HHS Public Access

Author manuscript

Psychol Test Assess Model. Author manuscript; available in PMC 2017 July 11.

Published in final edited form as:

Psychol Test Assess Model. 2016 ; 58(1): 79–98.

Differential item functioning magnitude and impact measures from item response theory models

Marjorie Kleinman[1] and **Jeanne A. Teresi**[2,3,4,5]

[2]New York State Psychiatric Institute

[3]Columbia University Stroud Center

[4]Research Division, Hebrew Home at Riverdale; RiverSpring Health

[5]Department of Geriatrics and Palliative Medicine, Weill Cornell Medical Center

Abstract

Measures of magnitude and impact of differential item functioning (DIF) at the item and scale level, respectively are presented and reviewed in this paper. Most measures are based on item response theory models. Magnitude refers to item level effect sizes, whereas impact refers to differences between groups at the scale score level. Reviewed are magnitude measures based on group differences in the expected item scores and impact measures based on differences in the expected scale scores. The similarities among these indices are demonstrated. Various software packages are described that provide magnitude and impact measures, and new software presented that computes all of the available statistics conveniently in one program with explanations of their relationships to one another.

Keywords

differential item functioning; magnitude; impact; item response theory

This paper describes magnitude and impact measures of item and scale level effects of differential item functioning (DIF) that are derived from item response theory (IRT) models, and specialized parameterizations of structural equation models (SEM).

The magnitude of DIF relates to item level effect sizes and refers to the degree of difference in item performance between or among groups, conditional on the trait assessed, denoted theta (θ). Magnitude has been defined as the weighted (by the trait distribution) group differences in the probability of an affirmative item response (Wainer, 1993). Magnitude measures are an essential component of examining DIF because reliance upon significance tests alone may result in identification of items with inconsequential DIF (Gómez-Benito, Dolores-Hidalgo, & Zumbo, 2013; Hambleton, 2006).

[1]*Correspondence concerning this article should be addressed to:* Marjorie Kleinman, M.S., New York State Psychiatric Institute, 1051 Riverside Drive, Unit 72, New York, NY, 10032, USA; kleinmam@nyspi.columbia.edu.

Impact refers to the influence of DIF on the scale score at the aggregate and individual level. In the context of IRT, aggregate DIF is evidenced by constructing and plotting differences in "test" response functions (Lord, 1980; Lord & Novick, 1968), also referred to as test characteristic curves (TCC) or expected scale score functions. Similar approaches are used in the context of SEM in which unadjusted and DIF adjusted means are compared. Individual-level impact is examined by comparing DIF adjusted and unadjusted trait (theta) estimates for individual respondents. These measures are described in more detail below.

True scores and expected scores

IRT-derived magnitude and impact measures described here are based on the notion of IRT true scores (Lord, Novick, & Birnbaum, 1968). The item true score function (also called the expected item score function or item characteristic curve) describes the relationship between the trait level and the person's expected value of the item score. A person's true score is his or her expected score, expressed in terms of probabilities for binary items and weighted probabilities for polytomous items. The test characteristic curve (expected scale score function) relates true scores (average expected scores) to theta.

DIF and expected scores

The use of expected scores to examine DIF magnitude for binary items was proposed by Wainer (1993), and expanded for polytomous items by Kim (see Kim, Cohen, Alagoz, & Kim, 2007). Raju, van der Linden and Fleer (1995) proposed a similar methodology to examine magnitude of DIF. These measures and/or related graphics are contained in several software packages described below: Differential Functioning of Items and Tests (DFIT; Oshima, Kushubar, Scott, & Raju, 2009; Raju et al., 2009); Item Response Theory for Patient Reported Outcomes (IRTPRO; Cai, duToit, & Thissen, 2009), logistic ordinal regression differential item functioning (lordif; Choi, Gibbons, & Crane, 2011), and a software package introduced here (MAGNITS). Additionally, Woods (2011) introduced the average unsigned difference (AUD), which is contained in two of the packages.

Effect size estimates and expected score calculation

Each respondent is posited to have two true (expected) scores, one as a member of the focal (studied) group and one as a member of the reference group. For each person in the focal group, the estimated theta is calculated so that the comparison groups are on the same scale (i.e., equated). Given the estimated theta for each person, his or her estimated true score on item *i* if a member of the reference group ($P_R(\theta)$) is calculated, and also the estimated true score if a member of the focal group ($P_F(\theta)$). For binary items, expected (true) scores equal the probability of item endorsement, conditional on trait (θ) level: $P_i(\theta_s)$. For example, for a binary item, $P_{iF}(\theta_s)$ is the expected score for individual *s* as a member of the focal (studied) group, reflecting the probability of a correct response to a test item.

Graded response model

For a polytomous item *i*, with K categories taking a graded response form, $P^*_{iK}(\theta)=\{1+\exp[-\alpha_i(\theta-\beta_{iK})]\}^{-1}$, the true score can be expressed as:

$$T_i(\theta) = \sum_{k=1}^{K_i} y_{ik} P_{ik}(\theta).$$

The expected score is thus the sum of the weighted probabilities of scoring in each of the possible categories for the item. For an item with 5 response categories, coded 0 to 4, for example, this sum would be:
$0 * [\mathrm{P_{i0}}(\theta_s)] + 1 * \mathrm{P_{i1}}(\theta_s) + 2 * \mathrm{P_{i2}}(\theta_s) + 3 * \mathrm{P_{i3}}(\theta_s) + 4 * \mathrm{P_{i4}}(\theta_s)$. For a detailed explication, see Collins, Raju, and Edwards (2000); Teresi et al. (2007).

Most of the effect size measures described here are based on examination of the differences in the expected item scores for two or more groups, as illustrated in Figure 1, panel B. For example, for item *i*, calculated is the average (expected value) of the squared difference between expected item scores for individuals as a member of the focal group and as a member of the reference group (see also Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006).

The precursors to these measures were various area and probability difference measures. The exact area methods compare the areas between the item characteristic functions estimated in different groups. Although a detailed explication of these measures is beyond the scope of this article, they are related to those described here (see also Raju, 1988; 1990; Teresi, Kleinman, & Ocepek-Welikson, 2000).

