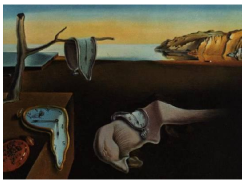

Journal of Statistical Software

January 2017, Volume 76, Issue 5. doi: 10.18637/jss.v076.i05

DFIT: An R Package for Raju's Differential Functioning of Items and Tests Framework

Víctor H. Cervantes
Instituto Colombiano para la Evaluación de la Educación

Abstract

This paper presents **DFIT**, an R package that implements the differential functioning of items and tests framework as well as the Monte Carlo item parameter replication approach for producing cut-off points for differential item functioning indices. Furthermore, it illustrates how to use the package to calculate power for the NCDIF index, both post hoc, as has regularly been the case in differential item functioning empirical and simulation studies, as well as a priori given certain item parameters. The version reviewed here implements all DFIT indices and Raju's area measures for tests comprised of items modeled with the same parametric item response unidimensional model (1-, 2-, and 3-parameters, generalized partial credit model or graded response model), the Mantel-Haenszel statistic with an underlying dichotomous item response model, and the item parameter replication method for any of the estimated indices with dichotomous item response models.

Keywords: DFIT framework, differential item functioning, type I error, power calculation, analytical standard error.

1. Introduction

Differential item functioning (DIF) has long been recognized as a threat to the validity of test scores and is an especially important issue to consider for fair comparisons between groups (Holland and Thayer 2009). In the past decades many procedures for identifying items with differential functioning have been envisioned and several of these methods have been implemented in R (R Core Team 2016): The package **difR** (Magis, Beland, Tuerlinckx, and De Boeck 2010) includes eleven different indices for dichotomous items; the package **lordif** (Choi, Gibbons, and Crane 2011) implements a modified logistic regression algorithm that allows for both dichotomous and polytomous items using estimated sum scores or item response theory (IRT) abilities as the matching criterion; packages like **mirt** (Chalmers 2012), **ltm** (Rizopoulos

2006), **lme4** (Bates, Mächler, Bolker, and Walker 2015), and **eRm** (Mair, Hatzinger, and Maier 2016; Mair and Hatzinger 2007) allow to test DIF using the likelihood ratio test through comparisons of multiple models; while packages like **psychomix** (Frick, Strobl, Leisch, and Zeileis 2012; Frick, Strobl, and Zeileis 2015), **mRm** (Preinerstorfer 2016), and **psychotree** (Komboz, Strobl, and Zeileis 2017; Strobl, Kopf, and Zeileis 2015), enable modeling and testing for test invariance using mixture Rasch models or through trees. A recent IRT parametric framework for detecting DIF that allows for detection for both dichotomous and polytomous items with unidimensional or multidimensional IRT models, known as the differential functioning of items and tests (DFIT) framework (Raju, Van der Linden, and Fleer 1995), was yet to be implemented.

This paper presents the R package **DFIT** (Cervantes 2017) which implements the framework indices for dichotomous and polytomous indices with several unidimensional models. It also implements the Monte Carlo item parameter replication (IPR) approach for hypothesis testing for the noncompensatory differential item functioning (NCDIF) index with dichotomous unidimensional IRT models. Section 2 presents a short overview of the DFIT framework and of the IPR approach, including an improvement of this approach to give correct NCDIF sampling distributions in the presence of group sample size differences. Section 3 illustrates the main functions in the package; they include functions for calculating Raju's area measures (Raju 1988), and the Mantel-Haenszel DIF statistic under IRT models assumptions (Roussos, Schnipke, and Pashley 1999). Section 4 presents how to use the package to calculate power for the NCDIF index. Finally, Section 5 concludes on the capabilities of the package and presents the future directions for its development.

2. The DFIT framework

The DFIT framework was proposed by Raju *et al.* (1995) as an improvement to the internal measures developed by Raju (1988) to detect items for which examinees from different groups responding to a test or a scale do not perform the same way (Holland and Thayer 2009). Originally proposed to identify DIF in dichotomous items and to be able to analyze differential functioning at the test level, it has subsequently been expanded to be used with bundles of items (Oshima, Raju, Flowers, and Slinde 1998), polytomous data (Raju, Fortmann-Johnson, Kim, Morris, Nering, and Oshima 2009), multidimensional models (Oshima, Raju, and Flowers 1997), and calculation of conditional DIF statistics (Oshima and Morris 2008). Within the DFIT framework differential functioning is analyzed between two groups of respondents: the reference group, and the focal group. It is generally considered that the reference group is the majority group and the focal group is the minority group or the group that might be at a disadvantage.

This framework proposes three indices for analyzing the differential functioning of items and tests:

DTF: The differential test functioning index.

CDIF: The compensatory DIF index.

NCDIF: The noncompensatory DIF index.

In order to define these indices for arbitrary IRT models and regardless of item responses being dichotomous or polytomous, let $S_i(\theta_j)$ be the expected score for examinee j with trait vector

θ_j on item i. The function S_i can stand for the expected score under either a one-, two-, or three-parameter model for dichotomous items (i.e., the probability of a correct response), or a rating scale, partial credit (PCM), or graded response model (GRM) for polytomous items, in their unidimensional or multidimensional expressions (Raju *et al.* 1995; Oshima *et al.* 1997; Oshima and Morris 2008). For a given examinee, the expected score on a test $T(\theta_j)$ (known as the "true" score in classical test theory) is equal to the sum of the expected scores for said examinee on the n items in the test. Note that different S_i do not need to have the same functional form for different items i and i'. Also, let $S_{i,g}$ and T_g represent these functions based on the true item parameters for group g.

