas.repec.org/s/boc/bocode.html). Type "ssc install difwithpar" from the STATA prompt to obtain the program files.

METHODS

Data Set

The MMSE dataset used for these and several other analyses prepared for this special issue is described elsewhere in this journal.

TABLE 1. Data Handling for Updated IRT Ability Estimation in PARSCALE When DIF Has Been Found in Some Items*

	Group 1	Group 2
Item without DIF 1	Present	Present
Item without DIF 2	Present	Present
. . .	Present	Present
Item without DIF n	Present	Present
Item with DIF n + 1	Present	Missing
Item with DIF n + 2	Present	Missing
. . .	Present	Missing
Item with DIF n + m	Present	Missing
Item with DIF n + 1	Missing	Present
Item with DIF n + 2	Missing	Present
. . .	Missing	Present
Item with DIF n + m	Missing	Present

The *n* items found without DIF are treated as present in both groups, and item parameters are derived using data from everyone in the data set. The *m* items detected with DIF are treated differently in the 2 groups. Group-specific item parameters are estimated for the *m* items with DIF by replicating the set of *m* items and treating each set as present in one group and missing in the other. The *n* items without DIF serve as an empirically determined DIF-free core for the purposes of anchoring the metric.

DIF Detection Software

DIFdetect was used for all analyses using STATA version 7 (College Station, TX: StataCorp, 2001). DIFdetect was downloaded from the National Alzheimer Coordinating Center website July 2003.

Specific Criteria for Nonuniform DIF

The -2 log likelihood of models with and without the interaction term (Models 2 and 3 above) were compared with a χ^2 (1 *df*) distribution with an α value of $(0.05/21) = 0.0024$. We show the *P* value for each interaction term.

Specific Criteria for Uniform DIF

The relative differences between the point estimates of the coefficients of the ability terms from Models 2 and 1* were compared with 10%, using the equation $(\beta_1 - \beta_1{}^*)/\beta_1$. For the main analysis of language of test administration, we show both the *P* value comparing Models 1 and 2 as well as the relative change in the β_1 coefficient.

Initial IRT Ability Estimates Used for DIF Detection

IRT ability estimates derived from PARSCALE version 3.0 (Scientific Software International, Lincolnwood, IL, 1997) were used. PARSCALE code used for these analyses is available from the author. We employed Samejima's graded response model[16,17] and expected a posteriori (EAP) scoring.

Iterative DIF Analysis

As described above, we updated IRT ability estimates based on DIFdetect findings. We evaluated false-positive and false-negative findings of DIF by repeating the ordinal LR analyses using the updated IRT ability estimates. We repeated these procedures until items identified with DIF were the same on successive runs. Results presented in the tables are from the final ordinal LR run with the updated IRT estimates.

Demographic Characteristics Evaluated for DIF

The primary analysis reported here is the analysis of DIF in this dataset with respect to language of test administration. Items were also investigated for DIF with respect to 5 other demographic variables: self-reported race (where the options were white, black, Asian, or Hispanic; analyzed as categorical comparisons of whites to blacks, whites to Hispanics, and blacks to Hispanics; the 7 Asian subjects were excluded from these categorical analyses); Hispanic ethnicity regardless of racial self-identification (an indicator variable); age (dichotomized at age 80); years of formal schooling (dichotomized at 8 years), and gender. Each demographic characteristic is treated independently; more complicated models incorporating multiple demographic characteristics simultaneously were not examined (see Crane et al[21] for an example of how one might do this).

Testing Assumptions of the OLR Models

For the principal analysis of language, we tested model fit for the 16 dichotomous items using Hosmer-Lemeshow statistics.[22] We were testing 3 models for each of the 16 items, so we used an α level of $0.05/48 = 0.0010$. We assessed the proportional odds assumption in the 5 polytomous items using Brant tests.[23]

Comparison of Findings With Findings of Standard MMSE Sum Scoring

For the principal analysis of language, we compare the findings of our iterative DIF detection and IRT ability estimation procedure with the findings of ordinal LR DIF detection using the standard MMSE sum score, including both the "serial 7 subtractions" and "WORLD" spelled backwards items. For analyses with the standard MMSE sum score, the item being assessed for DIF was not subtracted from the score (though this is an option available in DIFdetect).

