Published in final edited form as:
J Stat Softw. 2011 March 1; 39(8): 1–30.

lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations

Seung W. Choi,
Northwestern University

Laura E. Gibbons, and
University of Washington

Paul K. Crane
University of Washington

Abstract

Logistic regression provides a flexible framework for detecting various types of differential item functioning (DIF). Previous efforts extended the framework by using item response theory (IRT) based trait scores, and by employing an iterative process using group–specific item parameters to account for DIF in the trait scores, analogous to purification approaches used in other DIF detection frameworks. The current investigation advances the technique by developing a computational platform integrating both statistical and IRT procedures into a single program. Furthermore, a Monte Carlo simulation approach was incorporated to derive empirical criteria for various DIF statistics and effect size measures. For purposes of illustration, the procedure was applied to data from a questionnaire of anxiety symptoms for detecting DIF associated with age from the Patient–Reported Outcomes Measurement Information System.

Keywords

DIF; ordinal logistic regression; IRT; R

1. Introduction

Standardized tests and questionnaires are used in many settings, including education, psychology, business, and medicine. Investigations across numerous disciplines have identified respondent culture (more generally, any group membership irrelevant of the construct being measured) as a potential source of systematic measurement variability in survey research (Andersen 1967). Systematic measurement variability can lead to a number of problems including errors in hypothesis testing, flawed population forecasts, policy planning and implementation, and misguided research on disparities (Perkins *et al.* 2006). Ensuring equivalent measurement is important prior to making comparisons among individuals or groups (Gregorich 2006). Evaluations of item-level measurement equivalence have come to focus on DIF, defined as different probabilities of success or endorsement across construct-irrelevant groups when controlling for the underlying trait measured by the test (Camilli and Shepard 1994). There are many other frameworks for DIF detection, including explanatory item

Seung W. Choi, Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, 710 N. Lake Shore Dr., Chicago, Illinois 60611, United States of America, Telephone: +1/312/503-9803, s-choi@northwestern.edu, URL: http://www.mss.northwestern.edu/faculty/choi.html.

response model formulation (De Boeck and Wilson 2004), the multiple indicators multiple causes (MIMIC) formulation (Jones 2006), and the SIBTEST framework (Shealy and Stout 1993). This paper addresses the logistic regression framework, which provides a flexible model-based framework for detecting various types of DIF (Swaminathan and Rogers 1990; Zumbo 1999).

Previous efforts extended the logistic regression DIF technique into a framework known as **difwithpar** (Crane *et al.* 2006) by using IRT based trait estimates and employing an iterative process of accounting for DIF in the trait estimate with the use of group-specific IRT item parameters for items identified with DIF (Crane *et al.* 2006, 2007b,c, 2004). This framework has been found to be facile at accounting for multiple sources of DIF and paying specific attention to DIF impact. It is also able to address covariates with more than two categories, rather than limiting to only focal and reference groups.

The **difwithpar** software includes user-specified flagging criteria (or detection thresholds) for identifying items with DIF, and the developers have investigated the implications of different choices for these flagging criteria (Crane *et al.* 2007c). Several values may be used for flagging criteria in analyzing a single dataset, resulting in varying numbers of items identified with DIF, but fairly consistent DIF impact for individuals and groups across different values for the flagging criteria (Crane *et al.* 2007b). These observations suggest the need for empirical identification of flagging criteria.

To date, while the **difwithpar** software is freely distributed on the web (type `ssc install difwithpar` at the `Stata` prompt), it uses the proprietary software packages Stata (StataCorp. 2007) and **PARSCALE** (Muraki and Bock 2005). Recent developments of free IRT packages for `R` (`R` Development Core Team 2010), such as **eRm** (Mair and Hatzinger 2007) and especially the IRT/latent trait modeling package **ltm** (Rizopoulos 2006), suggested the possibility of integrating the framework in a freeware platform. The current investigation advances the **difwithpar** technique by creating a computational platform integrating both statistical and IRT procedures into a single freeware program. Furthermore, we provide a mechanism to evaluate statistical criteria proposed for detecting DIF using graphical approaches and Monte Carlo simulations. The resulting `R` package **lordif** is available from the Comprehensive `R` Archive Network at http://CRAN.R-project.org/package=lordif.

2. Logistic regression/IRT hybrid DIF detection method

2.1. Logistic regression DIF methods

Swaminathan and Rogers (1990) proposed the use of logistic regression in DIF detection for dichotomous items. Several researchers have extended the technique for polytomous items (French and Miller 1996; Miller and Spray 1993; Zumbo 1999). For polytomous items, the proportional-odds logistic regression model (Agresti 1990) is used with the assumption that the outcome variable is ordinal (as opposed to nominal). Let U_i denote a discrete random variable representing the ordered item response to item i, and u_i (= 0, 1, …, $m_i - 1$) denote the actual response to item i with m_i ordered response categories. Based on the proportional odds assumption or the parallel regression assumption, a single set of regression coefficients is estimated for all cumulative logits with varying intercepts (α_k). For each item, an intercept-only (null) model and three nested models are formed in hierarchy with additional explanatory variables as follows:

$$\begin{gathered}\text{Model 0:logit } P(u_i \geq k)=\alpha_k \\ \text{Model 1:logit } P(u_i \geq k)=\alpha_k+\beta_1 * \text{ability} \\ \text{Model 2:logit } P(u_i \geq k)=\alpha_k+\beta_1 * \text{ability}+\beta_2 * \text{group} \\ \text{Model 3:logit } P(u_i \geq k)=\alpha_k+\beta_1 * \text{ability}+\beta_2 * \text{group}+\beta_2 * \text{ability} * \text{group},\end{gathered}$$

where $P(u_i \geq k)$ denotes the cumulative probabilities that the actual item response u_i falls in category k or higher. The term "ability" is used broadly here to represent the trait measured by the test as either the observed sum score or a latent variable. Without loss of generality the term "trait level" may be substituted in each case. In the remainder of this paper, we use these terms somewhat interchangeably.

2.2. DIF detection

Testing for the presence of DIF (both uniform and non-uniform) under the logistic regression framework is traditionally based on the likelihood ratio χ^2 test (Swaminathan and Rogers 1990). DIF is classified as either uniform (if the effect is constant) or non-uniform (if the effect varies conditional on the trait level). Uniform DIF may be tested by comparing the log likelihood values for Models 1 and 2 (one degree of freedom, or $df = 1$), and non-uniform DIF by Models 2 and 3 ($df = 1$). An overall test of "total DIF effect" is tenable by comparing Models 1 and 3 ($df = 2$). The 2-*df* χ^2 test was designed to maximize the ability of this procedure to identify both uniform and non-uniform DIF and control the overall Type I error rate. However, the component uniquely attributable to either uniform or non-uniform DIF can be partitioned separately by the 1-*df* tests (Jodoin and Gierl 2001; Zumbo 1999).

The extension of this framework for multiple groups is also straightforward. The β_2 and β_3 terms from Models 2 and 3 are expanded to include binary indicator variables for all of the groups except one. For both uniform (Model 1 vs. 2) and non-uniform DIF (Model 2 vs. 3) twice the difference in log likelihoods is compared to a χ^2 distribution with degrees of freedom equal to the number of groups minus one.

2.3. DIF magnitude

Although the likelihood ratio test has been found to yield good Type I error control (Kim and Cohen 1998), some researchers have reported good power but inflated Type I error under the logistic regression likelihood ratio test (Li and Stout 1996; Rogers and Swaminathan 1993; Swaminathan and Rogers 1990). Because statistical power is dependent on sample size (Cohen 1988), a trivial but non-zero difference in population parameters will be found to be statistically significant given a large enough sample. In response to this concern, several effect size measures have been used to quantify the magnitude of DIF (Crane *et al.* 2004; Jodoin and Gierl 2001; Kim *et al.* 2007; Zumbo 1999). Zumbo (1999) suggested several pseudo R^2 statistics as magnitude measures and guidelines for classifying DIF as negligible (< 0.13), moderate (between 0.13 and 0.26), and large (> 0.26). Subsequent studies (Jodoin and Gierl 2001; Kim *et al.* 2007), however, found the proposed thresholds to be too large, resulting in under-identification of DIF. Kim *et al.* (2007) also found that the pseudo R^2 measures are closely related (with almost perfect rank order correlations) to some standardized impact indices (Wainer 1993).

Jodoin and Gierl (2001) noted that the regression coefficients β_2 and β_3 can be used as magnitude measures of uniform and non-uniform DIF, respectively. The difference in the β_1 coefficient from Models 1 and 2 has also been used to identify items with uniform DIF (Crane *et al.* 2004). Based on simulation studies in a different context (Maldonado and Greenland 1993), 10% differences in this coefficient from Models 1 and 2 was initially proposed as a practically meaningful effect (Crane *et al.* 2004). Subsequent studies used lower thresholds such as 5% and even 1% (Crane *et al.* 2007c).

2.4. Monte Carlo simulation approach to determining detection thresholds

Within the logistic regression DIF detection framework, there is considerable variability in specific criteria recommended for determining whether items exhibit DIF. Several experts have recommended magnitude measures with a plea towards "clinical relevance," though specific thresholds based on this plea are not clearly discernible. Some authors have recommended a flexible, almost analog procedure in which the threshold used for a given parameter to identify items with and without DIF is manipulated up and down, and the effect on DIF impact for individuals or groups is evaluated (Crane *et al.* 2007c,a, 2010, 2008b,a; Gibbons *et al.* 2009).

Given the variety of DIF magnitude measures and detection criteria, a Monte Carlo simulation approach may be useful. Two general approaches are feasible, one driven by Type I error and another by Type II error. The Type I error approach involves generating multiple datasets (of the same dimension as the real data) under the null hypothesis (i.e., no DIF), preserving observed group differences in ability (trait level). Various magnitude measures are computed repeatedly over the simulated datasets, from which the empirical distributions are obtained. The researcher can then be guided by these empirical distributions when making a determination with any particular magnitude measure whether items have DIF. The target threshold to use in the real data is one where the empirical probability of identifying an item as displaying DIF (i.e., false positives) is not greater than the preset nominal α level. The Type II error approach, which is not implemented in **lordif** for the reasons provided below, involves generating multiple datasets as before, preserving group differences. However, the Type II error approach also involves introducing known DIF of varying magnitude, deemed as minimally detectable (e.g., power ≥ 0.80), to a varying number of items. Again, the magnitude measures are computed over the simulated datasets and their empirical distributions are obtained. Reviewing the empirical distributions the researcher can determine a target threshold to use in the real data. Unlike in the Type I error approach, the target threshold corresponds to a typical value in the empirical distribution (e.g., the median) rather than an extreme one that cuts off the tail end of the distribution. The choice of the magnitude of DIF introduced and the specific items having DIF can affect the simulation outcome (Donoghue *et al.* 1993) and hence makes it difficult to implement the Type II error approach in a general simulation framework.