The signed and unsigned area differences for binary items are defined as follows (see Raju, 1990):

$$SA(\text{signed area}) = \hat{b}_2 - \hat{b}_1;$$
$$UA(\text{unsigned area}) = \left| \left[2(\hat{a}_2 - \hat{a}_1) / D\hat{a}_1\hat{a}_2 \right] \left[\ln\left\{ 1 + \exp\left[D\hat{a}_1\hat{a}_2 \left(\hat{b}_2 - \hat{b}_1 \right) / (\hat{a}_2 - \hat{a}_1) \right\} \right] - \left(\hat{b}_2 - \hat{b}_1 \right) \right|$$

The variance for the signed area is: $\sigma^2(SA) = \sigma^2\left(\hat{b}_2\right) + \sigma^2\left(\hat{b}_1\right)$;

where, $\sigma^2\left(\hat{b}_i\right) = I\hat{a}_i / \left(I\hat{a}_i I\hat{b}_i - I^2\hat{a}_i\hat{b}_i \right)$ and I is the information matrix.

The variance for the unsigned area is:

$$\sigma^2(UA) = B_1^2\sigma^2\left(\hat{b}_1\right) + B_2^2\sigma^2\left(\hat{b}_2\right) + A_1^2\sigma^2(\hat{a}_1) + A_2^2\sigma^2(\hat{a}_2) + 2B_1A_1\sigma\left(\hat{b}_1, \hat{a}_1\right) + 2B_2A_2\sigma\left(\hat{b}_2, \hat{a}_2\right),$$

where $B_1 = 1 - 2\exp(Y) / (1 + \exp(Y))$ and $B_2 = -B_1$

$$A_1 = 2/\hat{a}_1^2 \left\{ \hat{a}_1\hat{a}_2 \left(\hat{b}_2 - \hat{b}_1 \right) / (\hat{a}_2 - \hat{a}_1) \left(\exp(Y) / (1 + \exp(Y)) \right) - \ln\left[1 + \exp(Y) \right] / D \right\};$$
$$A_2 = (-\hat{a}_1^2 / \hat{a}_2^2) / A_1$$

and $Y = D\hat{a}_1\hat{a}_2\left(\hat{b}_2 - \hat{b}_1\right) / (\hat{a}_2 - \hat{a}_1)$. *D* is a scaling constant equal to 1.7 that places values of the logistic function to within 0.01 of those of the normal distribution function.

Wainer (1993) points out several problems with using area statistics to assess DIF. When ability distributions differ between reference and studied groups, area statistics may show large differences but affect very few subjects. Conversely, a small area difference may affect a relatively large number of people in the studied group. The area statistics do not take into account the distribution of the ability parameter within the focal group. Also, in a 3-parameter model the area differences cannot be computed precisely if the c (guessing) parameter differs in the two groups.

Wainer effect size estimates

Wainer (1993) proposed several effect size measures for binary items, denoted *T (1)* to *T (4).* *T (2)*, shown below is thus the sum of the differences $(\sum_{i=1}^{N}[P_R(\theta) - P_F(\theta)])$ across all members of the focal group.

Based on the estimated theta for each member of the focal group, for any given item, the true score for a subject is calculated based on the estimated parameters for the reference group ($T_R(\theta)$) and the true score is also calculated based on the estimated parameters for the focal group ($T_F(\theta)$). Let N_F denote the number of subjects in the focal group.

$$T(2) = \sum_{i=1}^{N}[T_R(\theta) - T_F(\theta)]$$
$$T(1) = T(2)/N_F$$

(In DFIT, this is called “Mean of D”. This can be either positive favoring the reference group or negative favoring the focal group.)

$$AUD = \sum_{i=1}^{N}|[T_R(\theta) - T_F(\theta)]|/N_F$$

(The AUD described below is the same as T (1) except that the absolute value of the difference is taken). This is also a graphic available in lordif (see Figure 2, lower right panel.)

$$T(4) = \sum_{i=1}^{N}[T_R(\theta) - T_F(\theta)]^2$$
$$T(3) = T(4)/N_F$$

(This is the same as non-compensatory DIF (NCDIF) index (Raju et al., 1995) described below). This is the sum of the unsigned (squared) differences weighted by the theta distribution for the focal group.

T (4) is the sum of the differences squared, resulting in an unsigned difference such that any difference, regardless of the direction is significant. Because *T (4)* is the sum of the squared differences and *T (3)* is the average squared difference, these indices capture the magnitude of non-uniform DIF, whereas *T (1)* and *T (2)* do not. Non-uniform DIF occurs when the probability of response is in a different direction for the reference and focal groups, at different levels of the latent ability (θ). For example, Black older persons may have a lower probability than White older persons of endorsing a depression item at low levels of the depression trait and higher probabilities of endorsement than White older persons at higher levels.

NCDIF

A method for quantification of the difference in the average expected item scores is the NCDIF index (Raju et al, 1995) used in DFIT (Oshima, et al., 2009; Raju et al., 2009). NCDIF is the average squared difference between the expected item scores for comparison groups, based on the actual distribution of thetas in the focal group. This is the same as Wainer's *T (3)* provided that one uses the actual distribution of estimated thetas in computing Wainer's statistics. These measures can be used with both binary and polytomous items and are thus appropriate for both Rasch (Rasch, 1980) and graded response (Samejima, 1969) models.

Prior to computations, it is necessary to equate the item parameters so that they are on the same scale for comparison groups. This is usually accomplished with a linear transformation of the item parameters for the focal (studied or targeted) group; however, see Chalmers, Counsell, and Flora (2016) for an alternative to linking. Scaling constants are defined and the linking constants are often computed using the test characteristic curve method (Stocking & Lord, 1983). Baker's Equate (Baker, 1995) program has typically been used for this procedure. (This procedure is explicated in more detail in the section on software below.) It is noted that just like other methods, it is necessary to provide iterative purification of the equating constants, after removing items with DIF (Huggins, 2014; Seybert & Stark, 2012).

NCDIF for item (i) is defined as the average squared difference between the true or expected scores for an individual (s) as a member of the focal group (F) and as a member of the reference group (R). For each subject in the focal group, two estimated scores are computed. One is based on the subject's ability estimate and the estimated item parameters for the focal group and the other based on the ability estimate and the estimated item parameters for the reference group. Each subject's difference score (d) is squared and summed for all subjects ($j = 1, N_F$) to obtain NCDIF.

$$NCDIF_{\mathrm{i}}=\left[\sum_{j=1}^{N_F}(ES_{siF}-ES_{siR})^{2}\right]/\mathrm{N_F}$$, which can also be expressed as $T(4)/N_F$, given above.