The three indices are defined on the differences in the values of S and of T between the focal and the reference groups. Thus, $d_i(\theta_j) = S_{i,F}(\theta_j) - S_{i,R}(\theta_j)$ is the difference on the expected score for item i and examinee j, while $D(\theta_j) = T_F(\theta_j) - T_R(\theta_j)$ is the difference on the expected test scores. The square of these differences is taken as a measure of differential item or test functioning at the examinee level (Raju *et al.* 1995). The test level statistic (DTF) is, then, defined as the expected value over the focal group of the squared differences of expected test scores. That is:

$$\text{DTF} := \mathsf{E}_F(D^2(\theta_j)), \qquad j \in F. \tag{1}$$

And the basic item level statistic (NCDIF) is the expected value over the focal group of the difference of expected item scores, i.e.,

$$\text{NCDIF}_i := \mathsf{E}_F(d_i^2(\theta_j)), \qquad j \in F. \tag{2}$$

This index, as most DIF statistics, does not consider DIF from other items (or assumes that the other items are DIF free; Raju *et al.* 1995) and does not sum to the test level statistic. The third index within this framework seeks to define a DIF index such that its sum is equal to the test level statistic. To do so, DTF is decomposed in the following manner:

$$\begin{aligned}
\text{DTF} &= \mathsf{E}_F\left(\left(\sum_{i=1}^{n} d_i(\theta_j)\right)^2\right) \\
&= \sum_{i=1}^{n} \left(\mathsf{COV}_F(d_i(\theta_j), D(\theta_j)) + \mathsf{E}_F(d_i(\theta_j))\mathsf{E}_F(D(\theta_j))\right) \\
&= \sum_{i=1}^{n} \mathsf{E}_F(d_i(\theta_j)D(\theta_j)) \\
&= \sum_{i=1}^{n} \text{CDIF}_i
\end{aligned} \tag{3}$$

and, thus, the item index is defined as

$$\text{CDIF}_i := \mathsf{E}_F(d_i(\theta_j)D(\theta_j)), \qquad j \in F. \tag{4}$$

It should be noted that the CDIF index is additive by definition and includes information from DIFs on the other items.

An important characteristic of the indices in the DFIT framework is that their definition is based directly on the IRT interpretation of DIF rather than, for instance, differences on item parameters. Although differences on item parameters imply differences on the expected scores both at the item and the test levels, it should be noted that different sets of item parameters may produce very similar item characteristic curves (ICC), and thus, very similar expected scores. Also, what best distinguishes this framework from other approximations is that it explicitly states that the focus of bias and fairness research is how the minority group's results are affected and incorporates that into the procedure. This latter characteristic also enables DIF to be calculated for the three parameters IRT model from the perspective of the differences on expected scores where the original area measures were infinite when pseudo-guessing parameters differed between both groups as demonstrated by Raju (1988).

2.1. The IPR approach

Currently, hypothesis testing for NCDIF, the main index within the framework, is done using the item parameter replication (IPR) approach which was proposed by Oshima, Raju, and Nanda (2006) to overcome the limitations of the χ^2 tests originally developed by Raju *et al.* (1995). This approach "us[es] the focal group's item parameters for a test item, [and] a large number of pairs (e.g., 1000) of these parameters are reproduced. NCIDF for each of these pairs is calculated. These replicated pairs represent the 'No DIF' condition, and hence, any extreme differences observed would be considered beyond chance. [. . .] The NCDIF values [. . .] [are] used to determine statistical significance (Wright and Oshima 2015, p. 6)". Algorithmically, the procedure may be described as follows:

1. Define the item parameter vector for the null hypothesis (μ) to be equal to the item parameter estimates from the focal group.
2. Take the estimated variance-covariance matrix of those estimates (Σ).
3. Use these item parameters and covariance matrix to simulate as many item parameter vectors as desired[1]. They are obtained from the multivariate normal distribution with mean vector μ from Step 1 and covariance matrix Σ from Step 2.
4. Obtain the NCDIF values for each pair of item parameter replications.
5. Calculate the cut-off point as the $(1 - \alpha)$ percentile of the NCDIF values.
6. Repeat Steps 1 to 5 for each item.

This approach basically implements a Monte Carlo algorithm to generate chains of parameter vectors from their sampling distribution. However, as specified in Oshima *et al.* (2006) and implemented by Oshima, Kushubar, Scott, and Raju (2009), the sampling distribution required to generate these chains uses only the estimates for item parameters and the covariance matrices of these estimates obtained with data from the focal group only. However, as it stands, the procedure is analogous to finding the degrees of freedom for a two sample t-test by always choosing twice the number of observations of one of the groups minus two, regardless of the effective sample sizes in each of them. Recently, Clark and LaHuis (2012) examined the effect on type I error for the current algorithm under unequal and small sample sizes (500 or 250 examinees in the focal group and 500 examinees in the reference group) and found it increased when the sample size for the focal group reduced from 500 to 250 (from 0.05 to

[1] Oshima *et al.* (2006) recommend 1000 as a minimum number of replications.

0.15). This effect is due to the different sampling distributions for the parameter estimates of each group[2]. These effects on type I error and power will be shown in detail in Section 4, where full power calculations for the NCDIF index will be presented.

Taking these considerations into account, Cervantes (2012) proposed a modification to the algorithm presented by Oshima *et al.* (2006) for obtaining cut-off points based only on sample estimates. The modified algorithm comprises the following steps:

1. Define the item parameter vector for the null hypothesis to be equal to the item parameter estimates from the focal group.
2. Take the variance-covariance matrix of those estimates given the respective sample sizes and ability distributions of the focal and the reference group.
3. Obtain as many pairs of item parameter vectors as desired from the multivariate normal distribution for each group.
4. Obtain the NCDIF values for each pair of item parameter replications.
5. Calculate the cut-off point as the $(1 - \alpha)$ percentile of the NCDIF values.
6. Repeat Steps 1 to 5 for each item.

The previous algorithm is easily adapted to be used with theoretical values for item parameters and variance-covariance matrices. The effects of the change in the algorithm will be illustrated in Section 4 in particular with respect to expected differences for type I error and power.