2.5. Iterative purification of the matching criterion

DIF refers to a difference in item performance between groups of respondents matched on the trait being measured. The matching criterion, the variable by which the respondents are matched, is important in order to distinguish between differences in item functioning from differences between groups (Dorans and Holland 1993). One of the potential limitations of logistic regression DIF detection was the reliance on the observed sum score as the matching criterion. As Millsap and Everson (1993) point out, the sum score is not a very good matching criterion unless statistical properties of the Rasch model hold (e.g., equal discrimination power for all items). Even if the Rasch model holds, using the sum score in a regression framework may not be ideal because the relationship between the sum score and the Rasch trait score is not linear, as evident in a test characteristic curve. In such situations, an IRT trait score is a more reasonable choice for regression modeling (such as DIF detection in the ordinal logistic regression framework) than an observed sum score (Crane *et al.* 2008a).

Another consideration for obtaining the matching criterion is related to purification. Zumbo (1999) advocated purifying the matching criterion by recalculating it using only the items that are identified as not having DIF. French and Maller (2007) reported that purification was beneficial under certain conditions, although overall power and Type I error rates did not substantially improve. Holland and Thayer (1988) suggested that the item under examination should be included in the matching criterion even if it was identified as having DIF but excluded from the criterion for all other items to reduce the Type I error rate. Zwick (1990) also proved theoretically that excluding the studied item from the matching variable leads to a bias (over detection) under the null condition.

Eliminating items found to have DIF is only one option for reducing the effect of DIF on the trait estimates used for the matching criterion. Reise *et al.* (1993) pointed out that although items with DIF measure differently in different groups, they are still measuring the same underlying construct. This point is especially relevant in psychological measures where some items can be considered crucial in measuring a certain construct (e.g., crying in measuring depression), even if they are known to function differently between some demographic groups (e.g., gender).

To address these issues, Crane *et al.* (2006) developed an iterative process to update trait estimates using group-specific IRT item parameter estimates for items found to have DIF. Specifically, each item with DIF is replaced by as many sparse items (response vectors) as there are groups. For example, if there are two groups, A and B, two new sparse item response vectors are formed to replace the original. In the first sparse item response vector, the responses are the same as the original item for group A, and missing for group B. In the second sparse item response vector, the pattern is reversed.

The non-DIF items have parameters estimated using data from the entire sample and are often referred to as anchor items because they ensure that scores for individuals in all of the groups are on the same metric. The group-specific items and the anchor items are used to obtain trait estimates that account for DIF which in turn are then used in subsequent logistic regression DIF analyses. This process is continued until the same set of items is found to have DIF over successive iterations.

This algorithm has much to recommend it compared with more traditional purification approaches. First, it is possible for items to have false positive identification with DIF at an early stage. Most traditional purification approaches would result in such an item being excluded from consideration for the matching criterion, even though in the subsequent iterations it may be found to not have DIF. Second, by including all of the items in the trait estimate, the measurement precision using this approach will be better than that for trait estimates that includes only a subset of the items. Finally, the iterative nature of this procedure avoids the forward stepwise nature of some algorithms, such as that used in the multiple indicators multiple causes framework (Jones 2006).

2.6. Fitting the graded response model

Unlike other IRT-based DIF detection methods focusing on tests of the equality of item parameters across groups (Lord 1980; Raju *et al.* 2009; Thissen 2001), the main objective of fitting an IRT model under **lordif** is to obtain IRT trait estimates to serve as the matching criterion. Therefore, the choice of a specific IRT model is of little consequence in the current application, because trait estimates for the same data based on different IRT models (e.g., graded response model vs. Generalized Partial Credit Model) are virtually interchangeable ($r > 0.99$) (Cook 1999). However, the graded response model might be preferred in the current context on the basis of its inherent connection to ordinal logistic regression. The model assumes any response to item i, u_i, can be scored into m_i ordered categories, e.g., $u_i \in \{0, 1, 2, \ldots, (m_i - 1)\}$. The model then defines $(m_i - 1)$ cumulative category response functions as follows:

$$\begin{array}{c} P_1^*(\theta) \equiv P(u_i \geq 1|\theta)=\{1+exp[-a_i(\theta-b_{i_1})]\}^{-1} \\ P_2^*(\theta) \equiv P(u_i \geq 2|\theta)=\{1+exp[-a_i(\theta-b_{i_2})]\}^{-1} \\ \cdots \\ P_{(m_i-1)}^*(\theta) \equiv P(u_i \geq m_i-1|\theta)=\{1+exp[-a_i(\theta-b_{i_{(m_i-1)}})]\}^{-1} \end{array}$$

where the item discrimination parameter a_i is finite and positive, and the location parameters, b_{i_1}, b_{i_2}, …, $b_{i(m_i-1)}$, satisfy

$$b_{i_1}<b_{i_2}<\ldots<b_{i_{(m_i-1)}}.$$

Further, $b_{i_0} \equiv -\infty$ and $b_{i_{m_i}} \equiv \infty$ such that $P_0^*=1$ and $P_{(m_i)}^*=0$. Finally, for any response category, $u_i \in \{0, 1, \ldots, m_i - 1\}$, the category response function can be expressed as

$$P_{u_i}(\theta)=P_{u_i}^*(\theta)-P_{(u_i+1)}^*(\theta)>0.$$

2.7. Scale transformation

The impact of DIF on scores can be determined by comparing the initial trait score to the final trait score that accounts for DIF. To compare scores, however, the IRT item parameter estimates from the initial and final calibrations should be placed on the same metric. The method developed by Stocking and Lord (1983) can be used to determine the appropriate transformation. Using the non-DIF items as anchor items, the procedure can equate the group-specific item parameter estimates from the final "matrix" calibration (M) to the metric of the initial "single-group" calibration (S). The Stocking-Lord equating procedure finds the linear transformation constants, A and B, that minimize the sum of squared differences between the test characteristic curves (TCCs) based on J non-DIF items over a θ grid (e.g., $-4 \leq \theta \leq 4$). The loss function (L) to be minimized can be expressed as follows:

$$\begin{array}{c} L=\sum_{q=1}^{Q}[TCC_S(\theta_q)-TCC_M(\theta_q)]^2, \\ TCC_S(\theta_q)=\sum_{i=1}^{J}\sum_{k\in u_i}k \cdot P(u_i=k|\theta_q, a_{i_S}, b_{i_{1_S}}, b_{i_{2_S}}, \ldots, b_{i_{(m_i-1)_S}}), \\ TCC_M(\theta_q)=\sum_{i=1}^{J}\sum_{k\in u_i}k \cdot P(u_i=k|\theta_q, a_{i_M}^*, b_{i_{1_M}}^*, b_{i_{2_M}}^*, \ldots, b_{i_{(m_i-1)_M}}^*), \end{array}$$

$$\begin{array}{c} a_{iM}^*=\frac{a_{iM}}{A}, \\ b_{i_{1_M}}^*=A \cdot b_{i_{1_M}}+B, \\ b_{i_{2_M}}^*=A \cdot b_{i_{2_M}}+B, \\ \cdots \\ b_{i_{(m_i-1)_M}}^*=A \cdot b_{i_{(m_i-1)_M}}+B, \end{array}$$

where Q is the number of equi-spaced quadrature points over the θ grid, J is the number of non-DIF items, a_{i_S}, $b_{i_{1_S}}$, $b_{i_{2_S}}$, …, $b_{i(m_i-1)_S}$ are the single-group item parameter estimates for the *i*th non-DIF item, and $a^*_{i_M}, b^*_{i_{1_M}}, b^*_{i_{2_M}}, \ldots, b_{i_{(m_i-1)_M}}$ are the matrix calibration item parameter estimates for the same item.

3. The lordif package

3.1. Overview

The **lordif** package is based on the **difwithpar** framework (Crane *et al.* 2006). We developed the **lordif** package (available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=lordif) to perform DIF detection with a flexible iterative hybrid OLR/IRT framework. The **lordif** package is also able to perform OLR with sum scores rather than IRT scores. **lordif** also incorporates a Monte Carlo simulation approach to derive empirical threshold values for various DIF statistics and magnitude measures. The **lordif** package generates DIF-free datasets of the same dimension as the empirical dataset using the purified trait estimates and initial single-group item parameter estimates obtained from the real data, preserving observed group differences and distributions. The user specifies the number of replications (`nr`) and the Type I error rate (e.g., `alpha = 0.01`). The program then applies the OLR/IRT procedure over the simulated DIF-free datasets and computes the statistics and magnitude measures. Finally, the program identifies a threshold value that cuts off the most extreme (α × 100)% of each of the statistics and magnitude measures.

The **lordif** package is built on two main packages: The **ltm** package (Rizopoulos 2006) to obtain IRT item parameter estimates according to the graded response model (Samejima 1969) and the **Design** package (Harrell Jr. 2009) for fitting (ordinal) logistic regression models. Both the **ltm** and **Design** packages can handle binary outcome variables as a special case and hence allow the **lordif** package to handle both dichotomous and polytomous items. The **Design** package also handles a grouping variable with more than two levels (e.g., Black, Hispanic and White) by automatically entering it into a model as a set of dummy variables.

The **lordif** package allows the user to choose specific criteria and their associated thresholds for declaring items to have uniform and non-uniform DIF. Items found displaying DIF are recalibrated in appropriate subgroups to generate trait estimates that account for DIF. These steps are repeated until the same items are identified with DIF on consecutive runs. The program uses the Stocking and Lord (1983) equating procedure to place the group-specific item parameter estimates onto the scale defined by the initial naive (i.e., no-DIF) run and to facilitate evaluations of the impact on individual trait estimates on the same scale. Items displaying no DIF serve as anchor items.

3.2. Algorithm

In what follows, we will describe the algorithm used in the **lordif** package in more detail.

1. Data preparation: Check for sparse cells (rarely observed response categories; determined by a minimum cell count specified by the user (e.g., `minCell = 5`); collapse/recode response categories as needed based on the minimum cell size requirement specified.

2. IRT calibration: Fit the graded response model (using the `grm` function in **ltm**) to obtain a single set of item parameters for all groups combined.

3. Trait estimation: Obtain trait (ability) estimates using the expected a posteriori (EAP) estimator with omitted responses treated as not presented.

4. Logistic regression: Fit three (ordinal) logistic models (Models 1, 2 and 3) on each item using the `lrm` function in **Design** (observe these are item-wise regressions); generate three likelihood-ratio χ^2 statistics for comparing three nested logistic

regression models (Models 1 vs. 3, Models 1 vs. 2, and Models 2 vs. 3); compute three pseudo R^2 measures – Cox & Snell (Cox and Snell 1989), Nagelkerke (Nagelkerke 1991), and McFadden (Menard 2000) – for three nested models and compute differences between them; compute the absolute proportional change in point estimates for β_1 from Model 1 to Model 2 as follows: $|(\beta_1 - \beta_1^*)/\beta_1^*|$, where β_1^* is the regression coefficient for the matching criterion (ability) from Model 1 and β_1 is the same term from Model 2.