While chi-square tests of significance are available for NCDIF, these were found to be too stringent, over identifying DIF when sample sizes are larger. Cutoff values established based on simulations (Fleer, 1993; Flowers, Oshima, & Raju, 1999) can be used in the estimation of the magnitude of item-level DIF. For example, the cutoff values recommended by Raju are 0.006 for binary items, and 0.024, 0.054 and 0.096 for polytomous items with three, four and five response options (Raju, 1999). Because NCDIF is expressed as the average squared difference in expected scores for individuals as members of the focal group and as members of the reference group, the square root of NCDIF provides an effect size in terms of the original metric. Thus, for a polytomous item with three response categories, the recommended cutoff of 0.024 would correspond to an average absolute difference greater than 0.155 (about 0.16 of a point) on a three point scale (see Raju, 1999; Meade, Lautenschlager, & Johnson, 2007).

Because of the sensitivity of cutoff thresholds to the distribution of parameter estimates, simulations to derive cutoffs based on empirical distributions have been incorporated into the latest versions of software such as DFIT (Raju et al., 2005) and ordinal logistic regression (Choi et al., 2011). Recently tables have been developed for use in determining moderate and large DIF for the NCDIF statistic; however, they apply only to dichotomous items (Wright & Oshima, 2015). Simulations by Meade et al. (2007) resulted in the recommendation to use empirically derived DIF cutoff values. Seybert and Stark (2012) examined item parameter replication (IPR) methods used in the newer software with testwide cutoffs developed by Flowers et al. (1999) based on simulations. It was found that it is critical to perform iterative linking involving purification of the equating constants in order to maintain a balance between power and type 1 error (excess DIF detection). The findings were that test-wide critical values were more powerful than IPR in detecting DIF, but at the expense of inflated type 1 error. However, under iterative linking procedures, the test-wide critical values given above appear to work sufficiently well to continue their use.

AUD

The average unsigned difference is the absolute value of the difference between the expected item response functions, weighted by the focal (studied) group distribution. The differences can be evaluated at various quadrature (theta) points (Woods, 2011). However, in many instances, evaluating the differences at theta points that reflect the observed densities in the current sample may provide a more accurate reflection of the magnitude of DIF for the data investigated.

For the AUD the absolute value of the difference between the expected item scores is calculated.

$$AUD = \sum_{i=1}^{N} |[T_R(\theta) - T_F(\theta)]| / N_F .$$

The AUD is the same as *T (1)* except that it is the *absolute value of the differences* across subjects that is summed, and divided by N. When the AUD is close to the value of *T (1)*, this is an indication of uniform DIF; that is, the probability of response is consistently higher for either the reference or focal group across all levels of the latent ability *θ*. For this reason it is helpful to report both *T (1)* and AUD in order to investigate instances of non-uniform DIF, in which case, *T (1)* and AUD could differ substantially.

Wainer's *T (2)* is the sum of the differences (both positive and negative), so that positive and negative differences can potentially cancel each other out. Thus it is possible to have a fairly large value for AUD, while *T (2)* could be close to zero in the case of nonuniform DIF. The AUD always takes a positive value, whereas *T (2)* and *T (1)* can be negative. Kim (2000) suggests that when the absolute value of *T (1)* is greater than .10, the item requires 'close examination', but the basis for this claim is unclear. Based on empirical data shown in the examples in this series of articles, this cutoff may be overly conservative.

Impact measures

IRT severity parameters are on the same (usually) z-score metric, as the latent construct, so that a severity or location (*b*) parameter difference of 0.50 represents one half standard deviation on the trait estimate (*θ*). Thus, effect sizes at the item level have a relationship to impact measures at the scale level (Steinberg & Thissen, 2006). Several impact measures based on IRT examine the group differences in the area between the expected scale score functions (see Figure 1, panel B and Figure 3).

The Differential Test Functioning (DTF) index (Raju et al., 1995) is a summary measure of these differences in expected scale (total test) scores that incorporates a weight, and reflects the aggregated net impact. The DTF is the sum of the item-level compensatory DIF indices, and as such reflects the results of DIF cancellation. Stark, Chernyshenko, and Drasgow (2004) describe a similar effect size measure, Differential Test Functioning R (DTFR), a total test score statistic. DTFR measures the difference between groups in the relationship of θ to the expected observed scores.

Defining *θ* as the estimated theta for a given subject in the focal group, $TCC_{EF}(\theta)$ is that subject's expected total test score based on the equated focal (EF) group parameter estimates of *a*'s and *b*'s for all items in the test. Similarly, $TCC_R(\theta)$ represents the same subject's expected total test score based on the reference group parameter estimates of *a*'s and *b*'s. Then the DTFR value in raw score points can be obtained by integration, as shown in Stark et al. (2004, pg. 499):

$DTFR = \int [TCC_R(\theta) - TCC_{EF}(\theta)] f_F(\theta)\, d(\theta)$, where $f_F(\theta)$ is the ability density for the focal group, which is assumed to be normally distributed, and TCC_{EF} is the expected total test score for the focal group based on focal group equated item parameters. TCC_R is the expected total test score based on reference group parameters.

The proposed calculation of DTFR described below is not based on the assumption that the ability density for the focal group is normally distributed; instead statistics are based on the actual distribution of estimated thetas in the focal group (after equating). This measure assesses the expected impact of DIF on scores in absolute group differences between item true-score functions and density-weighted differences between groups. The latter, shown below, adjusts for the actual distribution of individuals; if few respondents are located at the point where the differences are greatest, the weighted impact will be less.

For each member of the focal group, the expected total test score is calculated based on estimated parameters for the reference group and then based on estimated parameters for the focal group.

Then, $DTFR=\sum_{i=1}^{N}\left[TCC_{R}\left(\theta\right)-TCC_{EF}\left(\theta\right)\right]/N_{F}$, where N_F = the number of subjects in the focal group.

A positive value for DTFR is indicative of DIF impact against the focal group, and a negative value in favor of the focal group. A DTFR value represents the number of raw score points that are due to scale level DIF.