3. The DFIT package

The **DFIT** package is able to calculate all indices from the DFIT framework for logistic and normal ogive unidimensional one-, two-, and three-parameters IRT models for dichotomous items, as well as for the PCM and GRM for polytomous items. It is freely available for download via the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=DFIT`. The main functions in package **DFIT** are the following: (a) `Cdif()`, `Ncdif()`, and `Dtf()` for calculating the statistics from the DFIT framework; (b) `PlotNcdif` for illustrating NCDIF; (c) `SignedArea()`, `UnsignedArea()`, and `IrtMh()` for calculating Raju's area measures and the Mantel-Haenszel DIF statistic; and (d) `Ipr()` and `CutoffIpr()` for obtaining the cut-off points for any of the DIF statistics by using the IPR approach. This section shows how to use the functions implemented in the **DFIT** package. Section 3.1 presents the functions devoted to calculating DIF and DTF, while Section 3.2 presents the functions to obtain cut-off values by means of the IPR approach.

The following code loads the package and the data used in the examples.

```
R> library("DFIT")
R> data("dichotomousItemParameters", package = "DFIT")
R> data("polytomousItemParameters", package = "DFIT")
```

[2] For the analytical variance-covariance matrices, it may be shown that all entries are inversely proportional to sample size (see for example equation A3 in Li and Lissitz 2004 that shows the form of the elements of the information matrix), and as such the difference in these matrices will be proportional to the sample size ratio of both groups when the ability distributions are equal.

Item parameters from `data("dichotomousItemParameters", package = "DFIT")` are based on those used by Wright (2011), while those from `data("polytomousItemParameters", package = "DFIT")` are based on those used by Raju *et al.* (2009). In the following code, the subsets of data that will be used to illustrate DIF with Rasch models and with the three-parameter logistic model are selected.

```
R> raschParameters <- lapply(dichotomousItemParameters, function(x)
+    x[, 2, drop = FALSE])
R> raschParameters <- as.list(unique(as.data.frame(raschParameters)))
R> raschParameters <- lapply(raschParameters, function(x)
+    matrix(x, ncol = 1))
R> items3Pl <- c(2, 20, 22, 8, 10, 28, 46, 32)
R> threePlParameters <- lapply(dichotomousItemParameters, function(x)
+    x[items3Pl, ])
```

The format expected by all functions that use item parameters is a list with two named elements: focal and reference, each a matrix with a row for each item and columns according to the IRT model. Discrimination parameters (except for one-parameter IRT models for dichotomous items), must be included in the first column and pseudo-guessing parameters for three-parameters IRT models, in the third column. In the case of the polytomous models currently supported: generalized partial credit model (GPCM) or graded response model (GRM), the first column includes the discrimination parameters and the other columns the item categories difficulty parameters, that is the models should be parametrized as $a_i(\theta - b_{ik})$ rather than $a_i(\theta - b_i + d_k)$. The formatting is illustrated next by the headings of parameters for the three-parameter model and for polytomous models.

```
R> lapply(threePlParameters, head, 5)

$focal
     [,1] [,2] [,3]
[1,]  1.0 -3.0 0.00
[2,]  1.0 -2.7 0.05
[3,]  0.5 -2.4 0.00
[4,]  1.0 -2.2 0.00
[5,]  0.5  0.0 0.00

$reference
     [,1] [,2] [,3]
[1,]    1   -3 0.00
[2,]    1   -3 0.05
[3,]    1   -3 0.05
[4,]    1   -3 0.00
[5,]    1    0 0.00

R> lapply(polytomousItemParameters, head, 5)

$focal
        [,1]  [,2]  [,3] [,4] [,5]
```

```
[1,] 0.7300 -0.80  0.40 1.60 2.80
[2,] 1.2425 -0.80  0.40 1.60 2.80
[3,] 0.4289 -0.80  0.40 1.60 2.80
[4,] 0.5000 -1.82 -0.62 0.58 1.78
[5,] 0.8510 -1.82 -0.62 0.58 1.78

$reference
       [,1]  [,2]  [,3] [,4] [,5]
[1,] 0.7300 -1.80 -0.60 0.60 1.80
[2,] 1.2425 -1.80 -0.60 0.60 1.80
[3,] 0.4289 -1.80 -0.60 0.60 1.80
[4,] 1.0000 -2.32 -1.12 0.08 1.28
[5,] 1.7020 -2.32 -1.12 0.08 1.28
```

3.1. DIF functions

Using this package, the DFIT statistics may be estimated either as the means, given a vector of abilities from a sample of examinees from the focal group, of the values within the expectations in expressions (1), (2), and (4) – this is the method used for estimation by Raju *et al.* (1995) and Oshima *et al.* (2009), or by integrating over the specified ability distribution (standard normal by default). The following code shows the use of Ncdif and Cdif functions to obtain DFIT's item indices for the one-parameter logistic IRT model using the normal ogive metric (i.e., $D = 1.702$). It also shows how to use the Dtf() function to estimate the differential test functioning index.

```
R> ncdif1pl <- Ncdif(itemParameters = raschParameters, irtModel = "1pl",
+    focalAbilities = NULL, focalDistribution = "norm",
+    subdivisions = 5000, logistic = FALSE)
R> cdif1pl  <- Cdif(itemParameters = raschParameters, irtModel = "1pl",
+    focalAbilities = NULL, focalDistribution = "norm",
+    subdivisions = 5000, logistic = FALSE)
R> dtf1plWithCdif <- Dtf(cdif = cdif1pl)
R> dtf1plWithoutCdif <- Dtf(cdif = NULL, itemParameters = raschParameters,
+    irtModel = "1pl", focalAbilities = NULL, focalDistribution = "norm",
+    subdivisions = 5000, logistic = FALSE)
```

Functions to calculate Raju's area measures for logistic and normal ogive IRT models are also included in the package. These measures may be calculated for dichotomous models and for polytomous models (based on the expected scores for the GRM as proposed by Cohen, Kim, and Baker 1993). Also, through the function IrtMh, it is possible to estimate the Mantel-Haenszel DIF parameter when the given item parameters are assumed to hold with the one-, two-, or three-parameters IRT models for dichotomous responses. The function DeltaMhIrt transforms the Mantel-Haenszel statistic into the ETS delta metric that has long been used as the standard for DIF effect size assessment.

```
R> sam1pl <- SignedArea(itemParameters = raschParameters, irtModel = "1pl",
+    subdivisions = 5000, logistic = FALSE)
```

Item	Reference b	Focal b	MH	ETS Δ	SA	UA	NCDIF	CDIF
1	−3.0	−3.0	1.000	−0.00	0.00	0.00	0.00000	0.00000
2	−3.0	−2.7	1.666	−1.20	−0.30	0.30	0.00051	0.00604
3	−3.0	−2.4	2.777	−2.40	−0.60	0.60	0.00273	0.01471
4	−3.0	−2.2	3.902	−3.20	−0.80	0.80	0.00582	0.02235
5	0.0	0.0	1.000	−0.00	0.00	0.00	0.00000	0.00000
6	0.0	0.3	1.666	−1.20	−0.30	0.30	0.00841	0.05019
7	0.0	0.6	2.777	−2.40	−0.60	0.60	0.03228	0.09803
8	0.0	0.8	3.902	−3.20	−0.80	0.80	0.05503	0.12725

Table 1: DIF statistics for example items under the 1PL model.