5. **Detecting DIF:** Flag DIF items based on the detection criterion (`“Chisqr”`, `“R2”`, or `“Beta”`) and a corresponding flagging criterion specified by the user (e.g., `alpha = 0.01 for criterion = “Chisqr”`); `for criterion = “Chisqr”` an item is flagged if any one of the three likelihood ratio χ^2 statistics is significant (the 2-*df* test for non-uniform DIF, χ^2_{13}, as a sole criterion may lack power if DIF is attributable primarily to uniform DIF, although inflated Type I error might be of concern).

6. Sparse matrix: Treat DIF items as unique to each group and prepare a sparse response matrix by splitting the response vector for each flagged item into a set of sparse vectors containing responses for members of each group (e.g., males and females if DIF was found related to gender). In other words, each DIF item is split into multiple sparse variables such that each variable corresponds to the data of just one group and missing for all other groups. Note that sparse matrices are to account for DIF in the trait estimate; (ordinal) logistic regression DIF detection is performed on the original data matrix.

7. IRT recalibration: Refit the graded response model on the sparse matrix data and obtain a single set of item parameter estimates for non-DIF items and group-specific item parameter estimates for DIF items.

8. Scale transformation: Equate Stocking and Lord (1983) item parameter estimates from the matrix calibration to the original (single-group) calibration by using non-DIF items as anchor items (this step is necessary only when looking at DIF impact and can be deferred until the iterative cycles have concluded).

9. Trait re-estimation: Obtain EAP trait (ability) estimates based on item parameter estimates from the entire sample for items that did not have DIF and group-specific item parameter estimates for items that had DIF.

10. Iterative cycle: Repeat Steps 4 through 9 until the same items are flagged for DIF or a preset maximum number of iterations has been reached. Using the trait estimates from the previous round that account for DIF detected to that point, (ordinal) logistic regression DIF detection is repeated on all items including previously flagged items.

11. Monte Carlo simulation: Generate DIF-free datasets `nr` number of times (e.g., `nr = 1000`), using the final trait estimates accounting for DIF (Step 10) and the initial single-group item parameter estimates (Step 2). Each simulated dataset contains the same number of cases by group as the empirical dataset and reflects observed group differences in trait estimates. For each simulated dataset, obtain trait (ability) estimates based on the single-group item parameter estimates and run the OLR/IRT procedure. Compute the DIF statistics and magnitude measures for each simulated dataset and store the results for all replications. Identify a threshold value for each statistic/magnitude measure that cuts off the most extreme (defined by α) end of its cumulative distribution.

3.3. lordif vs. difwithpar

The **lordif** package differs in several ways from the previously developed **difwithpar** program. Improvements include the use of the **ltm** package (Rizopoulos 2006) rather than the proprietary software **PARSCALE** (Muraki and Bock 2005) for IRT item parameter estimation. The **lordif** package also includes the following important changes. First, **lordif** permits comparison of Model 1 with Model 3, facilitating a single omnibus test of both uniform and non-uniform DIF. Second, **lordif** automates the steps of DIF detection and subsequent IRT parameter estimation in a single invocation of the iterative algorithm; whereas **difwithpar** performs a single iteration and the user must continue the steps until the same items are identified on subsequent runs. Third, **lordif** performs the Stocking-Lord equating (Stocking and Lord 1983) that facilitates investigations of DIF impact on the same metric. Finally, and perhaps most important, **lordif** implements the Monte Carlo procedures described previously to identify empirically-based thresholds for DIF detection.

4. Illustration

To illustrate, the procedure was applied to a real dataset and the results were compared to the standard sum-score based approach. We analyzed a dataset ($N = 766$) on a 29-item anxiety bank (Pilkonis *et al.* 2011, see appendix) for DIF related to age using data from the Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS is an NIH Roadmap initiative designed to improve patient-reported outcomes using state-of-the-art psychometric methods (for detailed information, see http://www.nihpromis.org/). The reference and focal groups were defined as younger (< 65; $n = 555$) and older (; $n = 211$), respectively. All items shared the same rating scale with five response categories (Never, Rarely, Sometimes, Often, and Always). The scale was constructed such that higher scores mean higher levels of anxiety. The $S - X^2$ model fit statistics (Kang and Chen 2008; Orlando and Thissen 2003) were examined for the graded response model (Samejima 1969) using the **IRTFIT** (Bjorner *et al.* 2006) macro program. All 29 items had adequate or better model fit statistics ($p > 0.05$).

Running **lordif** requires a minimum level of competence in `R`, including reading external datasets using `R` syntax submitted via a command line interface or a syntax file. In what follows we present sample `R` code to demonstrate specifics of the interface with **lordif** and to generate output for discussion in the subsequent section:

```
R> library(“lordif”)
R> data(“Anxiety”)
R> Age <- Anxiety$age
R> Resp <- Anxiety[paste(“R”, 1:29, sep = “”)]
R> ageDIF <- lordif(Resp, Age, criterion = “Chisqr”, alpha = 0.01, + minCell
= 5)
R> print(ageDIF)
R> summary(ageDIF)
R> plot(ageDIF, labels = c(“Younger (<65)”, “Older (65+)”))
```

The `library ( “lordif” )` command loads the **lordif** package (and other dependent packages) into the `R` computing environment. The `data (  “Anxiety”)` command loads the Anxiety dataset containing 29 item response variables (named $R1, R2, \ldots, R29$) and three binary demographic indicators including the age group (0 = Younger and 1 = Older). The next two lines of commands extract those variables from the dataset and create a vector for the age indicator (`Age`) and a matrix for the item response variables (`Resp`). The `lordif ( Resp,` `Age, …`) command performs the OLR/IRT DIF procedure on the data with specified options (details provided below) and saves the output as `ageDIF`. The `print` (`ageDIF`) and `summary` (`ageDIF`) commands generate basic and extended output, respectively. The `plot` (`ageDIF`) command then takes the output (`ageDIF`) and generates diagnostic plots. An optional Monte Carlo simulation procedure (and the corresponding `print` and `summary` methods) can be invoked on the output (`ageDIF`) to obtain empirical threshold values by

```
R> ageMC <- montecarlo(ageDIF, alpha = 0.01, nr = 1000)
R> print(ageMC)
R> summary(ageMC)
```

Monte Carlo simulations generally require a large number of iterations and are computationally intensive – the above simulation run took approximately 30 minutes on an Intel Core2 Duo CPU at 2.53GHz running Windows Vista. Finally, the empirical threshold values can be displayed visually by

```
R> plot(ageMC)
```

We used the likelihood ratio (LR) χ^2 test (`criterion = “Chisqr”`) as the detection criterion at the α level of 0.01, and McFadden's pseudo R^2 (default) as the magnitude measure. With a minimum cell count of 5, all items ended up with one or more response categories collapsed. After recoding (done by **lordif**), four items ended up with four response categories, one item had two categories, and the rest had three. Using these settings, **lordif** terminated in two iterations flagging five items as displaying age-related DIF – #1 (“I felt fearful”), #9 (“I was anxious if my normal routine was disturbed”), #11 (“I was easily startled”), #18 (“I worried about other people's reactions to me”), and #24 (“Many situations made me worry”). The standard sum score-based method flagged the same items and one additional item – #7 (“I felt upset”). The `plot` function in **lordif** shows (see Figure 1) the theta distributions for the younger and older groups. Older people on average had lower mean scores than their younger counterparts (−0.57 vs. 0.04). The `plot` function then displays four diagnostic plots for each of the flagged items (see Figures 2–6). The top left plot in Figure 2 shows item true-score functions based on group-specific item parameter estimates. The slope of the function for the older group was substantially higher than that for the younger group, indicating non-uniform DIF. The LR χ^2 test for uniform DIF, comparing Model 1 and Model 2, was not significant (p = 0.42), whereas the 1-*df* test for comparing Model 2 and Model 3 was significant (p = 0.0004). It is interesting to note that had the 2-*df* test (comparing Models 1 and 3) been used as the criterion for flagging, this item would not have been flagged at $\alpha = 0.01$ (p = 0.011).

The bottom left plot in Figure 2 juxtaposes the item response functions for younger and older adults. The non-uniform component of DIF revealed by the LR χ^2 test can also be observed in the difference of the slope parameter estimates (3.04 vs. 1.95). Although there was no significant uniform DIF, on close inspection the difference in the second category threshold values (shown as hash marks immediately above the x-axis) for the two groups were noticeable (1.21 vs. 1.77). For polytomous items, a single item-level index of DIF may not provide adequate information concerning which response categories (or score levels) contribute to the DIF effect. The combination of visual and model-based approaches in **lordif** provide useful diagnostic information at the response category level, which can be systematically investigated under the differential step functioning framework (Penfield 2007; Penfield *et al.* 2009).

The top right plot in Figure 2 presents the expected impact of DIF on scores as the absolute difference between the item true–score functions (Kim *et al.* 2007). There is a difference in the item true–score functions peaking at approximately $\theta = 1.5$, but the density–weighted impact (shown in the bottom right plot) is negligible because few subjects have that trait level in this population. When weighted by the focal group trait distribution the expected impact became negligible, which is also apparent in the small McFadden's pseudo R^2 measures (printed on the top left plot), i.e., R^2_{13}=0.007 and R^2_{23}=0.006. Figure 3 displays the plots for item *#9 ("I was anxious if my normal routine was disturbed")*, which shows statistically significant uniform DIF, $Pr(\chi^2_{12}, 2)$<0.001. The LR χ^2_{13} was also significant; however, as the LR χ^2_{23} was non–significant this result suggests the DIF was primarily uniform. The item response functions suggest that uniform DIF was due to the first category threshold value for the focal group being smaller than that for the reference group (−0.31 vs. +0.23). Figure 4 displays slightly stronger uniform DIF for item *#11 ("I was easily startled")*. Again, both χ^2_{12} and χ^2_{13} were significant ($p < 0.001$) with non–significant χ^2_{23}. McFadden's R^2 change for uniform DIF was 0.009, which is considered a negligible effect size (Cohen 1988). The item response functions show that the category threshold parameters for the focal group were uniformly smaller than those for the reference group. Figure 5 displays another item with uniform DIF, item *#18 ("I worried about other people's reactions to me")*, but in the opposite direction. The item true–score functions reveal that older people are prone to endorse the item with higher categories compared to younger people with the same overall anxiety level. The item response functions also show that the category threshold parameters for the focal group were uniformly higher than they were for the reference group. Finally, Figure 6 displays uniform DIF for item *#24 ("Many situations made me worry")*–both χ^2_{12} and χ^2_{13} tests were statistically significant ($p < 0.001$) with a non–significant χ^2_{23}. However, the item response functions (and the item parameter estimates) revealed a somewhat different diagnosis–the difference in slope parameters (2.80 vs. 1.88) suggests non–uniform DIF.