To obtain a measure of magnitude of effect for DTFR, one can divide DTFR by the standard deviation for the focal group's observed scores or alternatively, the standard deviation of the expected scale scores. The MAGNITS program described below computes the standard deviation of the expected scale scores because DTFR is based on the difference in expected scale scores. Because this method yields a slightly smaller standard deviation, the effect size (*d*) would be slightly larger (Stark et al., 2004).

$$d_{DTF}=DTFR/SD_{F}$$

This statistic is closely related to DTF (Raju et al., 1995). Instead of the above,

$$DTF=\sum_{i=1}^{N}\left[TCC_{R}\left(\theta\right)-TCC_{EF}\left(\theta\right)\right]^{2}/N_{F}.$$

Because each difference is squared, the positive and negative values of the differences do not cancel each other out, which can happen with the DTFR statistic described above in the case of non-uniform DIF. It is because of this difference between DTFR and Raju's DTF that in the case of DTFR, individual items may show DIF but at the scale level DTFR may be quite small. Egberink, Meijer and Tendeiro (2015) point out that if the total test score is the main concern, a test statistic such as DTFR will only be large if many items in the scale exhibit uniform DIF in the same direction; that is, favoring either the reference or the studied group.

Individual impact

The impact measures just described are all at the aggregate or group level rather than the individual level. The impact on specific individuals rather than on the group as a whole can also be examined. Individual impact can be assessed through an examination of changes in theta estimates with and without adjustment for DIF. The unadjusted thetas are produced from a model with all item parameters set equal for the two groups. The adjusted thetas are produced from a model with parameters that showed DIF based on the IRT results estimated separately (freed) for the groups. The capacity to fix and free parameters based on DIF, and compare theta estimates is incorporated into software packages such as IRTPRO (Cai et al., 2009). This method permits comparisons of trait measure estimates that are DIF free (adjusted) to those with parameters estimated without DIF adjustment. This methodology has been used by several authors to examine the individual impact of DIF (e.g., Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Teresi et al., 2009).

Crane and colleagues (2007) used a similar method in calculating the difference between naïve scores that ignore DIF and scores that account for DIF to examine cumulative impact of DIF on individual participants. The distribution of these difference scores is then examined; for individual-level DIF impact, a box-and-whiskers plot of the difference scores is constructed. The differences due to DIF are plotted against the median standard error of measurement (SEM). Differences larger than that value are termed salient individual-level DIF impact (see Figure 4).

Software

DFIT

The DFIT methodology (Flowers et al., 1999; Raju, 1999; Oshima, & Morris, 2008; Raju, et al., 1995; Raju et al., 2009) permits examination of the magnitude of the gap between the ICCs (in this case boundary response functions) for two or more groups, such as illustrated in Figure 1. Non-Compensatory DIF (NCDIF) is an effect size measure that is weighted by the focal group density such that more weight is given to differences in the region of the trait with the highest frequency in the targeted group. As reviewed above, because simulation studies found over-identification of DIF with the use of Chi-square tests, cutoff values are used instead to identify DIF (see Morales, et al., 2006). DFIT yields both magnitude and impact measures.

lordif

Psychometric software is available in R (Rizopoulus, 2006, 2009). The R package lordif (Choi et al., 2011), similar to DFIT is based on the notion of true (expected) scores, from which various DIF magnitude and impact measures are derived. lordif uses *mirt* in R (Chalmers, 2012; see also https://cran.r-project.org/web/packages/mirt/index.html) to obtain IRT item parameter estimates for the Graded Response Model (Samejima, 1969) or the Generalized Partial Credit Model (Muraki, 1992), and the *Design* package for OLR. The R-Project website is https://www.r-project.org/. The link is: http://watson.nci.nih.gov/cran_mirror/.lordif is free software and can be found under “Packages” at the following link:

https://cran.r-project.org/web/packages/lordif/index.html. Figures 2–4 provide illustrations from lordif.

DIF magnitude measures—One measure shown in the lower right panel of Figure 2 is similar to the non-compensatory DIF test of Wainer (1993). However, the metric in this panel (shown for a range of theta values) is summed, and is the AUD described by Woods (2011) as the unsigned difference between the true scores for the comparison groups weighted by the studied (focal) group density. The unweighted unsigned differences appear in the upper right panel of Figure 2. As shown in Figure 2, the expected item scores (upper left panel) and category response functions (lower left panel) are also presented.

DIF aggregate impact measures—lordif also shows the graphic of the expected item and test score functions for groups, including differences for items with DIF. Expected test score plots show the expected total (sum scores) for groups for different theta levels (See Figure 3.)

DIF individual impact—Individual theta scores are calculated after fixing and freeing parameters for items without and with DIF respectively. Graphics are displayed (see Figure 4). Theta estimates by group before and after accounting for DIF and median differences before and after DIF adjustment (fixing and freeing parameters based on DIF) are plotted. A dotted line shows the mean difference between the initial and DIF-adjusted theta estimates.

Just as with most such software, the Stocking-Lord (1983) equating procedure identifies the equating constants to place the IRT item parameters for groups on the same metric. The items without DIF are used as anchors (see below). The software package, lordif also allows the user to designate a specific subset of items s an anchor set.

MAGNITS

This program (available upon request from the authors) computes item-level magnitude and scale level impact measures of DIF (see Figure 5). The magnitude measures are as follows: T (1) through T (4) (Wainer, 1993); NCDIF (Raju et al., 1995) and AUD (Woods, 2011). The magnitude measures are all closely related to each other. Impact measures are DTF (Raju et al) and DTFR from Stark et al. (2004).

As discussed above, prior to running this program, it is necessary to equate parameters and theta estimates for the focal group so that they are on the same metric scale as the reference group. There are several methods for doing so. Baker's EQUATE (Baker, 1995) software, which employs the test characteristic curve method of Stocking and Lord (1983) is the most popular method of equating, and is used in Magnits; lordif also uses the Stocking and Lord (1983) procedure, which is explicated as follows, based on Teresi et al. (2000; pg 1663–1664).

In order to compare the a_i's (denoted a's) and the b_i's (denoted b's) of one group with those of another on the same set of items, a linear transformation of the a's and b's for the second group (usually the focal group) is performed so that the parameters for both groups are on the same scale. Two scaling constants, α and β, are defined : $b_i^*=\alpha b_i+\beta$, where b_i is the b for

the item in the second group and b_i^* is the equated *b* and $a_i^*=a_i/\alpha$, where a_i is the *a* value for item *i* in the second group and a_i^* is the equated *a* value.

More than two groups can be compared by equating each group to the reference group, determining the constants α and β for each of the two *or more* groups. The constants α and β then can be used to place the ability values of the subjects in group 2 on the same scale as the ability values for group 1. Equated θ's for group 2 are calculated in a fashion similar to that used for the *b*'s: $\theta_j^*=\alpha\theta_j+\beta$, where θ_j is the original θ, and θ_j^* is the equated θ for individual *j*.