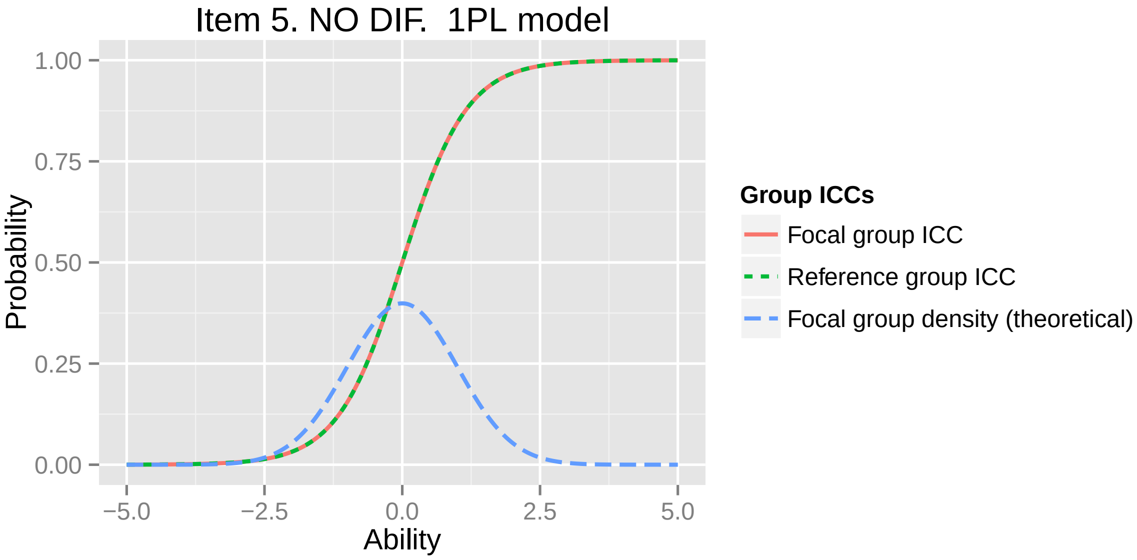

Figure 1: Plot of an item with no DIF using `PlotNcdif` showing the density.

```
R> uam1pl <- UnsignedArea(itemParameters = raschParameters, irtModel = "1pl",
+    subdivisions = 5000, logistic = FALSE)
R> mh1pl <- IrtMh(itemParameters = raschParameters, irtModel = "1pl",
+    focalDistribution = "norm", referenceDistribution = "norm",
+    focalDistrExtra = list(mean = 0, sd = 1),
+    referenceDistrExtra = list(mean = 0.5, sd = 1), groupRatio = 1,
+    logistic = FALSE)
R> delta1pl <- DeltaMhIrt(mh1pl)
```

Table 1 presents the DIF statistics for the selected items, assuming a standard normal distribution for the abilities of examinees from the focal group. Item difficulties for both groups are presented in columns "Reference b" and "Focal b." Next to the Mantel-Haenszel (MH) statistic, appears the associated effect size measure (ETS Δ). The signed (SA) and unsigned (UA) area measures are shown along with the DFIT item statistics. The DTF statistic for a test composed of this set of items would be 0.3186. It should be noted that Items 1 and 5 do not present any differential functioning, while all other items show some amount that would