The diagnostic plots for individual DIF items (see Figures 2 through 6) are followed in **lordif** by two test–level plots. Figure 7 shows the impact of all of the DIF items on test characteristic curves (TCCs). The left plot is based on item parameter estimates for all 29 items including the group–specific parameter estimates for the five items identified with DIF. The plot on the right is based only on the group–specific parameter estimates. Although the impact shown in the plot on the right is very small, the difference in the TCCs implies that older adults would score slightly lower (less anxious) if age group–specific item parameter estimates were used for scoring. When aggregated over all the items in the test (left plot) or over the subset of items found to have DIF (right plot), differences in item characteristic curves (Figure 7) may become negligibly small due to canceling of differences in opposite directions, which is what appears to have happened here. However, it is possible for the impact on trait estimates to remain.

For the impact at the individual score level, **lordif** compares the initial naive theta estimate and the "purified" theta estimates from the final run accounting for DIF as shown in Figure 8. Notice that the item parameter estimates from the final run were equated (using non–DIF items as anchor items) to the initial, single–group calibration and not re–centered to 0.0 (see Step 8), and hence the mean difference ("initial minus purified") is not necessarily 0.0. This is a modification from the original **difwithpar** framework (Crane *et al.* 2006). The Box–and–Whisker plot on the left shows the median difference (over all examinees) is about 0.1 and the differences ranged from −0.176 to +0.263 with a mean of 0.073. The scatter plot on the right shows that the final theta estimates had a slightly larger standard deviation (1.122 vs. 1.056). The dotted horizontal reference line is drawn at the mean difference between the initial and purified estimates (i.e., 0.073). With the inclusion of five items with group–specific item parameters, scores at both extremes became slightly more extreme. Accounting for DIF by using group–specific item parameters had negligible effects on individual scores. In the absence of a clinical effect size, we labeled individual changes as “salient” if they exceeded the median standard error (0.198) of the initial score. About 1.96% (15 of 766) of the subjects had salient changes. About 0.52% (4 out of 766) had score changes larger than their initial standard error estimates. Cohen’s effect size *d* for the difference between the two group means (Younger minus Older) was nearly unchanged after identifying and accounting for DIF (from 0.544 to 0.561).

Table 1 shows the Monte Carlo threshold values for the statistics and magnitude measures by item, based on `nr = 1000` and `alpha = 0.01`. On average, the empirical threshold values for the probability associated with the χ^2 statistic were close to the nominal α level–the mean probability threshold values across items were 0.010, 0.011, and 0.011 for χ^2_{12},χ^2_{23}, and χ^2_{23}, respectively. Figure 9 displays the probability thresholds for the three χ^2 statistics by item. The horizontal reference line is drawn at the nominal α level (i.e., 0.01). There is no indication that the empirical threshold values are systematically deviated from the nominal level, which is congruent with previous research showing that the Type I error rate is well controlled under the likelihood ratio test (Kim and Cohen 1998).

Figure 10 presents the threshold values on pseudoR^2 measures. As expected, data generated under no DIF conditions produced negligibly small pseudo R^2 measures, i.e., considerably smaller than Cohen’s guideline for a small effect size (0.02). Although some fluctuations are visible across items, the pseudo R^2 thresholds were unmistakably smaller than any guidelines for non–trivial effects. Unlike the ordinary least squares R^2, pseudo R^2 measures may lack a reasonable interpretation. For instance, the Cox & Snell (Cox and Snell 1989) pseudo R^2 measure cannot attain the value of 1 even if the model fits perfectly and residuals are zero (Mittlböck and Schemper 1996). Although the Nagelkerke formula for pseudo R^2 corrects the scale issue, it may still lack an immediate interpretation (Mittlböck and Schemper 1999). Mc-Fadden’s pseudo R^2 measure, on the other hand, offers intuitively meaningful interpretations, e.g., proportional reduction in the −2 log–likelihood statistic. However, since the primary interest in the current context is the change in the pseudo R^2 measures between two nested models, the scale issue may not be a serious concern. Although further study is needed, it is interesting to note that the empirical thresholds based on Cox & Snell displayed the least amount of variation across items (see Figure 10 and standard deviations at the bottom of Table 1).

The threshold on proportionate β_1 change was fairly consistent over items (mean= 0.0323, SD= 0.0063). The maximum change across items was 0.0538 (i.e., about 5% change) and was from item #17, which was also the item with the largest pseudo R^2 measures (see Figure 11). A 10% change in β_1 (i.e., 0.1) has been used previously as the criterion for the presence of uniform DIF (Crane *et al.* 2004). The proportionate β_1 change effect size is closely related to the pseudo R^2_{12} measures (comparing Model 1 vs. Model 2). The correlation coefficients between the three R^2_{12} measures and the proportionate β_1 change thresholds across items were 0.855, 0.928, and 0.784 for Cox & Snell, Nagelkerke, and McFadden, respectively. The correlation between Nagelkerke’s R^2_{12} and the proportionate β_1 change effect size was especially high. It is interesting to note that when the proportionate β_1 change thresholds were linearly interpolated (based on the threshold values in Table 1), a 10% change in β_1 is roughly equivalent to 0.02 in Nagelkerke’s R^2_{12}. Although the two effect size measures and associated agging criteria originated in different disciplines, they appear to be consistent in this context.

5. Conclusion

The **lordif** package is a powerful and flexible freeware platform for DIF detection. Ordinal logistic regression (OLR) provides a flexible and robust framework for DIF detection, especially in conjunction with trait level scores from IRT as the matching criterion (Crane *et al.* 2006). This OLR/IRT hybrid approach implemented in **lordif** provides statistical criteria and various magnitude measures for detecting and measuring uniform and non–uniform DIF. Furthermore, the use of an IRT trait score in lieu of the traditional sum score makes this approach more robust and applicable even when responses are missing by design, e.g., block–testing, because unlike raw scores comparable IRT trait scores can be estimated based on different sets of items. The **lordif** package also introduces Monte Carlo procedures to facilitate the identification of reasonable detection thresholds to determine whether items have DIF based on Type I error rates empirically found in the simulated data. This functionality was not available in **difwithpar** (Crane *et al.* 2006).

A multitude of DIF detection techniques have been developed. However, very few are available as an integrated, non–proprietary application, and none offers the range of features of **lordif**. Of the non–proprietary programs, **DIFAS** (Penfield 2005) and **EZDIF** (Waller 1998) are based on the sum score. **DIFAS** implements a variety of DIF detection techniques based on raw scores for both dichotomous and polytomous items. **EZDIF** only allows dichotomous items, although it employs a two–stage purification process of the trait estimate. **IRTLRDIF** (Thissen 2001) uses the IRT parameter invariance framework and directly tests the equality of item parameters, but does not allow for empirical determination of DIF detection criteria. It should be noted that it is on the basis of its features that we recommended **lordif**; we have not conducted any simulations to compare its findings with these programs.

In our illustration, five of 29 items were found to have modest levels of DIF related to age. Findings were very similar between the standard sum score–based method and the iterative hybrid OLR/IRT algorithm. The IRT model–based OLR approach provides a mechanism to diagnose DIF in terms of the impact on IRT parameters. The impact of DIF on the TCC was minimal, though some item characteristic curves (ICCs) clearly demonstrated differences. When accounting for DIF, a very small percentage of the subjects had "salient" score changes. This definition of salience is based on the median standard error of measurement (SEM) for the scale. In this instance, the "scale" is an entire item bank with a relatively small median SEM. When a minimal clinically important difference (MCID) is available for a scale, Crane and colleagues recommend a similar approach, but use the MCID and refer to differences beyond the MCID as "relevant" DIF impact (Crane *et al.* 2007b). While the MCID for the PROMIS anxiety scale has yet to be determined, it will likely be larger than the value used to indicate salience here (0.198). In that case, the proportion of subjects who will have *relevant* DIF will be even smaller than that found to have *salient* DIF, further buttressing our view that DIF related to age is negligible in this dataset.

The Monte Carlo simulation results confirmed that the likelihood ratio χ^2 test maintains the Type I error adequately in this dataset. Some pseudo R^2 values varied across items, but overall they were very small under simulations that assume no DIF. Some pseudo R^2 values may vary from item to item depending on the number of response categories and the distribution within each response category (Menard 2000), so using a single pseudo R^2 threshold may result in varying power across items to detect DIF (Crane *et al.* 2007b). Monte Carlo simulations can help inform the choice of reasonable thresholds. If a single threshold is to be used across all items, it should be set above the highest value identified in Monte Carlo simulations. For instance, the maximum pseudo R^2 in Table 1 was 0.015, and thus a reasonable lower bound that would avoid Type I errors might be 0.02, which interestingly corresponds to a small, non–negligible effect size (Cohen 1988).

Subsequent development will be facilitated by the algorithm's ability to account for DIF using group specific item parameters. Future studies may focus on examining the potential greater impact of DIF in a computer adaptive testing (CAT) framework, and developing a CAT platform that can account for DIF in real time. It will also be interesting to compare the OLR/IRT framework implemented in **lordif** to other DIF detection techniques based on the IRT parameter invariance assumption, such as **IRTLRDIF** (Thissen 2001) and **DFIT** (Raju *et al.* 2009). For instance, it will be interesting to see how those procedures would diagnose item #24 (see Figure 6). As noted previously, this item displayed no non–uniform DIF (p = 0.85); however, the slope parameter estimates appeared quite different (1.88 vs. 2.80).

In conclusion, in this paper we introduce **lordif**, a new freeware package for DIF detection that combines IRT and ordinal logistic regression. The Monte Carlo simulation feature facilitates empirical identification of detection thresholds, which may be helpful in a variety of settings. Standard output graphical displays facilitate sophisticated understanding of the nature of and impact of DIF. We demonstrated the use of the package on a real dataset, and found several anxiety items to have DIF related to age, though they were associated with minimal DIF impact.

Acknowledgments

This study was supported in part by the Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS is an NIH Roadmap initiative to develop a computerized system measuring PROs in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS II was funded by cooperative agreements with a Statistical Center (Northwestern University, PI: David F. Cella, PhD, 1U54AR057951), a Technology Center (Northwestern University, PI: Richard C. Gershon, PhD, 1U54AR057943), a Network Center (American Institutes for Research, PI: Susan (San) D. Keller, PhD, 1U54AR057926) and thirteen Primary Research Sites (State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, 1U01AR057948; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, 1U01AR057954; University of Washington, Seattle, PIs: Dagmar Amtmann, PhD and Karon Cook, PhD1U01AR052171; University of North Carolina, Chapel Hill, PI: Darren A. DeWalt, MD, MPH, 2U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, 1U01AR057956; Stanford University, PI: James F. Fries, MD, 2U01AR052158; Boston University, PIs: Stephen M. Haley, PhD and David Scott Tulsky, PhD1U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD and Brennan Spiegel, MD, MSHS, 1U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, 2U01AR052155; Georgetown University, Washington DC, PIs: Carol. M. Moinpour, PhD and Arnold L. Potosky, PhD, U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan Dewitt, MD, 1U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, 1U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, 2U01AR052186). NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD, Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein, MD, Ellen Werner, PhD and James Witter, MD, PhD. This manuscript was reviewed by PROMIS reviewers before submission for external peer review. See the Web site at http://www.nihpromis.org/ for additional information on the PROMIS initiative.