The linking constants α and β are computed using the characteristic curve method (Stocking & Lord, 1983). If t_{1i} is the estimated true score of individual *i* of estimated ability level θ_i in group 1, and t_{2i} is the estimated true score of the same individual *i* if a member of group 2, then α and β are chosen to minimize the difference between t_{1i} and t_{2i} for all members of group 1. The function to be minimized is $f=1/N\sum_{i=1}^{N}(t_{1i}-t_{2i})^2$, where *N* is the number of subjects in group 1.

It is noted that such linking algorithms may result in some error in estimation as contrasted with simultaneous linking in which parameter estimation and DIF testing is conducted holding constant the latent scale across groups (Woods, Cai, & Wang, 2013). As presented above, many magnitude measures such as those in this program and lordif rely on equating algorithms to link the groups on a common latent trait metric. DIF in the equating anchor set could compromise this process (Dorans, 2004). Recent simulations showed that directional DIF favoring one group over the other had a greater impact on equating (Huggins, 2014). While the Stocking and Lord equating method used in MAGNITS and lordif was more robust to the effects of DIF than an alternative method, in general and particularly under conditions of group differences in the mean trait level, as often occurs in health and mental health-related applications, DIF in the anchor set can affect equating (Huggins, 2014). Thus, purification of equating constants and careful selection of anchor items for each subgroup comparison is important. Such purification can be conducted using most software reviewed here.

Summary and conclusions

This paper is meant to orient the reader to magnitude and impact measures associated with IRT-based methods for DIF detection. Magnitude measures are an essential part of DIF detection because of the need to avoid false positives (Seybert & Stark, 2012), particularly in an environment in which items have been studied carefully and subjected to qualitative and quantitative analyses prior to DIF detection. It is desirable to identify and flag only items with salient DIF. Impact is essential at the aggregate level to determine if items identified with DIF have an appreciable impact at the scale level. However, also important is an examination of DIF at the level of the person.

Individual-level DIF impact may be important to clinicians who wish to know how different an unadjusted score might be from a score that accounted for DIF for a particular person. An

example of this is shown, using the Patient Reported Outcomes Measurement Information System® (PROMIS®) anxiety item bank data (Choi et al., 2011), and in the depression item bank (Teresi, et al., 2009). Additional examples, using depression, anxiety and cognitive measures are found in the articles in this two-part series.

DIF magnitude and impact analyses are conducted to help answer the question of what difference is meaningful and makes a practical difference. Recent work on effect sizes in the context of DIF is presented in Stark et al. (2004); Steinberg and Thissen (2006); and Kim et al. (2007); and Seybert and Stark (2012). More work is needed in order to determine optimal cutoff values. For example, an effect size measure for NCDIF has been proposed recently (Wright & Oshima, 2015); however, only binary item response models were included. Most recently, Chalmers et al. (2016) have developed differential test functioning statistics for both polytomous and binary data that consider sampling variation and are not reliant upon cutoff scores. Empirically derived thresholds based on Monte Carlo simulations to detect optimal cutoffs for the sample investigated were embedded into lordif for various chi-square tests associated with ordinal logistic regression, based on a latent conditioning variable. Most software does not permit such calculations. In the context of DFIT, test-wide thresholds such as those developed by Flowers et al., (1999) may be sufficiently accurate (Seybert & Stark, 2012). Further work is needed to compare their performance with that of other methods of threshold derivation.

Acknowledgments

Partial funding for these analyses was provided by the National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971 (PI: A. Potosky, C. Moinpour) and by the National Institute on Aging, 1P30AG028741-01A2 (PI: A. Siu).

The authors thank Seung Choi, Ph.D. for his helpful comments on an earlier version of this manuscript.

References

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

Cai, L., duToit, SHL., Thissen, D. IRTPRO: flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago: Scientific Software International; 2009.

Chalmers RP. Mirt. A multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012; 48:1–29. DOI: 10.18637/jss.v048.i06

Chalmers RP, Counsell A, Flora DB. It might not make a big DIF: improved differential test statistics that account for sampling variability. Educational and Psychological Measurement. 2016; 76:114–140. DOI: 10.1177/0013164415584576

Choi SW, Gibbons LE, Crane PK. lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression / item response theory and Monte Carlo simulations. Journal of Statistical Software. 2011; 39:1–30. DOI: 10.18637/jss.v039.i08

Collins WC, Raju NS, Edwards JE. Assessing differential item functioning in a satisfaction scale. Journal of Applied Psychology. 2000; 85:451–461. DOI: 10.1037//0021-9010.85.3.451 [PubMed: 10900818]

Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, Teresi JA. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Quality of Life Research. 2007; 16:69–84. DOI: 10.1007/s11136-007-9185-5 [PubMed: 17554640]

Dorans NJ. Using subpopulation invariance to assess test score equity. Journal of Educational Measurement. 2004; 41:43–68. DOI: 10.1111/j.1745-3984.2004.tb01158.x

Egberink IJL, Meijer RR, Tendeiro JN. Investigating measurement invariance in computer-based personality testing: the impact of using anchor items on effect size indices. Educational and Psychological Measurement. 2015; 75(1):126–145. DOI: 10.1177/0013164414520965

Fleer PF. A Monte Carlo assessment of a new measure of item and test bias. Illinois Institute of Technology. Dissertation Abstracts International. 1993; 54(04B):2266.

Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999; 23:309–326. DOI: 10.1177/01466219922031437

Gómez-Benito J, Dolores-Hidalgo M, Zumbo BD. Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items. Educational and Psychological Measurement. 2013; 73:875–897. DOI: 10.1177/0013164413492419

Hambleton RK. Good practices for identifying differential item functioning. Medical Care. 2006; 44(Suppl. 11):S182–S188. DOI: 10.1097/01.mlr.0000245443.86671.c4 [PubMed: 17060826]

Huggins AC. The effect of differential item functioning in anchor items on population invariance of equating. Educational and Psychological Measurement. 2014; 74:627–658. DOI: 10.1177/0013164413506222

Kim S, Cohen AS, Alagoz C, Kim S. DIF detection and effect size measures for polytomously scored items. Journal of Educational Measurement. 2007; 44(2):93–116. DOI: 10.1111/j.1745-3984.2007.00029.x

Kim, S-H. An investigation of the likelihood ratio test, the Mantel test, and the generalized Mantel-Haenszel test of DIF. Paper presented at the annual meeting of the American Educational Research Association; New Orleans, LA. Apr. 2000

Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential functioning of the Beck Depression Inventory in late-life patients: use of item response theory. Psychology and Aging. 2002; 17(3):379–391. [PubMed: 12243380]

Lord, FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.