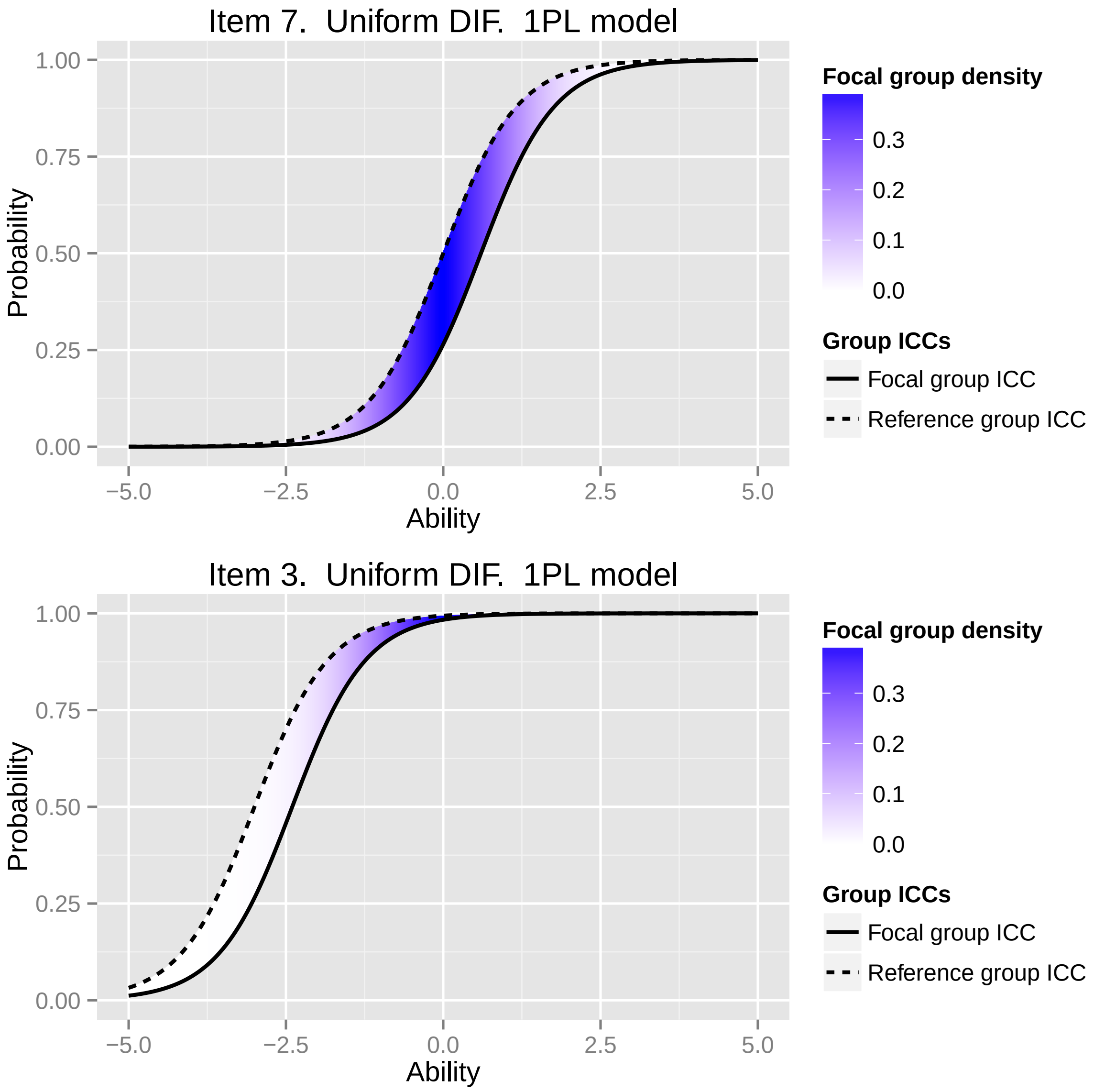

Figure 2: DIF plots of items with DIF and difficulty close (up) and far (down) the focal group mean.

be categorized as moderate (or B, Items 2 and 6) or large (or C, all other items) given their ETS Δ measure.

The use of the `PlotNcdif` function is illustrated below. Figure 1 presents an item with no DIF (Item 5) for which the normal density is plotted. Figure 2 presents the plots for two items with DIF, one with item difficulties for both groups close to the mean of the focal group ability, and one for which they are far. This figure also shows an alternative representation of the weighting given by the density of the abilities of examinees from the focal group, as obtained by setting `plotDensity = FALSE`.

```
R> it5PlotD <- PlotNcdif(iiItem = 5, itemParameters = raschParameters,
+    irtModel = "1pl", plotDensity = TRUE, logistic = FALSE,
```

Item	Reference	Focal	MH	ETS Δ	SA	UA	NCDIF	CDIF
2	(1, −3, 0)	(1, −3, 0)	1.00	−0.000	0.000	0.000	0.0000	0.000
20	(1, −3, 0.05)	(1, −2.7, 0.05)	1.66	−1.190	−0.285	0.285	0.0005	0.004
22	(1, −3, 0.05)	(0.5, −2.4, 0)	10.08	−5.430	Inf	Inf	0.0193	0.054
8	(1, −3, 0)	(1, −2.2, 0)	3.90	−3.200	−0.800	0.800	0.0058	0.016
10	(1, 0, 0)	(0.5, 0, 0)	1.00	0.000	0.000	0.815	0.0121	0.018
28	(1, 0, 0.05)	(1, 0.3, 0.05)	1.58	−1.075	−0.285	0.285	0.0076	0.046
46	(1, 0, 0.1)	(0.5, 0.6, 0.15)	1.38	−0.753	Inf	Inf	0.0179	0.055
32	(1, 0, 0.05)	(1, 0.8, 0.05)	3.25	−2.771	−0.760	0.760	0.0497	0.122

Table 2: DIF statistics for example items under the 3PL model.

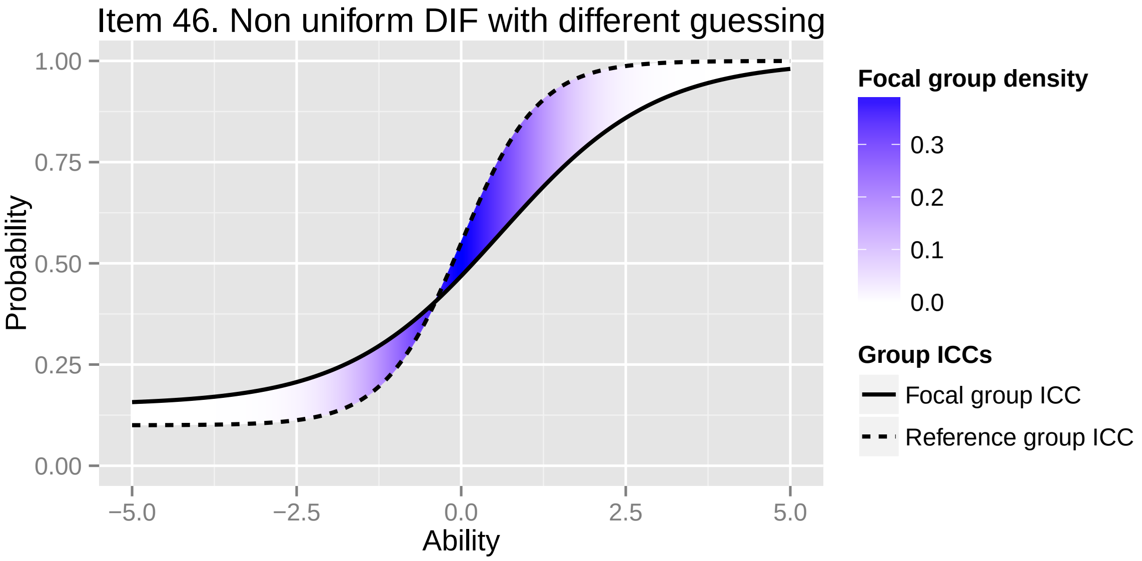

Figure 3: Plot of an item with nonuniform DIF and different pseudo-guessing parameter between groups.

```
+    focalDensityText = "Focal group density (theoretical)",
+    main = "Item 5. NO DIF.  1PL model")
R> it7PlotS <- PlotNcdif(iiItem = 7, itemParameters = raschParameters,
+    irtModel = "1pl", plotDensity = FALSE, logistic = FALSE,
+    main = "Item 7.  Uniform DIF.  1PL model")
```