Additional support for the second and third authors were provided by P50 AG05136 (Raskind) and R01 AG 029672 (Crane). The authors appreciate the careful efforts of Shubhabrata Mukherjee, PhD, to format the manuscript in LATEX.

References

Agresti, A. Categorical Data Analysis. John Wiley & Sons; New York: 1990.

Andersen RB. On the Comparability of Meaningful Stimuli in Cross-Cultural Research. Sociometry. 1967; 30:124–136. [PubMed: 6044173]

Bjorner, JB.; Smith, KJ.; Orlando, M.; Stone, C.; Thissen, D.; Sun, X. IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models. Quality Metric Inc; Lincoln, RI: 2006.

Camilli, G.; Shepard, LA. Methods for Identifying Biased Test Items. Sage; Thousand Oaks: 1994.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2. Lawrence Earlbaum Associates; Hillsdale, NJ: 1988.

Cook K. A Comparison of Three Polytomous Item Response Theory Models in the Context of Testlet Scoring. Journal of Outcome Measurement. 1999; 3:1–20. [PubMed: 10063769]

Cox, DR.; Snell, EJ. The Analysis of Binary Data. 2. Chapman and Hall; London: 1989.

Crane PK, Cetin K, Cook K, Johnson K, Deyo R, Amtmann D. Differential Item Functioning Impact in a Modified Version of the Roland-Morris Disability Questionnaire. Quality of Life Research. 2007a; 16(6):981–990. [PubMed: 17443419]

Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis with Ordinal Logistic Regression Techniques: DIF Detect and difwithpar. Medical Care. 2006; 44 (11 Supp 3):S115–S123. [PubMed: 17060818]

Crane PK, Gibbons LE, Narasimhalu K, Lai JS, Cella D. Rapid Detection of Differential Item Functioning in Assessments of Health-Related Quality of Life: The Functional Assessment of Cancer Therapy. Quality of Life Research. 2007b; 16(1):101–114. [PubMed: 17111233]

Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D. A Comparison of Three Sets of Criteria for Determining the Presence of Differential Item Functioning Using Ordinal Logistic Regression. Quality of Life Research. 2007c; 16(Supp 1):69–84. [PubMed: 17554640]

Crane PK, Gibbons LE, Willig JH, Mugavero MJ, Lawrence ST, Schumacher JE, Saag MS, Kitahata MM, Crane HM. Measuring Depression and Depressive Symptoms in HIV-Infected Patients as Part of Routine Clinical Care Using the 9-Item Patient Health Questionnaire (PHQ-9). AIDS Care. 2010; 22(7):874–885. [PubMed: 20635252]

Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, Kuller L, Hall K, van Belle G. Item Response Theory Facilitated Cocalibrating Cognitive Tests and Reduced Bias in Estimated Rates of Decline. Journal of Clinical Epidemiology. 2008a; 61(10):1018–1027. [PubMed: 18455909]

Crane PK, Narasimhalu K, Gibbons LE, Pedraza O, Mehta KM, Tang Y, Manly JJ, Reed BR, Mungas DM. Composite Scores for Executive Function Items: Demographic Heterogeneity and Relationships with Quantitative Magnetic Resonance Imaging. Journal of International Neuropsycholigal Society. 2008b; 14(5):746–759.

Crane PK, van Belle G, Larson EB. Test Bias in a Cognitive Test: Differential Item Functioning in the CASI. Statistics in Medicine. 2004; 23:241–256. [PubMed: 14716726]

De Boeck, P.; Wilson, M. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer-Verlag; New York: 2004.

Donoghue, JR.; Holland, PW.; Thayer, DT. A Monte Carlo Study of Factors That Affect the Mantel-Haenszel and Standardization Measures of Differential Item Functioning. In: Holland, P.; Wainer, H., editors. Differential Item Functioning. Erlbaum; Hillsdale, NJ: 1993. p. 137-166.

Dorans, NJ.; Holland, PW. DIF Detection and Description: Mantel-Haenszel and Standardization. In: Holland, P.; Wainer, H., editors. Differential Item Functioning. Erlbaum; Hillsdale, NJ: 1993. p. 35-66.

French AW, Miller TR. Logistic Regression and Its Use in Detecting Differential Item Functioning in Polytomous Items. Journal of Educational Measurement. 1996; 33:315–332.

French BF, Maller SJ. Iterative Purification and Effect Size Use with Logistic Regression for Differential Item Functioning Detection. Educational and Psychological Measurement. 2007; 67(3):373–393.

Gibbons LE, McCurry S, Rhoads K, Masaki K, White L, Borenstein AR, Larson EB, Crane PK. Japanese-English Language Equivalence of the Cognitive Abilities Screening Instrument among Japanese-Americans. International Psychogeriatrics. 2009; 21(1):129–137. [PubMed: 18947456]

Gregorich SE. Do Self-Report Instruments Allow Meaningful Comparisons across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. Medical Care. 2006; 44(11 Supp 3):S78–S94. [PubMed: 17060839]

Harrell, FE, Jr. Design: Design Package. R package version 2.3-0. 2009. URL http://CRAN.R-project.org/package=Design

Holland, PW.; Thayer, DT. Differential Item Prformance and the Mantel-Haenszel Procedure. In: Wainer, H.; Braun, HI., editors. Test Validity. Erlbaum; Hillsdale, NJ: 1988. p. 129-145.

Jodoin MG, Gierl MJ. Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection. Applied Measurement in Education. 2001; 14:329–349.

Jones R. Identification of Measurement Differences between English and Spanish Language Versions of the Mini-Mental State Examination: Detecting Differential Item Functioning Using MIMIC Modeling. Medical Care. 2006; 44:S124–S133. [PubMed: 17060819]

Kang T, Chen T. Performance of the Generalized $S-X^2$ Item Fit Index for Polytomous IRT Models. Journal of Educational Measurement. 2008; 45(4):391–406.

Kim SH, Cohen AS. Detection of Differential Item Functioning under the Graded Response Model with the Likelihood Ratio Test. Applied Psychological Measurement. 1998; 22:345–355.

Kim SH, Cohen AS, Alagoz C, Kim S. DIF Detection Effect Size Measures for Polytomously Scored Items. Journal of Educational Measurement. 2007; 44(2):93–116.

Li H, Stout W. A New Procedure for Detection of Crossing DIF. Psychometrika. 1996; 61:647–677.

Lord, FM. Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates; Hillsdale, NJ: 1980.

Mair, P.; Hatzinger, R. Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R; Journal of Statistical Software. 2007. p. 1-20.URL http://www.jstatsoft.org/v20/i09/

Maldonado G, Greenland S. Simulation Study of Confounder-Selection Strategies. American Journal of Epidemiology. 1993; 138(11):923–936. [PubMed: 8256780]

Menard S. Coefficients of Determination for Multiple Logistic Regression Analysis. The American Statistician. 2000; 54:17–24.

Miller TR, Spray JA. Logistic Discriminant Function Analysis for DIF Identi3cation of Polytomously Scored Items. Journal of Educational Measurement. 1993; 30:107–122.

Millsap RE, Everson HT. Methodology Review: Statistical Approaches for Assessing Measurement Bias. Applied Psychological Measurement. 1993; 17(4):297–334.

Mittlböck M, Schemper M. Explained Variation for Logistic Regression. Statistics in Medicine. 1996; 15:1987–1997. [PubMed: 8896134]

Mittlböck M, Schemper M. Computing Measures of Explained Variation for Logistic Regression Models. Computer Methods and Programs in Biomedicine. 1999; 58:17–24. [PubMed: 10195643]

Muraki, E.; Bock, D. PARSCALE. Vol. 4. Scientific Software International, Inc; Lincolnwood, IL: 2005. URL http://www.ssicentral.com/

Nagelkerke NJD. A Note on a General Definition of the Coefficient of Determination. Biometrika. 1991; 78:691–692.

Orlando M, Thissen D. Further Investigation of the Performance of $S-X^2$: An Item Fit Index for Use with Dichotomous Item Response Theory Models. Applied Psychological Measurement. 2003; 27 (4):289–298.

Penfield RD. DIFAS: Differential Item Functioning Analysis System. Applied Psychological Measurement. 2005; 29(2):150–151.

Penfield RD. Assessing Differential Step Functioning in Polytomous Items Using a Common Odds Ratio Estimator. Journal of Educational Measurement. 2007; 44:187–210.

Penfield RD, Gattamorta K, Childs RA. An NCME Instructional Module on Using Differential Step Functioning to Refine the Analysis of DIF in Polytomous Items. Educational Measurement: Issues and Practice. 2009; 28(1):38–49.

Perkins AJ, Stump TE, Monahan PO, McHorney CA. Assessment of Differential Item Functioning for Demographic Comparisons in the MOS SF-36 Health Survey. Quality of Life Research. 2006; 15 (3):331–348. [PubMed: 16547771]

Pilkonis, PA.; Choi, SW.; Reise, SP.; Stover, AM.; Riley, WT.; Cella, D. Item Banks for Measuring Emotional Distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): Depression, Anxiety, and Anger. 2011. Under review

Raju NS, Fortmann-Johnson KA, Kim W, Morris SB, Nering ML, Oshima TC. The Item Parameter Replication Method for Detecting Differential Functioning in the DFIT Framework. Applied Measurement in Education. 2009; 33(2):133–147.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2010. URL http://www.R-project.org/

Reise SP, Widaman KF, Pugh RH. Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance. Psychological Bulletin. 1993; 114(3):552–566. [PubMed: 8272470]

Rizopoulos, D. ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses; Journal of Statistical Software. 2006. p. 1-25.URL http://www.jstatsoft.org/v17/i05/

Rogers HJ, Swaminathan H. A Comparison of Logistic Regression and Mantel-Hanenszel Procedures for Detecting Differential Item Functioning. Applied Psychological Measurement. 1993; 17:105–116.

Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometrika Monograph. 1969:17.

Shealy, RT.; Stout, WF. An Item Response Theory Model for Test Bias and Differential Test Functioning. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Erlbaum; Hillsdale, NJ: 1993. p. 197-239.

StataCorp. Stata Statistical Software: Release 10. StataCorp LP; College Station, TX: 2007. URL http://www.stata.com/

Stocking ML, Lord FM. Developing a Common Metric in Item Response Theory. Applied Psychological Measurement. 1983; 7(2):201–210.

Swaminathan H, Rogers HJ. Detecting Differential Item Functioning Using Logistic Regression Procedures. Journal of Educational Measurement. 1990; 27:361–370.

Thissen, D. IRTLRDIF: Version v.2.0b. L.L. Thurstone Psychometric Laboratory, University of North Carolina; Chapel Hill: 2001. URL http://www.unc.edu/~dthissen/dl.html

Wainer, H. Model-Based Standardized Measurement of an Item's Differential Impact. In: Holland, P.; Wainer, H., editors. Differential Item Functioning. Erlbaum; Hillsdale, NJ: 1993.