RESULTS

DIF Related to Language

Results for language (Spanish vs. English) are shown in Table 2 for the IRT iterative DIF detection algorithm and the ordinal LR approach using the standard MMSE score. Three orientation items (identification of the season, city, and state), and repeating a phrase were found to have uniform DIF. Recall of 3 objects was found to have nonuniform DIF. Findings using the same criteria but using standard sum scoring were almost identical.

Comparison of DIF Detection Criteria

The differences between the criteria we have used and the traditional ordinal LR criteria[9] also can be seen in Table 2. For nonuniform DIF, the only difference between our criteria and the traditional ordinal LR criteria is that we have used Bonferroni adjustment, resulting in a more stringent criterion for declaring an item to have DIF. Without this adjustment, 4 additional items would be declared to have nonuniform DIF, including 2 items that also had uniform DIF (correct identification of the state and repeating a phrase) and 2 additional items that didn't have uniform DIF (recall of apple, table, and penny, and naming a pencil). The difference between the uniform DIF criteria is much greater between our technique and traditional ordinal LR criteria: 14 of the items have a *P* value of <0.05 when comparing Models 1 and 2, while only 4 of these had a relative difference of the β_1 term greater than 10%. The next largest relative change in the β_1 term was for the item that assesses the subject's following a written command to close their eyes, with a 2.9% difference. Thus the criterion we use is much more stringent for uniform DIF detection than the traditional ordinal LR criterion.

To demonstrate the difference between our criteria and the traditional criteria for ordinal LR approaches, we performed an additional PARSCALE analysis. For the purposes of this illustration, we declared 6 items to have DIF: the 5 shown in Table 2 as well as the "close your eyes" item. We prepared a data set for PARSCALE analysis that freed parameters for those 6 items to be estimated separately in different language groups, and fixed the parameters from the other 15 items to be estimated from all of the subjects. We then plotted the item characteristic curves for the item with the lowest amount of uniform DIF that we found to be of clinical importance (repeating a phrase, which had a relative difference of 11.5% and a *P* value of <0.001) and the "close your eyes" item (which had a relative difference of 2.9% and a *P* value of <0.001). The item characteristic curves for

TABLE 2. Differential Item Functioning With Respect to Language In MMSE Items*

	Iterative DIF Detection and IRT Ability Estimation Technique			Standard MMSE Sum Score		
	Uniform		Non-uniform	Uniform		Nonuniform
Item	Percent Change in β_1	*P* Value, Model 1 vs. Model 2	*P* Value, Model 2 vs. Model 3	Percent Change in β_1	*P* Value, Model 1 vs. Model 2	*P* Value, Model 2 vs. Model 3
Year	0.5	0.223	0.143	0.5	0.215	0.111
Season	**13.1**	<0.001	0.067	**12.7**	<0.001	0.297
Day of month	1.4	<0.001	0.070	0.7	<0.001	0.050
Day of week	0.0	0.972	0.396	0.1	0.869	0.900
Month	0.9	0.059	0.073	0.9	0.061	0.108
State	**24.1**	<0.001	0.007	**25.7**	<0.001	0.074
City	**12.3**	<0.001	0.385	**12.0**	<0.001	0.304
Two nearby streets	1.9	0.003	0.189	2.0	0.003	0.178
Floor	0.1	0.464	0.630	0.1	0.555	0.547
Ident type place	0.3	0.019	0.511	0.0	0.030	0.750
Apple table penny	$0.4^{\dagger}$	0.118	0.008	$0.3^{\dagger}$	0.143	0.009
Subtractions	$0.1^{\dagger}$	0.687	0.134	$0.2^{\dagger}$	0.265	0.353
World backwards	$1.3^{\dagger}$	<0.001	0.610	1.3	<0.001	0.775
Recall of 3 objects	0.3	<0.001	**<0.001**	0.3	<0.001	**<0.001**
Pencil	0.8	0.004	0.032	$0.9^{\ddagger}$	0.004	0.022
Wristwatch	1.3	0.003	0.358	1.6	0.003	0.279
Repeat a phrase	**11.5**	<0.001	0.002	**12.3**	<0.001	0.037
Close your eyes	2.9	<0.001	0.354	3.5	<0.001	0.337
Three-part command	0.9	<0.001	0.897	$0.6^{\dagger}$	<0.001	0.870
Writing sentence	$1.7^{\dagger\dagger}$	0.002	0.335	$2.1^{\ddagger}$	0.001	0.398
Copying design	$0.1^{\dagger\dagger}$	0.730	0.882	$0.2^{\ddagger}$	0.549	0.348