Lord, FM., Novick, MR. Statistical theories of mental test scores. Reading Massachusetts: Addison-Wesley Publishing Company Inc.; 1968. (with contribution by A. Birnbaum)

Meade A, Lautenschlager G, Johnson E. A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. Applied Psychological Measurement. 2007; 31:430–455. DOI: 10.1177/0146621606297316

Morales LS, Flowers C, Gutierrez P, Kleinman M, Teresi JA. Item and scale differential functioning of the Mini-Mental State Exam assessed using the Differential Item and Test Functioning (DFIT) framework. Medical Care. 2006; 44(11, Suppl. 3):S143–S151. DOI: 10.1097/01.mlr.0000245141.70946.29 [PubMed: 17060821]

Oshima, TC., Kushubar, S., Scott, JC., Raju, NS. DFIT8 for Window User's Manual: differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation; 2009.

Oshima TC, Morris SB. Raju's Differential Functioning of Items and Tests (DFIT). Educational Measurement Issues and Practice. 2008; 27:43–50. DOI: 10.1111/j.1745-3992.2008.00127.x

Raju NS. Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement. 1990; 14:197–207. DOI: 10.1177/014662169001400208

Raju, NS. DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer program]. Chicago: Illinois Institute of Technology; 1999.

Raju NS. The area between two item characteristic curves. Psychometrika. 1988; 53:495–502. DOI: 10.1007/BF02294403

Raju NS, Fortmann-Johnson KA, Kim W, Morris SB, Nering ML, Oshima TC. The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. Applied Psychological Measurement. 2009; 33:133–147. DOI: 10.1177/0146621608319514

Raju, NS., Oshima, TC., Wolach, A. Differential functioning of items and tests (DFIT): dichotomous and polytomous [Computer Program]. Chicago: Illinois Institute of Technology; 2005.

Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995; 19:353–368. DOI: 10.1177/014662169501900405

Rasch, G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1980. (original work published in 1960).

Rizopoulus D. *Ltm*: an R package for latent variable modeling and item response theory analyses. Journal of Statistical Software. 2006; 17:1–25. DOI: 10.18637/jss.v017.i05

Rizopoulus, D. Ltm: Latent Trait Models under IRT. 2009. http://cran.rproject.org/web/packages/ltm/index.html

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 34:100–114. DOI: 10.1002/j.2333-8504.1968.tb00153.x

Seybert J, Stark S. Iterative linking with the differential functioning of items and tests (DFIT) Method: comparison of testwide and item parameter replication (IPR) critical values. Applied Psychological Measurement. 2012; 36:494–515. DOI: 10.1177/0146621612445182

Stark S, Chernyshenko OS, Drasgow F. Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? Journal of Applied Psychology. 2004; 89:497–508. DOI: 10.1037/0021-9010.89.3.497 [PubMed: 15161408]

Steinberg L, Thissen D. Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. Psychological Methods. 2006; 11:402–415. [PubMed: 17154754]

Stocking ML, Lord FM. Developing a common metric in item response theory. Applied Psychological Measurement. 1983; 7:201–210. DOI: 10.1177/014662168300700208

Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Statistics in Medicine. 2000; 19:1651–1683. DOI: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H [PubMed: 10844726]

Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Cella D. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. Quality Life Research. 2007; 16:43–68. DOI: 10.1007/s11136-007-9186-4

Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke J, Crane PK, Jones RN, Cella D. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly. 2009; 51(2):148–180. DOI: 10.1007/978-94-007-0753-5_728 [PubMed: 20336180]

Wainer, H. Model-based standardization measurement of an item's differential impact. In: Holland, PW., Wainer, H., editors. Differential Item Functioning. Hillsdale NJ: Lawrence Erlbaum Inc; 1993. p. 123-135.

Woods CM. DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH tests and IRTLRDIF when the latent distribution is nonnormal for both groups. Applied Psychological Measurement. 2011; 35:145–164.

Woods CM, Cai L, Wang M. The Langer-improved Wald test for DIF testing with multiple groups: evaluation and comparison to two-group IRT. Educational and Psychological Measurement. 2013; 73:532–547. DOI: 10.1177/0013164412464875

Wright KD, Oshima TC. An effect size measure for Raju's Differential Functioning for Items and Tests. Educational and Psychological Measurement. 2015; 75:338–358. DOI: 10.1177/0013164414532944

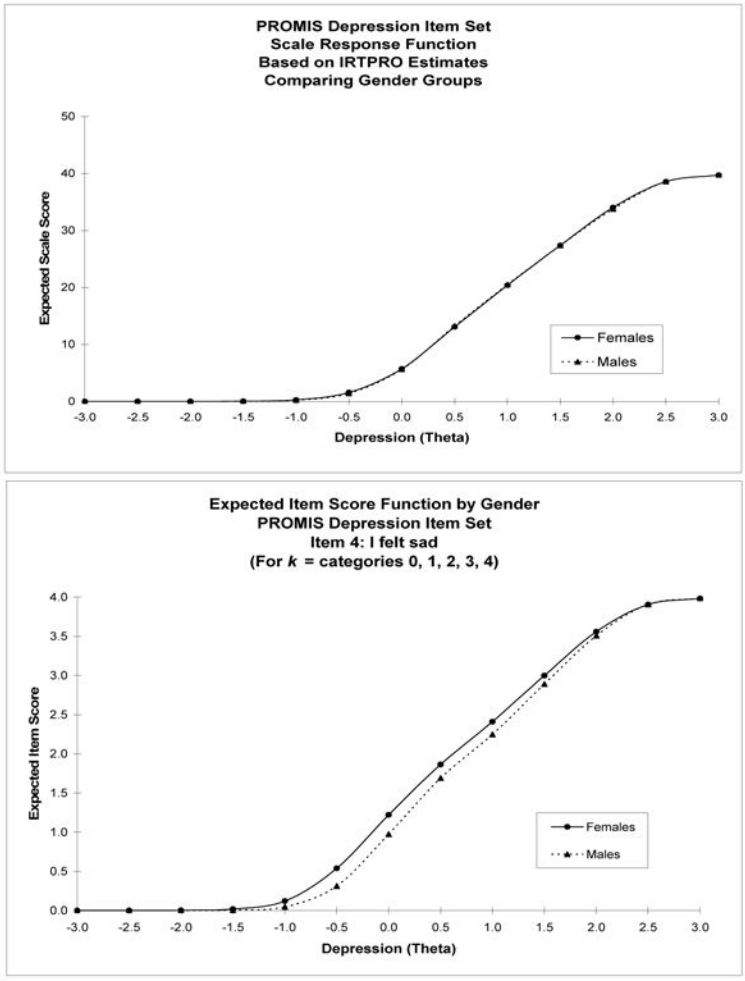

Figure 1.