Waller NG. EZDIF: Detection of Uniform and Nonuniform Differential Item Functioning With the Mantel-Haenszel and Logistic Regression Procedures. Applied Psychological Measurement. 1998; 22(4):391.

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. Directorate of Human Resources Research and Evaluation, Department of National Defense; Ottawa, ON: 1999.

Zwick R. When Do Item Response Function and Mantel-Haenszel Definition of Differential Item Functioning Coincide? Journal of Educational Statistics. 1990; 15(3):185–197.

A. PROMIS Anxiety Bank

See Pilkonis *et al.* (2011). In the past 7 days …

1. I felt fearful
2. I felt frightened
3. It scared me when I felt nervous
4. I felt anxious
5. I felt like I needed help for my anxiety
6. I was concerned about my mental health
7. I felt upset
8. I had a racing or pounding heart
9. I was anxious if my normal routine was disturbed
10. I had sudden feelings of panic
11. I was easily startled

12. I had trouble paying attention
13. I avoided public places or activities
14. I felt fidgety
15. I felt something awful would happen
16. I felt worried
17. I felt terrified
18. I worried about other people’s reactions to me
19. I found it hard to focus on anything other than my anxiety
20. My worries overwhelmed me
21. I had twitching or trembling muscles
22. I felt nervous
23. I felt indecisive
24. Many situations made me worry
25. I had difficulty sleeping
26. I had trouble relaxing
27. I felt uneasy
28. I felt tense
29. I had difficulty calming down

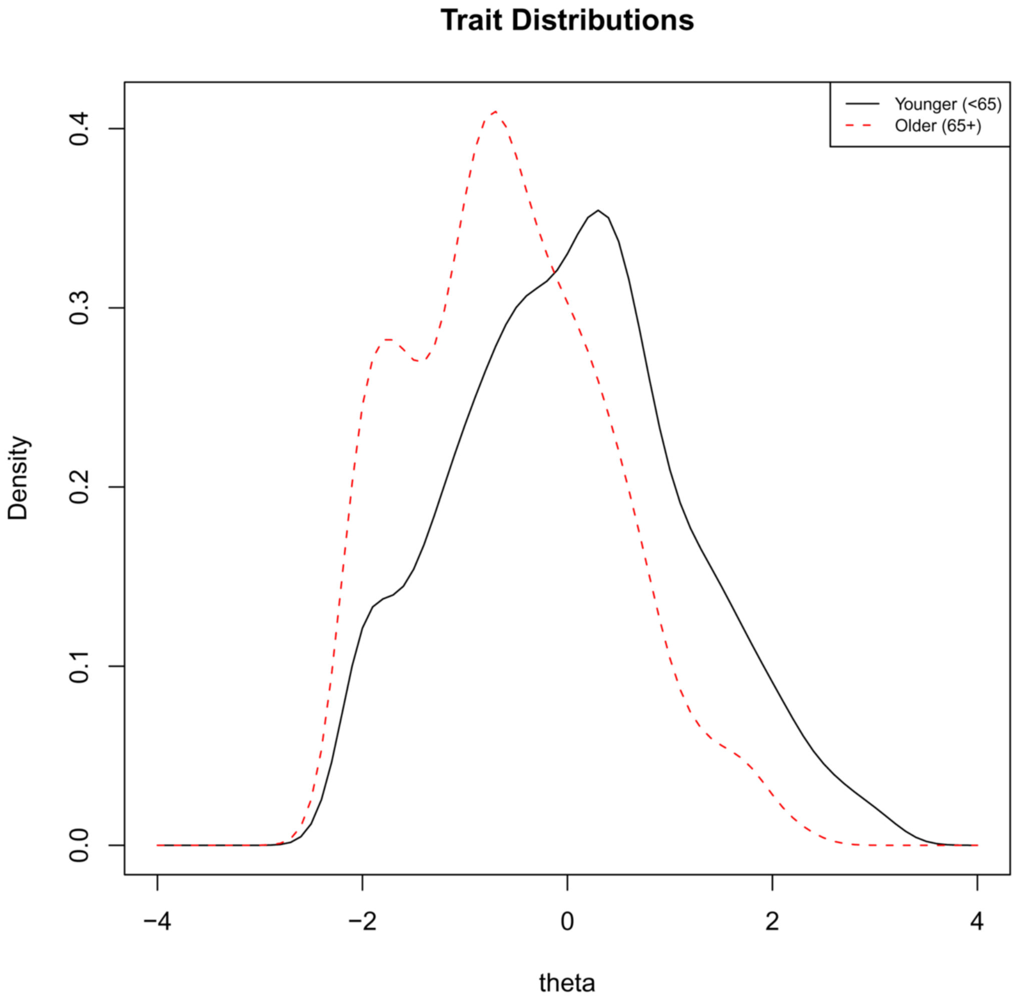

Figure 1.

Trait distributions – younger (< 65) vs. older (65 and up). Note: This graph shows smoothed histograms of the anxiety levels of older (dashed line) and younger (solid line) study participants as measured by the PROMIS Anxiety scale (theta). There is broad overlap in the distributions, though older individuals in general demonstrated lower levels of anxiety than younger individuals.

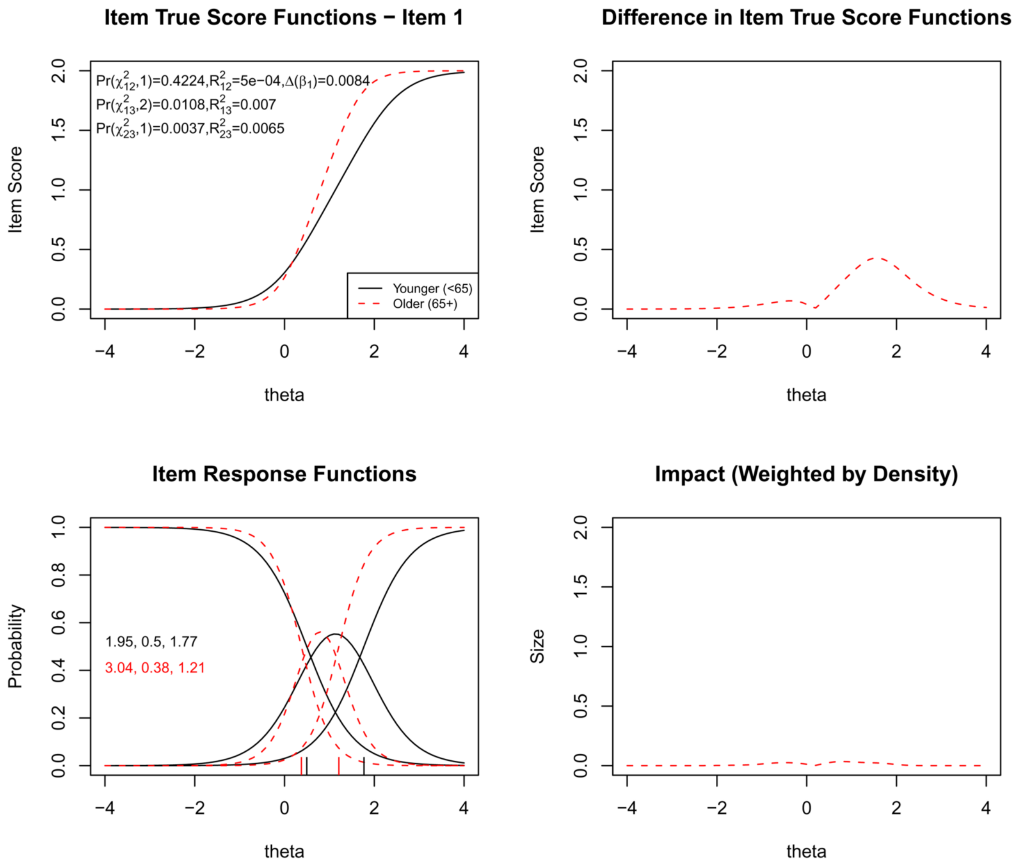

Figure 2.

Graphical display of the item "I felt fearful" which shows non–uniform DIF with respect to age. Note: This item retained three response categories (0, 1, and 2) from the original five–point rating scale after collapsing the top three response categories due to sparseness. The program by default uses a minimum of five cases per cell (the user can specify a different minimum) in order to retain each response category. The upper–left graph shows the item characteristic curves (ICCs) for the item for older (dashed curve) vs. younger (solid curve). The upper–right graph shows the absolute difference between the ICCs for the two groups, indicating that the difference is mainly at high levels of anxiety (theta). The lower–left graph shows the item response functions for the two groups based on the demographic–specific item parameter estimates (slope and category threshold values by group), which are also printed on the graph. The lower–right graph shows the absolute difference between the ICCs (the upper–right graph) weighted by the score distribution for the focal group, i.e., older individuals (dashed curve in Figure 1), indicating minimal impact. See text for more details.

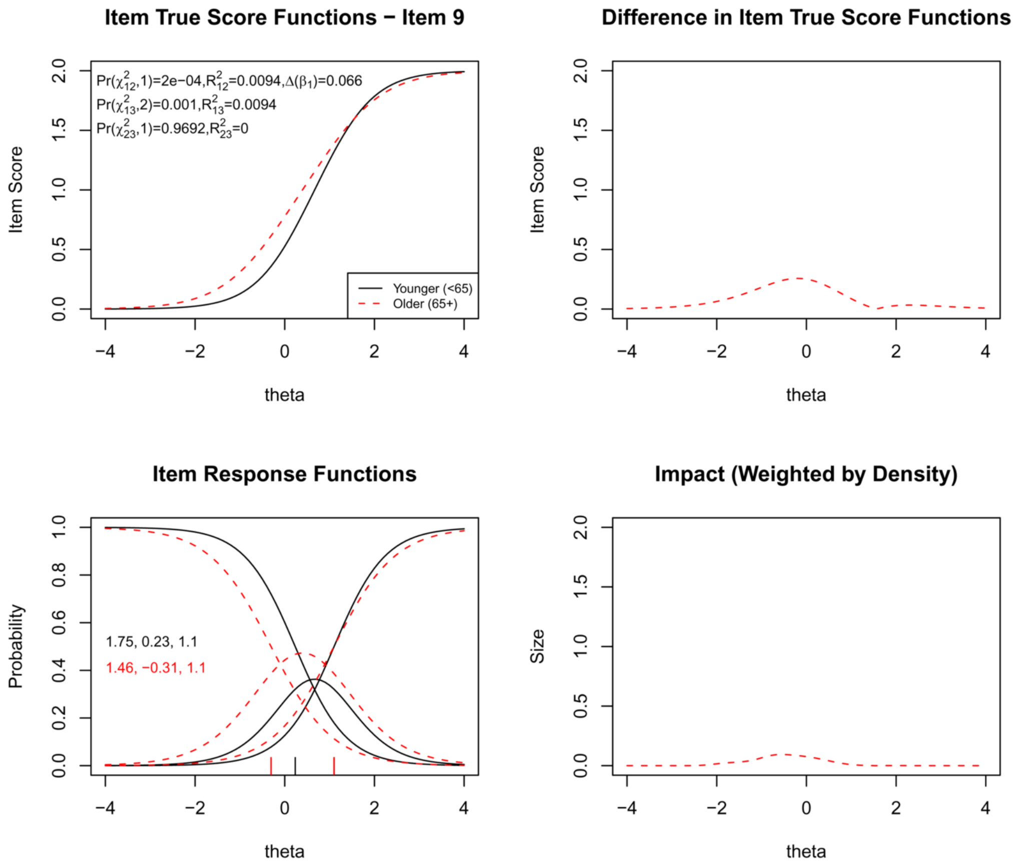

Figure 3.