All these measures may also be obtained for items under the two- and three-parameters IRT models. The argument `irtModel` should be, respectively, `"2pl"` and `"3pl"`. Table 2 presents the statistics for the items selected above under a three-parameters model. For a test composed of this set of items, the DTF statistic would be 0.3151. Also, Figure 3 shows the plot for an item with nonuniform DIF where the guessing parameters for the focal and the reference groups are also different

For polytomous IRT models, the DIF statistics (except for the Mantel-Haenszel and the associated delta measure) may be calculated by setting `irtModel = "grm"` or `"pcm"`. Table 3 presents the DIF statistics assuming that a GRM holds. For a test composed of this set of

Item	Reference	Focal	SA	UA	NCDIF	CDIF
1	(0.73, −1.8, −0.6, 0.6, 1.8)	(0.73, −0.8, 0.4, 1.6, 2.8)	−4.0	4.00	0.477	2.938
2	(1.24, −1.8, −0.6, 0.6, 1.8)	(1.24, −0.8, 0.4, 1.6, 2.8)	−4.0	4.00	0.595	3.282
3	(0.43, −1.8, −0.6, 0.6, 1.8)	(0.43, −0.8, 0.4, 1.6, 2.8)	−4.0	4.00	0.298	2.314
4	(1, −2.32, −1.12, 0.08, 1.28)	(0.5, −1.82, −0.62, 0.58, 1.78)	−2.0	2.98	0.166	1.756
5	(1.7, −2.32, −1.12, 0.08, 1.28)	(0.85, −1.82, −0.62, 0.58, 1.78)	−2.0	2.22	0.166	1.749
6	(0.59, −2.32, −1.12, 0.08, 1.28)	(0.29, −1.82, −0.62, 0.58, 1.78)	−2.0	5.11	0.150	1.605
7	(1, −1.28, −0.08, 1.12, 2.32)	(1, −0.78, 0.42, 1.62, 2.82)	−2.0	2.00	0.134	1.571
8	(1.7, −1.28, −0.08, 1.12, 2.32)	(1.7, −0.78, 0.42, 1.62, 2.82)	−2.0	2.00	0.154	1.674
9	(0.59, −1.28, −0.08, 1.12, 2.32)	(0.59, −0.78, 0.42, 1.62, 2.82)	−2.0	2.00	0.098	1.342
10	(1, −1.8, −0.6, 0.6, 1.8)	(0.5, −1.8, −0.6, 0.6, 1.8)	0.0	2.30	0.025	0.179
11	(1.7, −1.8, −0.6, 0.6, 1.8)	(0.85, −1.8, −0.6, 0.6, 1.8)	0.0	0.93	0.005	0.082
12	(0.59, −1.8, −0.6, 0.6, 1.8)	(0.29, −1.8, −0.6, 0.6, 1.8)	−0.0	4.80	0.048	0.249

Table 3: DIF statistics for example items under the GRM model.

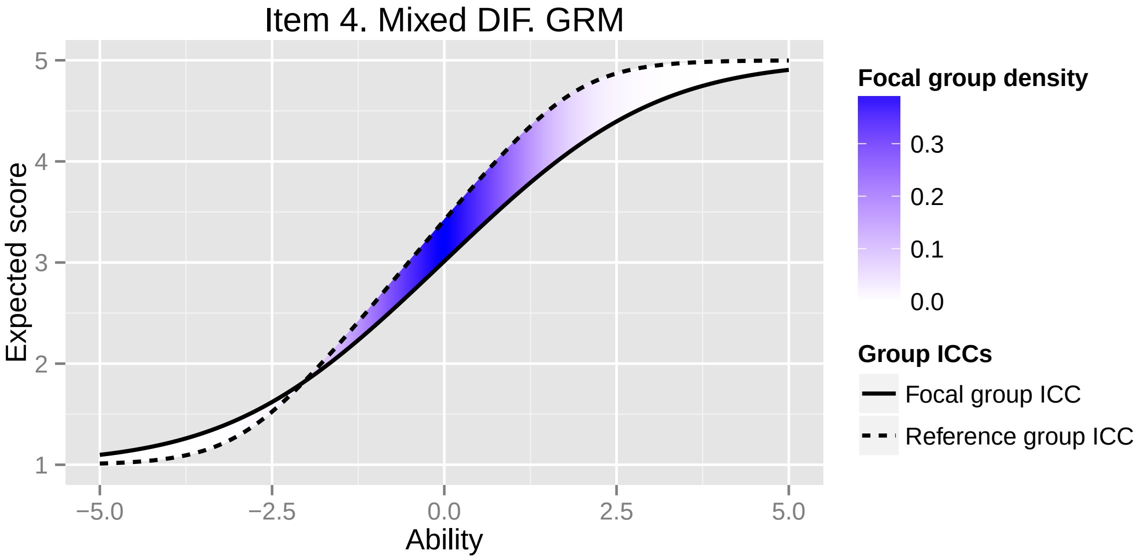

Figure 4: Plot of an item with mixed DIF under the GRM model parameter between groups.

items, the DTF statistic would be 18.7411. Figure 4 shows the plot for one of these items which exhibits mixed DIF.

3.2. IPR cut-off points

The **DFIT** package is able to perform the IPR Monte Carlo procedure as described in Section 2.1 under any unidimensional IRT model if the item parameters and their variance-covariance matrices are provided; it is also able to calculate the NCDIF index, the SA and UA measures, and the Mantel-Haenszel DIF statistic with the generated item parameter pairs with the same models as shown in Section 3.1. Currently (version 1.0-3), the **DFIT** package is able to calculate the asymptotic variance-covariance matrices for item parameter estimates under the three logistic IRT models for dichotomous responses (1PL, 2PL, and 3PL) via the function `AseIrt`. The main function to perform the IPR algorithm in order to obtain cut-off points is `CutoffIpr`; however, the function `Ipr` will be presented too since it allows greater flexibility. The following code illustrates their use to obtain the cut-off values under the Rasch model (i.e., $D = 1.0$ under the one-parameter model) and the different ways the function `CutoffIpr` may be used.