*Results are in bold if the percent change in β_1 was >10% for uniform DIF, or the *P* value for the language-ability interaction is <0.0024 for nonuniform DIF.

†Reject proportional odds assumptions with Brant test. Dichotomous findings still not significant.

‡Reject goodness-of-fit with Hosmer-Lemeshow test. Findings still not significant with outliers omitted.

English and Spanish speakers for "Repeat a phrase" are shown in Figure 1. The slopes of the curves are not the same (consistent with the finding that the *P* value for the interaction term is almost as small as 0.0024, the criterion for identifying nonuniform DIF), and the difficulties of the items are very different in the 2 different language groups. The item characteristic curves for English and Spanish speakers for the "Close your eyes" item are shown in Figure 2. The slopes of these 2 curves are the same, and the difficulties are very close to each other. Despite the *P* value of <0.001 for the difference between models 1 and 2, corresponding to the statistical significance of β_2 in model 2, the curves shown in Figure 2 are negligibly different from each other. Curves from the other items with statistically significant *P* values but smaller

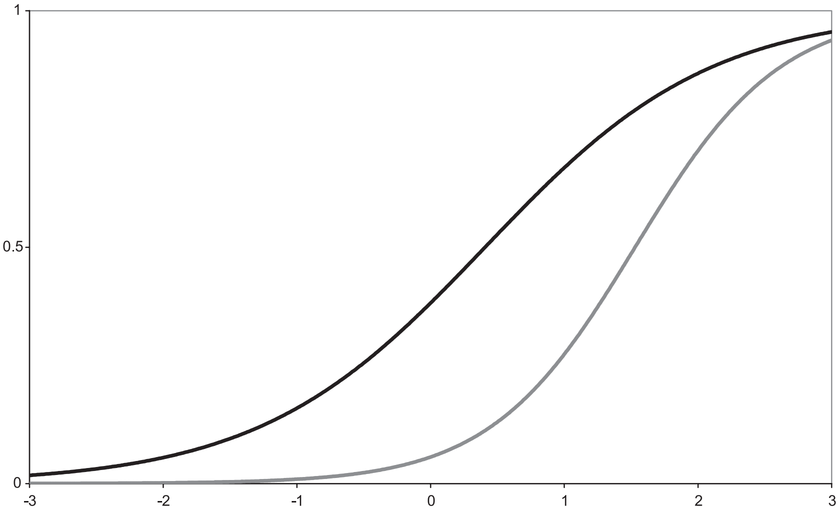

FIGURE 1. Shown are item characteristic curves for English (black line) and Spanish (gray line) speakers for the item "Repeat a phrase." The item characteristic curve depicts the probability of success on the item for individuals with a given ability level, here shown ranging from –3 to +3. Shown is an item depicting both uniform and nonuniform DIF: the 2 curves are far apart and are not parallel to each other. See text for further details.