Patient Reported Outcomes Measurement Information System® (PROMIS®) depression short form item set: Expected scale and item scores for gender subgroups from which differential item functioning magnitude and impact measures can be constructed

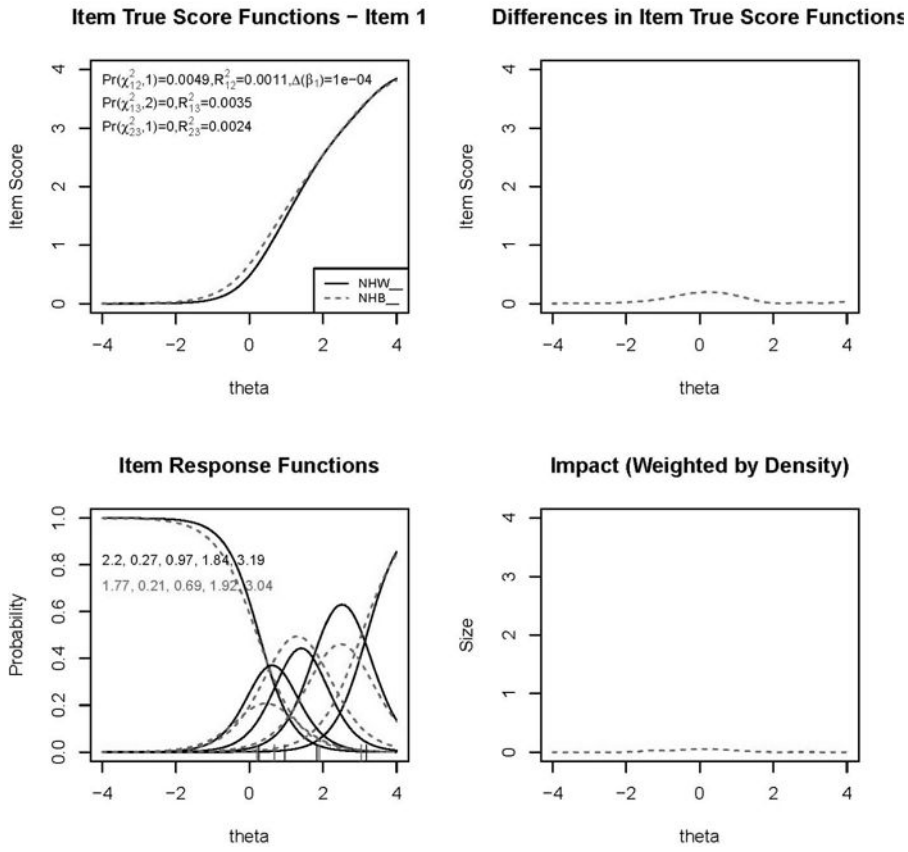

Figure 2.

Example graphics depicting magnitude of differential item functioning (from lordif) comparing non-Hispanic Blacks (NHB) and non-Hispanic Whites (NHW)

Expected item scores; category response functions; differences in expected scores and density weighted differences.

(The latter are similar to Wainer's (1993) magnitude measures and are unsigned differences weighted by the theta distributions of the focal group) (lordif depression NHB vs NHW)

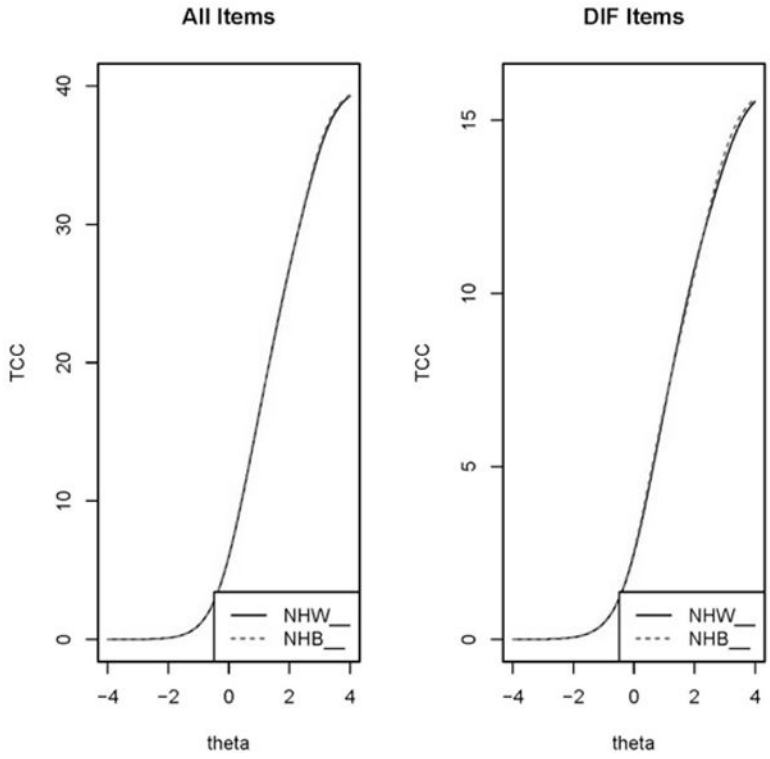

Figure 3.