Firstly, it is shown how to directly obtain the cut-off values from only the information on the estimated item parameters and sample sizes for each group. This is done by letting `CutoffIpr` calculate the asymptotic variance-covariance matrices for the models. For this, and the next examples, it is assumed that the reference group has a mean ability of 0.5 logits greater than the focal group.

```
R> set.seed(89334828)
R> cutoffRaschNcdif1 <- CutoffIpr(quantiles = 0.95,
+    itemParameters =  raschParameters, itemCovariances = "asymptotic",
+    nullGroup = "focal", irtModel = "1pl", focalSampleSize = 500,
+    referenceSampleSize = 1500,
+    referenceDistrExtra = list(mean = 0.5, sd = 1),
+    logistic = TRUE, statistic = "ncdif", nReplicates = 1000)
```

The following illustrates how to calculate the asymptotic matrices manually for the items under the Rasch model. Keeping on with current practice, the parameters from the focal group will be used as the expected item difficulties for both groups, but as pointed out in Section 2.1, the asymptotic covariance matrices will be calculated for each group according to their sample size and ability distribution.

```
R> nullParameters <- list()
R> nullParameters[["focal"]] <- raschParameters[["focal"]]
R> nullParameters[["reference"]] <- raschParameters[["focal"]]
R> raschAse <- list()
R> raschAse[["focal"]] <- AseIrt(itemParameters = nullParameters[["focal"]],
+    distribution = "norm", sampleSize = 500, irtModel = "1pl",
+    distributionParameters = list(mean = 0, sd = 1),
+    logistic = TRUE)
R> raschAse[["reference"]] <- AseIrt(
+    itemParameters = nullParameters[["reference"]],
+    distribution = "norm", sampleSize = 1500, irtModel = "1pl",
+    distributionParameters = list(mean = 0.5, sd = 1),
+    logistic = TRUE)
```

With these asymptotic variances, it is then possible to obtain the cut-off values by setting the `itemCovariances` to those obtained manually, as shown next. Also, a different random seed is used in order to assess the stability of the estimated cut-off values.

```
R> set.seed(29834328)
R> cutoffRaschNcdif2 <- CutoffIpr(quantiles = 0.95,
+    itemParameters = nullParameters, itemCovariances = raschAse,
+    irtModel = "1pl", logistic = TRUE, statistic = "ncdif",
+    nReplicates = 1000)
```

Additionally, obtaining the variance-covariance matrices independently from the cut-off function (whether calculating the asymptotic ones as presented or the estimated ones, if available), allows further flexibility to the algorithm, as well as reducing simulation time when the researcher is interested in calculating several indices and cut-off values. The following code shows how to apply the IPR algorithm to the other indices.

```
R> set.seed(29833326)
R> raschIpr <- Ipr(itemParameters = nullParameters,
+    itemCovariances = raschAse, nReplicates = 1000)
R> raschNcdifIpr <- IprNcdif(itemParameterList = raschIpr, irtModel = "1pl",
+    logistic = TRUE)
R> raschUamIpr <- IprUam(itemParameterList = raschIpr, irtModel = "1pl",
+    logistic = TRUE)
R> raschSamIpr <- IprSam(itemParameterList = raschIpr, irtModel = "1pl",
+    logistic = TRUE)
R> raschMhIpr <- IprMh(itemParameterList = raschIpr, irtModel = "1pl",
+    logistic = TRUE)
```

Given these chains of estimated statistics, the cut-off values may be obtained by providing the `iprStatistics` argument with the corresponding chain. Note that when `itemCovariances`, `itemParameterList`, or `iprStatistics` are directly provided to the `CutoffIpr` function, one must check that the desired null condition is the one being specified. The following code obtains the cut-off values for NCDIF (`statistic = "ncdif"`), Raju's signed (`statistic = "sam"`) and unsigned (`statistic = "uam"`) area measures, and Mantel-Haenszel (`statistic = "mh"`).

```
R> cutoffRaschNcdif3 <- CutoffIpr(quantiles = 0.95,
+    iprStatistics = raschNcdifIpr, itemParameterList = raschIpr,
+    itemParameters = nullParameters, itemCovariances = raschAse,
+    irtModel = "1pl", statistic = "ncdif")
R> cutoffRaschUam <- CutoffIpr(quantiles = 0.95, iprStatistics = raschUamIpr,
+    itemParameterList = raschIpr, itemParameters = nullParameters,
+    itemCovariances = raschAse, irtModel = "1pl", statistic = "uam")
R> cutoffRaschSam <- CutoffIpr(quantiles = c(0.025, 0.95),
+    iprStatistics = raschSamIpr,itemParameterList = raschIpr,
+    itemParameters = nullParameters, itemCovariances = raschAse,
+    irtModel = "1pl", statistic = "sam")
R> cutoffRaschMh <- CutoffIpr(quantiles = c(0.025, 0.975),
+    iprStatistics = raschMhIpr, itemParameterList = raschIpr,
+    itemParameters = nullParameters, itemCovariances = raschAse,
+    irtModel = "1pl", statistic = "mh")
```

Table 4 shows the different cut-off points obtained for the 8 items with the calls to the `CutoffIpr` function. It may be appreciated that, although differences between different simulations are very small, given the magnitude the NCDIF index may take, the variations occur even on the second significant digit. This suggests that more than 1000 replications might be desirable when obtaining cut-off points through this approach.