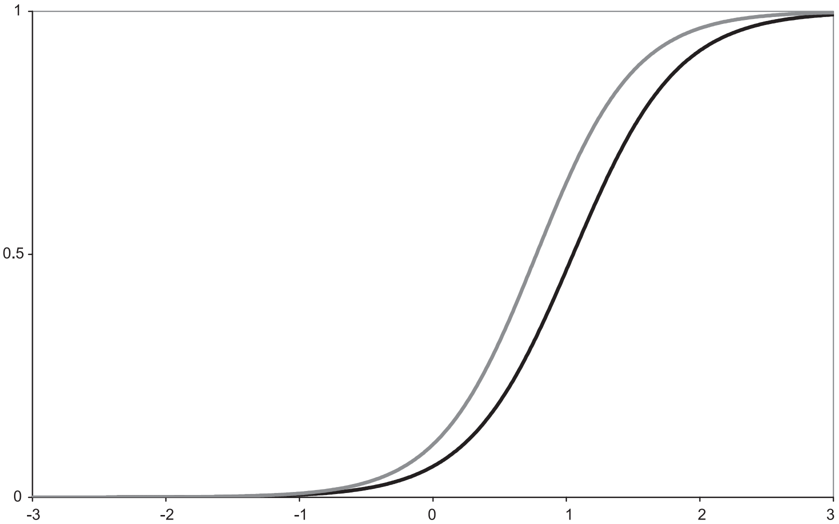

FIGURE 2. Shown are item characteristic curves for English (black line) and Spanish (gray line) speakers for the item "Close your eyes." The item characteristic curve depicts the probability of success on the item for individuals with a given ability level, here shown ranging from –3 to +3. Shown is an item that showed uniform DIF using a statistical significance criterion but not using the change in β_1 criterion. The 2 curves are much closer together and are nearly parallel to each other, as expected because no nonuniform DIF was found. Item characteristic curves from items that showed uniform DIF using the statistical significance criterion and even smaller changes in β_1 coefficients between modes will have curves for the 2 groups that are even closer together; the item depicted here had the largest change in β_1 coefficients that did not meet the cutoff value of 10%. See text for further details.

relative changes in the β_1 coefficients would have curves even closer to each other than the "close your eyes" item as indicated by even smaller relative differences.

Impact of Model Fit on Results

There are 5 polytomous MMSE items: subtraction of serial 7s, spelling the word "world" backwards, following a 3-step command, repeating 3 words, and short-term recall of those 3 words. We performed Brant tests for DIF analyses of these items. As shown in Table 2, 3 of the items had significant Brant tests when using IRT scoring. None of the items with significant Brant tests were found to have DIF. The other 16 items in the MMSE are dichotomous. As shown in Table 2, 3 of these items had significant violations of goodness of fit using Hosmer-Lemeshow criteria. We identified outliers and removed them, and results were broadly similar, with none of the items identified with either uniform or nonuniform DIF.

Comparison of DIF Findings Using IRT and Traditional Sum Score Ability Estimates

The final IRT score and the standard sum score for the MMSE were closely related to each other, with a correlation of 0.94. As shown in Table 2, findings when using IRT-derived ability estimates were broadly similar.

DIF Findings Related to Other Covariates

Findings with respect to other demographic characteristics are summarized in Table 3. Analyses of each characteristic were run separately, with iterative updating of ability scores until the same results were obtained on 2 successive runs. Orientation to the season of the year was found to have uniform DIF related to the comparison of blacks to Hispanics, and also related to Hispanic ethnicity. Correct identification of the state was found to have both uniform and nonuniform DIF related to the comparisons of blacks to Hispanics and whites to Hispanics, as well as related to Hispanic ethnicity. Correct identification of the city was found to have uniform DIF related to the comparison of blacks to Hispanics and related to Hispanic ethnicity. Repeating a phrase was found to have uniform DIF related to the comparison of whites to Hispanics. When we dichotomized education at less than 8 years versus 8 or more years, none of the items was found to have DIF. When we dichotomized age at younger than 80 years versus 80 years or older, none of the items was found to have DIF.

DISCUSSION

Ordinal LR approaches to analyzing test items for the presence of DIF are illustrated herein. The primary analysis of DIF with respect to language of test administration was easily accomplished. Items identified as exhibiting uniform DIF with respect to language of test administration included correct identification of the season, the state, the city, and repeating a phrase. The ability to recall 3 objects showed nonuniform DIF with both IRT and standard MMSE sum score ability estimates.