Graph depicting differential item functioning impact with all items included and with only items with DIF included (from lordif) comparing non-Hispanic Blacks (NHB) and non-Hispanic Whites (NHW)

Aggregate DIF Impact:

Expected Scale Scores for groups at each theta level (Test Characteristic Curves) (lordif depression NHB vs NHW)

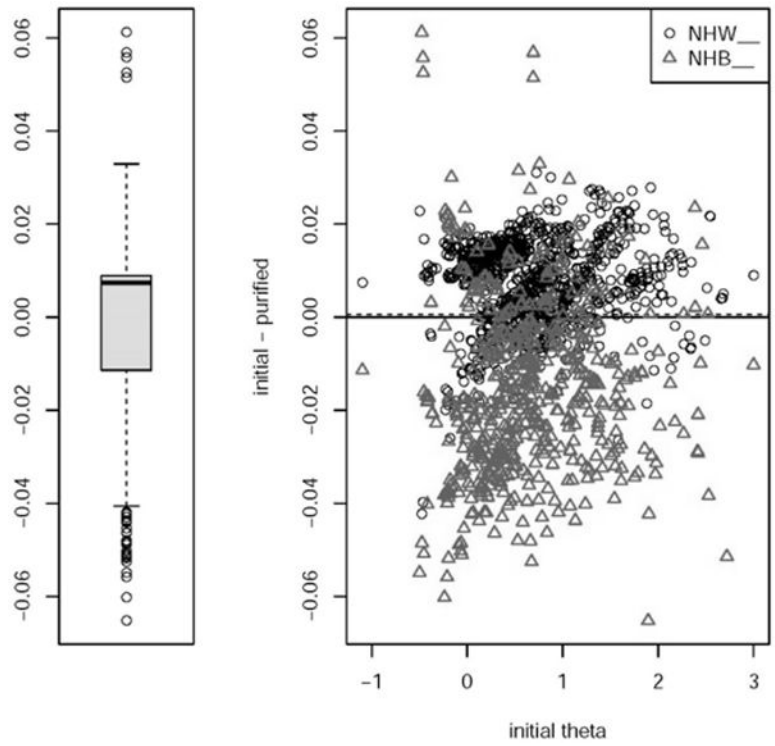

Figure 4.

Graphs depicting individual-level differential item functioning impact (from lordif) comparing non-Hispanic Blacks (NHB) and non-Hispanic Whites (NHW)

Individual Impact (Theta estimates by group before and after accounting for DIF and median differences before and after DIF adjustment (fixing and freeing parameters based on DIF)

*Dotted line is the mean difference between the initial and DIF-adjusted theta estimates (lordif depression NHB vs NHW)

PROMIS III Depression Scale by ethnicity NHW is reference, NHB is focal

REFERENCE GROUP ITEM PARAMETERS

1	4.214	.454	.941	1.619	2.462
2	4.878	.438	.904	1.554	2.299
3	4.315	.421	.877	1.524	2.243
4	4.018	-.171	.468	1.353	2.204
5	4.286	.604	1.105	1.738	2.359
6	4.642	.087	.639	1.394	2.167
7	4.496	-.135	.534	1.406	2.339
8	6.774	.495	.949	1.577	2.290
9	3.830	.072	.642	1.372	2.140
10	3.268	.223	.809	1.546	2.413

FOCAL GROUP ITEM PARAMETERS

1	2.942	.428	.858	1.705	2.596
2	4.532	.523	.938	1.660	2.281
3	3.971	.382	.808	1.566	2.375
4	4.009	-.134	.424	1.323	2.174
5	4.880	.542	.965	1.709	2.493
6	4.362	.104	.580	1.354	2.101
7	4.651	-.072	.481	1.371	2.222
8	6.906	.513	.915	1.528	2.264
9	4.428	.234	.713	1.469	2.192
10	3.938	.343	.784	1.531	2.312

NUMBER OF CASES: 1116

FOLLOWING ARE MEASURES OF ITEM LEVEL EFFECT SIZE DESCRIBED BY WAINER T(1) THROUGH T(4) (WAINER, 1993) AND WOODS (AUD) (WOODS, 2011)

ITEM	T(1)	T(2)	T(3)	T(4)	AUD
1	-.0577	-64.4467	.0109	12.2029	.0825
2	.0444	49.5241	.0042	4.6869	.0444
3	-.0291	-32.4993	.0034	3.8411	.0440
4	-.0106	-11.8267	.0006	.6648	.0194
5	-.0414	-46.1491	.0065	7.2579	.0468
6	-.0272	-30.3682	.0014	1.5160	.0272
7	-.0068	-7.6217	.0018	2.0554	.0339
8	-.0081	-9.0011	.0008	.9128	.0173
9*	.1074	119.8884	.0181	20.1459	.1076
10	.0467	52.1378	.0070	7.7935	.0637

*Kim (2000) suggests that if T(1) exceeds .1, the item requires 'close examination'. This is the case for item #9 here.

FOLLOWING IS THE TOTAL TEST MEASURE OF BIAS (DTFR) AND MAGNITUDE OF IMPACT (STARK ET AL, 2004)

DTFR = .017595 SD OF TRUE SCORES (FOCAL) = 9.557122

MAGNITUDE OF EFFECT FOR DTF (STARK) .001841

BELOW ARE RAJU DFIT STATISTICS

ITEM	MEAN OF D	SD	COVAR.	C-DIF	NC-DIF	SQ-RT NC-DIF	CUT-OFF
1	-.0577	.0872	-.0029	-.0039	.0109	.1044	.0960
2	.0444	.0472	-.0041	-.0033	.0042	.0648	.0960
3	-.0291	.0509	.0011	.0005	.0034	.0583	.0960
4	-.0106	.0220	.0022	.0020	.0006	.0245	.0960
5	-.0414	.0692	.0081	.0074	.0065	.0806	.0960
6	-.0272	.0249	.0019	.0015	.0014	.0374	.0960
7	-.0068	.0424	.0041	.0040	.0018	.0424	.0960
8	-.0081	.0274	.0018	.0017	.0008	.0283	.0960
9	.1074	.0807	.0023	.0042	.0181	.1345	.0960
10	.0467	.0693	.0049	.0057	.0070	.0837	.0960

DTF = .0197
DTF CUT-OFF = .9600

NOTE THAT MEAN OF D IS IDENTICAL TO T(1) AND NC-DIF IS IDENTICAL TO T(3)
The square root of NC-DIF is given because this provides an effect size in terms of the original metric (Raju, 1999). A recommended cut-off of 0.0960 would correspond to an average absolute difference of about 0.31 of a point on an item scored from 0 to 4. A DTFR value represents the number of raw score points for the total scale due to measurement bias (Stark, 2004). In this case, about 0.02 of a point on a 40-point scale.

Observe that for any item where the absolute value of T(1) equals AUD, this indicates items with no crossing DIF, since AUD is the absolute value of the difference between true scores summed over all Focal members(and divided by N), and T(1) is the sum of the differences both positive and negative, divided by N. If T(1) is equal to AUD, there is no 'canceling out' of differences.

Figure 5.
Output from MAGNITS showing magnitude and impact indices