Graphical display of the item “I was anxious if my normal routine was disturbed” which shows uniform DIF with respect to age. Note: See detailed comments accompanying Figure 2. Here the differences between younger and older individuals appear to be at lower anxiety levels.

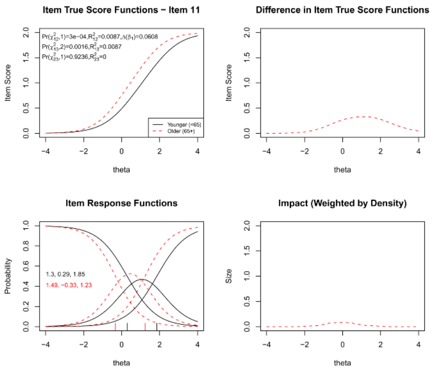

Figure 4.

Graphical display of the item “I was easily startled” which shows uniform DIF with respect to age. Note: See detailed comments accompanying Figure 2. Here the differences between younger and older individuals are across almost the entire spectrum of anxiety measured by the test.

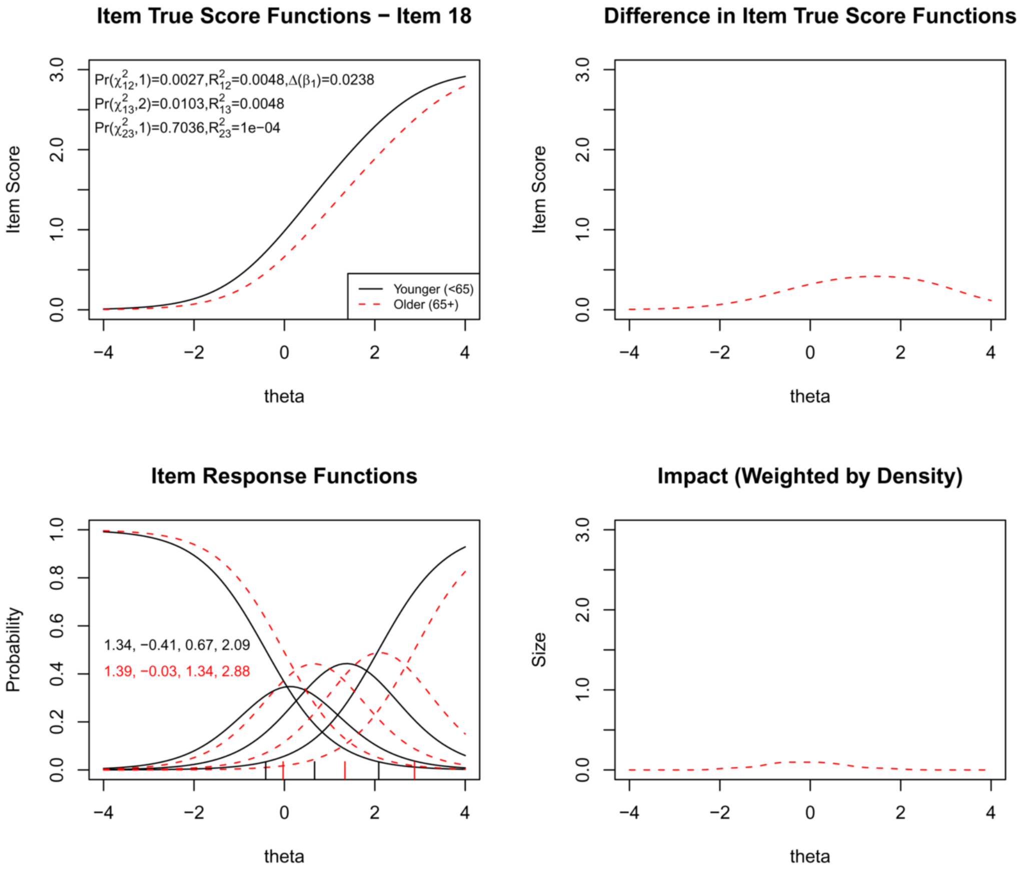

Figure 5.

Graphical display of the item “I worried about other people’s reactions to me”which shows uniform DIF with respect to age. Note: See detailed comments accompanying Figure 2.

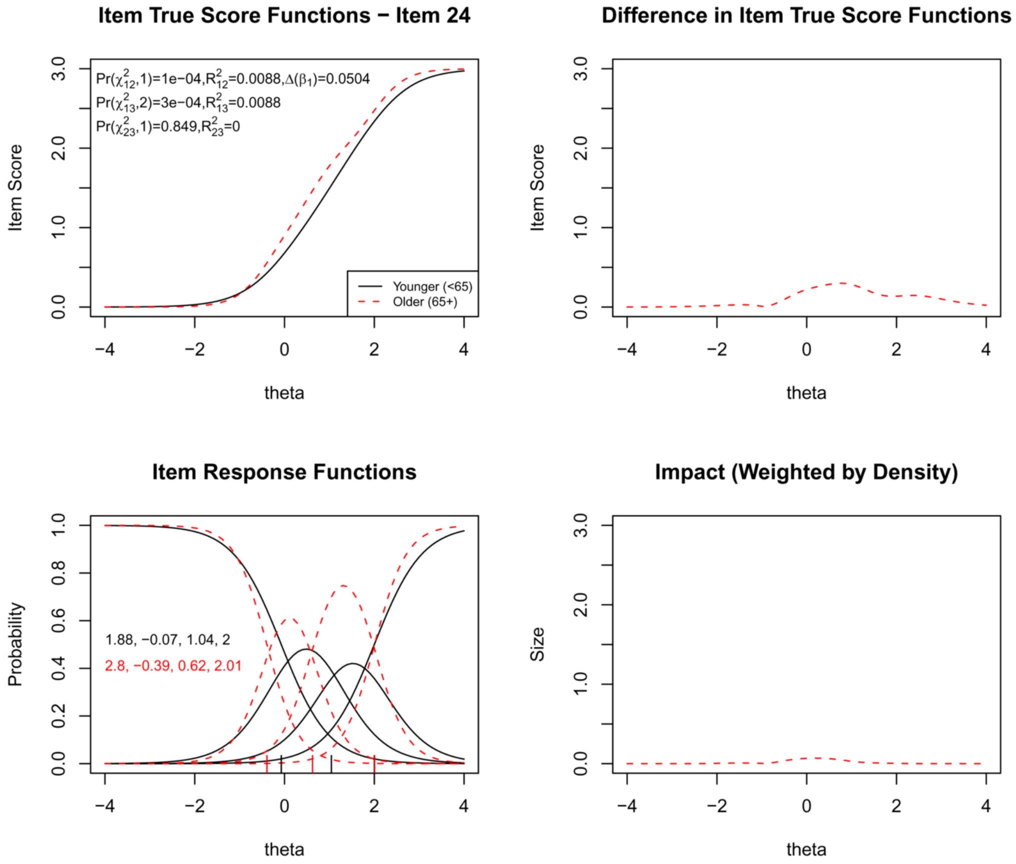

Figure 6.

Graphical display of the item “Many situations made me worry” displaying uniform DIF with respect to age. Note: See detailed comments accompanying Figure 2.

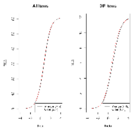

Figure 7.

Impact of DIF items on test characteristic curves. Note: These graphs show test characteristic curves (TCCs) for younger and older individuals using demographic–specific item parameter estimates. TCCs show the expected total scores for groups of items at each anxiety level (theta). The graph on the left shows these curves for all of the items (both items with and without DIF), while the graph on the right shows these curves for the subset of these items found to have DIF. These curves suggest that at the overall test level there is minimal difference in the total expected score at any anxiety level for older or younger individuals.

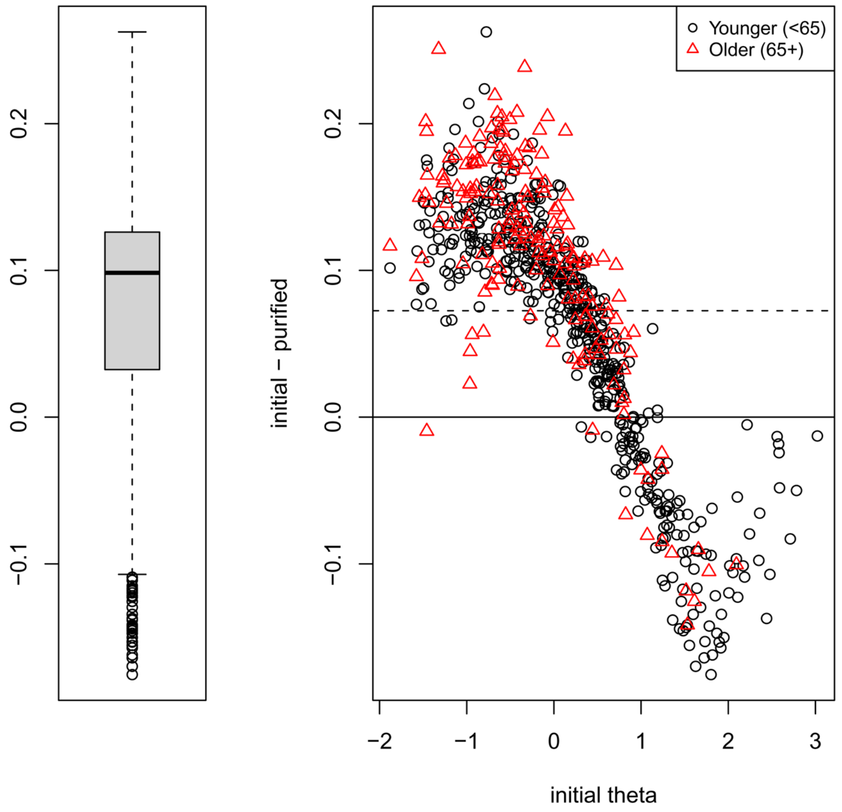

Figure 8.