In addition to the primary analysis, we illustrate a further capability of the DIFdetect approach to analyzing test items for the presence of DIF. Because of the speed with which the package works, and the flexibility of the ordinal LR framework, many other demographic characteristics can also be evaluated to determine whether items may exhibit DIF.

Several previous analyses have examined the MMSE for DIF with respect to a number of different demographic characteristics. One of the powerful advantages of DIFdetect is the ability to quickly assess items for DIF with respect to any particular covariate, facilitating assessment in turn for a large number of covariates in the same amount of time required to perform a single analysis using some other techniques. Thus, in their article, using LR-based DIF detection on the MMSE, Marshall and colleagues looked only at language of test administration and Hispanic ethnicity; no assessments were published on DIF with respect to years of schooling, age, or gender.[24] In a previous study,[25] writing a sentence, repeating a phrase, repeating names of objects, closing eyes, folding a paper, and calculating serial 7s were found to show either evidence of poor discrimination or DIF with respect to education. Both that study and the present

TABLE 3. Presence of Differential Item Functioning Related to Several Covariates in MMSE Items as Assessed by DIFdetect*

	Black-White		Black-Hispanic		White-Hispanic		Hispanic Ethnicity		Education (>8 Yr)		Age (80+ Yr)	
Item	**Uniform**	**Nonuniform**	**Uniform**	**Nonuniform**	**Uniform**	**Nonuniform**	**Uniform**	**Nonuniform**	**Uniform**	**Nonuniform**	**Uniform**	**Nonuniform**
Year	1.0	0.881	2.1	0.524	0.1	0.527	1.6	0.774	0.2	0.543	1.3	0.003
Season	0.1	0.410	**17.7**	0.068	8.4	0.845	**16.2**	0.151	0.6	0.344	7.1	0.945
Day of month	2.4	0.207	0.4	0.433	4.8	0.135	0.2	0.140	6.8	0.104	0.6	0.003
Day of week	0.3	0.263	0.1	0.212	0.1	0.511	0.2	0.686	7.0	0.766	2.1	0.147
Month	0.1	0.749	2.0	0.209	0.1	0.155	1.1	0.048	1.3	0.096	0.3	0.233
State	0.3	0.798	**26.6**	**0.002**	**12.4**	**<0.001**	**27.0**	**<0.001**	4.4	0.256	6.2	0.692
City	1.1	0.285	**12.5**	0.029	8.6	0.251	**13.3**	0.176	0.4	0.174	6.4	0.956
Two nearby streets	0.0	0.215	3.2	0.459	0.3	0.040	2.8	0.162	0.2	0.546	1.9	0.043
Floor	0.3	0.456	0.1	0.497	0.2	0.794	0.0	0.562	4.3	0.533	2.9	0.182
Ident type place	0.1	0.287	0.1	0.373	1.8	0.189	0.2	0.189	1.2	0.986	5.0	0.850
Apple,table, penny	0.5	0.900	0.1	0.025	0.1	0.012	0.3	0.005	3.6	0.037	0.4	0.094
Subtractions	2.1	0.720	0.6	0.711	0.0	0.008	0.6	0.061	2.0	0.252	5.1	0.098
World backwards	0.2	0.421	1.5	0.701	1.3	0.142	2.4	0.624	2.2	0.081	3.9	0.023
Recall of 3 objects	1.1	0.397	1.5	0.033	1.4	0.073	0.4	0.009	1.9	0.928	3.7	0.699
Pencil	0.3	0.101	1.5	0.006	1.1	0.298	0.8	0.015	0.4	0.734	3.7	0.552
Wristwatch	0.8	0.198	0.2	0.884	3.8	0.283	0.8	0.573	1.3	0.974	6.0	0.905
Repeat a phrase	1.9	0.625	6.6	0.170	**14.3**	0.172	7.0	0.072	5.9	0.564	8.3	0.020
Close your eyes	2.8	0.967	2.8	0.947	6.8	0.924	4.2	0.573	0.2	0.934	3.3	0.355
Three-part command	0.2	0.446	0.5	0.731	1.7	0.787	0.5	0.846	6.0	0.110	2.8	0.988
Writing sentence	0.0	0.266	3.9	0.102	0.6	0.809	3.7	0.196	0.6	0.398	2.1	0.764
Copying design	0.5	0.956	0.7	0.723	0.2	0.601	0.5	0.773	2.2	0.750	0.4	0.982

*Items with uniform DIF (change in β_1 coefficient between Models 1 and 2 of at least 10%) and nonuniform DIF (P value comparing Models 2 and 3 <0.024) are shown in bold. Iterative item response theory ability estimation was used for all of these analyses. See text for details.

study found DIF related to repeating names of objects and repeating a phrase. Interestingly, in Teresi's earlier study, none of the orientation items were found to have DIF,[25] whereas in the present study, 3 orientation items were found to exhibit DIF. A third study found several items showed DIF with respect to education in the 5 Epidemiologic Catchment Area sites, including serial 7s, repeating a phrase, writing a sentence, naming the season, and copying a design.[26] The DIFdetect approach illustrated with this dataset, with the assessment of DIF with respect to a large number of demographic characteristics at the same time, provides a more complete picture of DIF than these prior studies.

Differences between the present findings and previous studies[24–27] may be attributable to some combination of differences between DIF detection techniques and differences in populations. The current project represents an advantage over these earlier studies because many different DIF detection techniques have been performed on the same data set, eliminating that source of variation.

However, a significant limitation of the present study is that the use of a real data set will enable only a comparison of observed findings, without knowledge of the true DIF status. Only through simulation studies will the relative merits of different DIF detection techniques be established empirically. A related and under-developed area is empiric comparisons of various criteria for DIF detection within a particular method. Should some form of α allocation be pursued in analyses of nonuniform DIF? Should significance tests alone be sufficient for uniform DIF detection? Is a 10% change in the β_1 coefficient criterion optimal for uniform DIF, or is some other value a better choice? What are the tradeoffs in terms of accuracy, false-positive findings, and false-negative findings, when different choices are made regarding specific criteria chosen to determine whether an item has DIF? Only with carefully designed simulation studies will we be able to provide insight into these issues. A number of proposals for simulated DIF analyses are in circulation.

Advantages and Disadvantages of the Ordinal LR Framework for DIF Detection

The ordinal LR framework, as we have used it, has a number of advantages over other techniques. The most important advantage is flexibility. We are able to incorporate IRT ability estimates, and to allow item parameters to be separately determined for certain items but not others. This enables us to demonstrate the consequences of declaring items to have DIF, as shown in Figures 1 and 2. Our technique empirically determines a set of items that can be considered to be DIF-free anchors, removing some of the subjectivity and labor from this process involved with some other DIF detection techniques.

A second advantage of the ordinal LR algorithm we have developed is speed. Ordinal LR is quickly accomplished, and the tools we have developed permitted the completion of all the main analyses presented here in a single day (model checking took more time). A significant limitation of the ordinal LR framework for DIF detection employing IRT ability estimates is the necessity to be familiar with both ordinal LR as well as IRT. The software tools we have developed facilitate communication between PARSCALE and STATA, which goes a long way toward addressing these issues. However, it is still necessary to have access to and familiarity with 2 software packages rather than just one.

At present, we are unaware of a way to incorporate a parameter for guessing into the ordinal LR framework. This is generally not an important problem in the setting of medical tests, where one rarely sees items with the multiple choice format that led to the need for models that could account for the possibility that those with very low ability levels had a nonzero probability of a correct response. The importance of the violation of the proportional odds assumption is unknown. Simulation studies would help to develop a sense of when violation of this assumption might matter. At present, model fit for polytomous IRT models is also an under-developed area that cries out for further theoretical and practical work. In the present analyses related to language of test administration, DIF findings did not appear to be related to violations of the proportional odds assumption or to poor model fit.