Individual–level DIF impact. Note: These graphs show the difference in score between using scores that ignore DIF and those that account for DIF. The graph on the left shows a box plot of these differences. The interquartile range, representing the middle 50% of the differences (bound between the bottom and top of the shaded box), range roughly from +0.03 to +0.12 with a median of approximately +0.10. In the graph on the right the same difference scores are plotted against the initial scores ignoring DIF ("initial theta"), separately for younger and older individuals. Guidelines are placed at 0.0 (solid line), i.e., no difference, and the mean of the differences (dotted line). The positive values to the left of this graph indicate that in almost all cases, accounting for DIF led to slightly lower scores (i.e., naive score ignoring DIF minus score accounting for DIF > 0, so accounting for DIF score is less than the naive score) for those with lower levels of anxiety, but this appears to be consistent across older and younger individuals. The negative values to the right of this graph indicate that for those with higher levels of anxiety, accounting for DIF led to slightly higher scores, but this again was consistent across older and younger individuals.

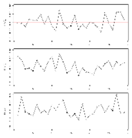

Figure 9.

Monte Carlo thresholds for χ^2 probabilities (1,000 replications). Note: The graphs show the probability values for each of the items (shown along the x–axis) associated with the 99th quantile (cutting the largest 1% over 1,000 iterations) of the χ^2 statistics generated from Monte Carlo simulations under the no DIF condition (data shown in Table 1). The lines connecting the data points are placed to show the fluctuation across items and not to imply a series. The horizontal reference line is placed at the nominal alpha level (0.01).

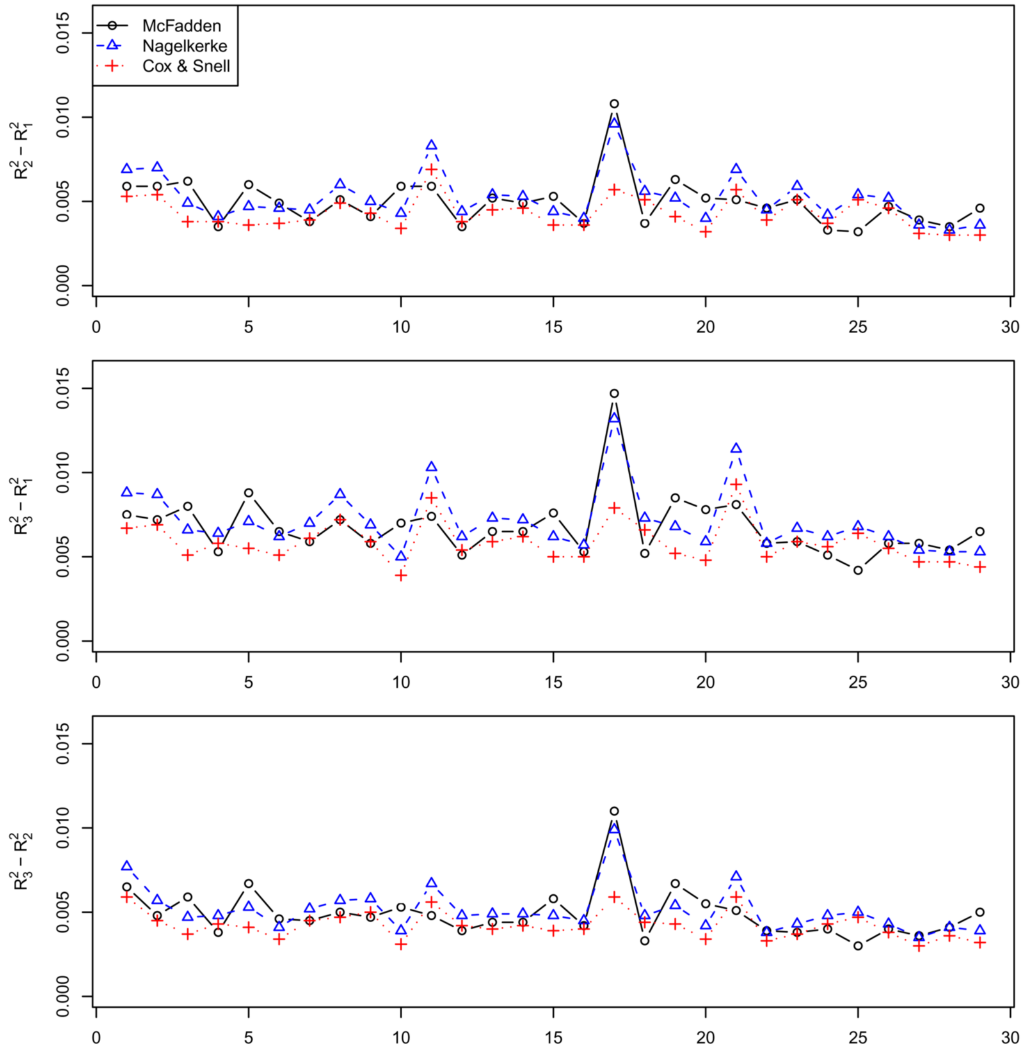

Figure 10.

Monte Carlo thresholds for pseudo R^2 (1,000 replications). Note: The graphs show the pseudo R^2 measures for each of the items (shown along the x–axis) corresponding to the 99th quantile (cutting the largest 1% over 1,000 iterations) generated from Monte Carlo simulations under the no DIF condition. The lines connecting the data points are placed to show the fluctuation across items and not to imply a series.

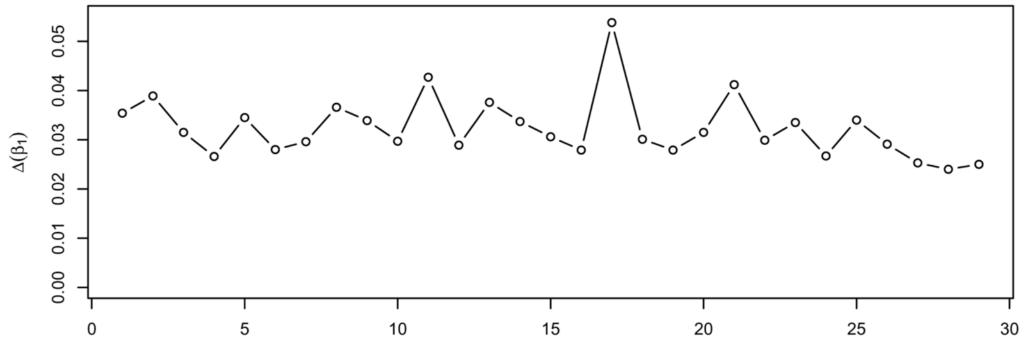

Figure 11.

Monte Carlo thresholds for proportional beta change (1,000 replications). Note: The graphs show the proportionate β_1 change measures for each of the items (shown along the x–axis) corresponding to the 99th quantile (cutting the largest 1% over 1,000 iterations) generated from Monte Carlo simulations under the no DIF condition. The lines connecting the data points are placed to show the uctuation across items and not to imply a series.

Table 1

Empirical threshold values from Monte Carlo simulations (`nr = 1000, alpha = 0.01`).

Item	no. cat.	Prob.			Cox & Snell			Nagelkerke			McFadden			%Δβ₁
		χ^2_{12}	χ^2_{13}	χ^2_{23}	R^2_{12}	R^2_{13}	R^2_{23}	R^2_{12}	R^2_{13}	R^2_{23}	R^2_{12}	R^2_{13}	R^2_{23}	$\%\Delta\beta_1$
1	3	.010	.016	.008	.005	.007	.006	.007	.009	.008	.006	.008	.007	.035
2	3	.009	.014	.018	.005	.007	.005	.007	.009	.006	.006	.007	.005	.039
3	3	.008	.009	.010	.004	.005	.004	.005	.007	.005	.006	.008	.006	.032
4	4	.012	.009	.009	.004	.006	.004	.004	.006	.005	.004	.005	.004	.027
5	3	.011	.008	.007	.004	.006	.004	.005	.007	.005	.006	.009	.007	.035
6	3	.011	.014	.014	.004	.005	.003	.005	.006	.004	.005	.007	.005	.028
7	3	.015	.010	.009	.004	.006	.005	.005	.007	.005	.004	.006	.005	.030
8	3	.009	.008	.010	.005	.007	.005	.006	.009	.006	.005	.007	.005	.037
9	3	.014	.013	.009	.004	.006	.005	.005	.007	.006	.004	.006	.005	.034
10	3	.008	.016	.014	.003	.004	.003	.004	.005	.004	.006	.007	.005	.030
11	3	.005	.008	.011	.007	.009	.006	.008	.010	.007	.006	.007	.005	.043
12	3	.017	.018	.014	.004	.005	.004	.004	.006	.005	.004	.005	.004	.029
13	3	.009	.013	.017	.005	.006	.004	.005	.007	.005	.005	.007	.004	.038
14	3	.007	.007	.009	.005	.006	.004	.005	.007	.005	.005	.007	.004	.034
15	3	.008	.008	.006	.004	.005	.004	.004	.006	.005	.005	.008	.006	.031
16	3	.014	.013	.008	.004	.005	.004	.004	.006	.005	.004	.005	.004	.028
17	2	.006	.005	.005	.006	.008	.006	.010	.013	.010	.011	.015	.011	.054
18	4	.009	.009	.014	.005	.007	.004	.006	.007	.005	.004	.005	.003	.030
19	3	.007	.007	.005	.004	.005	.004	.005	.007	.005	.006	.009	.007	.028
20	3	.010	.007	.008	.003	.005	.003	.004	.006	.004	.005	.008	.006	.032
21	3	.009	.006	.009	.006	.009	.006	.007	.011	.007	.005	.008	.005	.041
22	3	.007	.011	.013	.004	.005	.003	.005	.006	.004	.005	.006	.004	.030
23	3	.004	.009	.014	.005	.006	.004	.006	.007	.004	.005	.006	.004	.034
24	4	.016	.012	.009	.004	.006	.004	.004	.006	.005	.003	.005	.004	.027
25	4	.010	.014	.013	.005	.006	.005	.005	.007	.005	.003	.004	.003	.034

Item	no. cat.	Prob.			Cox & Snell			Nagelkerke			McFadden			
		χ^2_{12}	χ^2_{13}	χ^2_{23}	R^2_{12}	R^2_{13}	R^2_{23}	R^2_{12}	R^2_{13}	R^2_{23}	R^2_{12}	R^2_{13}	R^2_{23}	$\%\Delta\beta_1$
26	3	.006	.009	.010	.005	.006	.004	.005	.006	.004	.005	.006	.004	.029
27	3	.016	.013	.020	.003	.005	.003	.004	.005	.004	.004	.006	.004	.025
28	3	.016	.011	.009	.003	.005	.004	.003	.005	.004	.004	.005	.004	.024
29	3	.012	.012	.009	.003	.004	.003	.004	.005	.004	.005	.007	.005	.025
	Mean	.010	.011	.011	.004	.006	.004	.005	.007	.005	.005	.007	.005	.032
	SD	.004	.003	.004	.001	.001	.001	.001	.002	.001	.001	.002	.002	.006