Ordinal LR techniques require sufficient sample sizes for stable estimation of model parameters. A useful rule of thumb is to have 10 observations in the smaller group for each parameter being estimated.[28] In the case of dichotomous items, each item requires estimation of 3 parameters (β_1, β_2, and β_3). In the case of polytomous items, each item also requires estimation of parameters for each of the cut points, with a requirement of one fewer cut point than there are response options. Thus, for a 6-response item (such as spelling WORLD backwards), there are $6 - 1 = 5$ cut points in addition to 3 β parameters that need to be estimated, for a total of 8 parameters. For such an item, roughly 80 observations in the smaller group would be required for stable coefficients. Of course, if responses are highly skewed, more observations would be needed to have stable coefficient estimates across all response levels.

In summary, DIFdetect was able to rapidly provide answers regarding the presence or absence of DIF in this data set. The speed of DIFdetect enabled the evaluation of all 6 demographic categories available for analysis in the dataset. DIFdetect also facilitates the use of IRT-derived ability estimates, thus avoiding limitations involved with use of traditional sum scores as the ability estimates. The validity of the specific criteria used to determine the presence or absence of uniform and nonuniform DIF in test items is unknown. More work is needed to validate the choice of specific criteria for uniform and nonuniform DIF. The ordinal LR framework provides a flexible and attractive backbone for the assessment of DIF in test items. The DIFdetect package successfully implements these procedures and is available for free download from the web.

REFERENCES

1. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. *Stat Med*. 2004;23:241–256.
2. Mellenbergh GJ. Item bias and item response theory. *Intern J Educational Res*. 1989;13:127–143.
3. Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. *J Educational Measure*. 1990;27: 361–370.

4. Rogers HJ, Swaminathan H. A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Appl Psychol Measure*. 1993;17:105–116.
5. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Appl Measure Education*. 2001;14:329–349.
6. Miller TR, Spray JA. Logistic discriminant function analysis for DIF identification of polytomously scored items. *J Educational Measure*. 1993;30:107–122.
7. French AW, Miller TR. Logistic regression and its use in detecting differential item functioning in polytomous items. *J Educational Measure*. 1996;33:315–332.
8. Welch C, Hoover HD. Procedures for extending item bias detection techniques to polytomously scored items. *Appl Measure Ed*. 1993;6:1–19.
9. Zumbo BD. *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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.
10. Stata Statistical Software [computer program]. Release 8.0. College Station, TX: Stata Corporation; 2003.
11. Hsu J. *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall; 1996.
12. Schouten HJ. Combined evidence from multiple outcomes in a clinical trial. *J Clin Epidemiol*. 2000;53:1137–44.
13. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol*. 1993;138:923–36.
14. Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. *Appl Psychol Measure*. 1993;17(4):297–334.
15. Camilli G, Shepard LA. *Methods for Identifying Biased Test Items*. Thousand Oaks: Sage; 1994.
16. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr*. 1969; No. 17.
17. Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory*. New York: Springer; 1997:85–100.
18. Linn RL. The use of differential item functioning statistics: a discussion of current practice and future implications. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993:349–366.
19. Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993.
20. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull*. 1993;114:552–566.
21. Crane PK, Hart DL, Gibbons LE, et al. A 37-item shoulder functional status item pool had negligible differential item functioning. *J Clin Epidemiol*. 2006;59:478–484.
22. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York: Wiley; 2000.
23. Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*. 1990;46:1171–1178.
24. Marshall SC, Mungas D, Weldon M, et al. Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. *Psychol Aging*. 1997;12:718–725.
25. Teresi JA, Golden RR, Cross P, et al. Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *J Clin Epidemiol*. 1995;48:473–483.
26. Jones RN, Gallo JJ. Education and sex differences in the Mini-Mental State Examination: effects of differential item functioning. *J Gerontol B Psychol Sci Soc Sci*. 2002;57B:P548–P558.
27. Jones RN, Gallo JJ. Education bias in the mini-mental state examination. *Int Psychogeriatr*. 2001;13:299–310.
28. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373–1379.