Table 5 presents the cut-off points obtained for the other DIF statistics along with their true values given the item parameters and the defined distributions for each group (standard

Item	Reference b	Focal b	True NCDIF	Cut-off 1	Cut-off 2	Cut-off 3
1	−3	−3.0	0.00000	0.00125	0.00113	0.00116
2	−3	−2.7	0.00065	0.00129	0.00128	0.00135
3	−3	−2.4	0.00310	0.00153	0.00155	0.00161
4	−3	−2.2	0.00619	0.00152	0.00149	0.00162
5	0	0.0	0.00000	0.00200	0.00213	0.00218
6	0	0.3	0.00400	0.00229	0.00228	0.00217
7	0	0.6	0.01565	0.00210	0.00209	0.00216
8	0	0.8	0.02716	0.00221	0.00212	0.00193

Table 4: NCDIF cut-off points under the Rasch model through the IPR approach.

Item	MH lower	True MH	MH upper	SA lower	True SA	SA upper	True UA	UA
1	0.662	1.000	1.573	−0.4533	0.0	0.3437	0.0	0.4404
2	0.671	1.350	1.478	−0.3907	−0.3	0.3372	0.3	0.3967
3	0.696	1.822	1.424	−0.3532	−0.6	0.3090	0.6	0.3603
4	0.729	2.226	1.378	−0.3208	−0.8	0.2746	0.8	0.3200
5	0.803	1.000	1.251	−0.2240	0.0	0.1887	0.0	0.2207
6	0.804	1.350	1.256	−0.2280	−0.3	0.1910	0.3	0.2224
7	0.803	1.822	1.268	−0.2375	−0.6	0.1933	0.6	0.2314
8	0.796	2.226	1.248	−0.2215	−0.8	0.1972	0.8	0.2241

Table 5: Cut-off points for different DIF statistics under the Rasch model through the IPR approach.

normal for the focal group, and normal with mean 0.5 and standard deviation 1 for the reference group). Given the cut-off values, for all DIF indices, none of the items without DIF would be identified. Also Item 2 which presents moderate DIF, according to the usual effect size (after correcting for the difference in the constant D), is not detected by any index, although for all measures (except NCDIF) the true value for Item 2 is the same than that for Item 7. This difference is directly related to the distance between the item difficulties and the mean ability for the focal group; these differences are part of DIF definition in the DFIT framework.

4. Power calculation

This section illustrates how to use the functions in package **DFIT** to calculate power for the NCDIF index; the procedure may be used to obtain power for the other DIF measures; however, due to the interaction between the CDIF index with parameters of other items and DIF presence on the test level, the procedure is not recommended for this index. Section 4.1 presents how to obtain power curves, given both the cut-off points obtained by the current IPR procedure (as implemented in Oshima *et al.* 2009 and presented by Oshima *et al.* 2006, and under the modification presented in Section 2.1. These curves are useful during the planning of DIF analyses to avoid underpowered studies, and may help in defining effect sizes for the DFIT statistics.

Section 4.2 shows how to calculate power for a given set of item parameters using the items presented in Table 1 under the 1PL model. The shown procedure may be used to perform post-hoc power analyses, or a priori power calculations against a given difference of item parameters, as those presented by Wright and Oshima (2015). For all these examples, it is assumed that the ability for the focal and the reference groups is distributed as a standard normal (i.e., without impact).

The examples in this section will further illustrate the effects on power of the selected algorithm to obtain the cut-off points for the NCDIF statistic. Thus, in each case, both the results using the current algorithm as presented by Oshima *et al.* (2006) and the modified proposal are calculated and plotted.

4.1. Power curves

The power curves for uniform DIF and nonuniform DIF for an item under the two-parameters logistic IRT model with difficulty 0 and discrimination 1 are presented; these curves were originally presented by Cervantes (2012). In order to show how the proposed modified IPR algorithm compares to the current one, different sample sizes are used for each group. The general conditions for both uniform and nonuniform DIF power curves examples are set as follows,

```
R> nReplicates <- 3000
R> nFocal <- 800
R> nReference <- 2500
R> kRatio <- nReference / nFocal
R> focalParam <- list(mean = 0, sd = 1)
R> referenceParam <- list(mean = 0, sd = 1)
```

First, the code to obtain power curves for uniform DIF is presented. The item parameters for the null and alternative hypotheses are generated for a number of equally spaced item difficulties lesser and greater to 0 for the focal group, while the difficulty remains constant for the reference group.

```
R> itemParameters <- list(focal = cbind(rep(1, 51),
+    seq(-0.5, 0.5, length = 51)), reference = cbind(rep(1, 51), rep(0, 51)))
R> nullFocal <- which(itemParameters[["focal"]][, 2] ==
+    itemParameters[["reference"]][, 2])
R> itemParametersNull <- lapply(itemParameters, function(x)
+    x[nullFocal, , drop = FALSE])
R> names(itemParametersNull) <- c("focal", "reference")
```

Also, for each of these parameter vectors, we obtain the actual value of the NCDIF statistic. These will serve to compare power as a function of the true item parameters as well as as a function of the true NCDIF statistic.

```
R> twoPlUniNcdifTrue <- Ncdif(itemParameters, irtModel = "2pl",
+    focalDistribution = "norm", focalDistrExtra = focalParam,
+    logistic = FALSE)
```

Next, we obtain the asymptotic variance-covariance matrices of parameter estimates given the known item parameters for each group.

```
R> twoPlUniAse <- list()
R> twoPlUniAse[["focal"]] <- AseIrt(
+    itemParameters = itemParameters[["focal"]],
+    distribution = "norm", distributionParameters = focalParam,
+    sampleSize = nFocal, irtModel = "2pl", logistic = FALSE)
R> twoPlUniAse[["reference"]] <- AseIrt(
+    itemParameters = itemParameters[["reference"]],
+    distribution = "norm", distributionParameters = referenceParam,
+    sampleSize = nReference, irtModel = "2pl", logistic = FALSE)
```

Using the IPR Monte Carlo approach, the simulated item parameters are obtained for the null and the alternative hypotheses with the asymptotic covariance matrices calculated in the previous step. Then, the NCDIF statistic on each pair of parameter vectors is calculated.

```
R> set.seed(29834328)
R> twoPlUniIpr <- Ipr(itemParameters = itemParameters,
+    itemCovariances = twoPlUniAse, nReplicates = nReplicates)
R> twoPlUniNcdif <- IprNcdif(itemParameterList = twoPlUniIpr,
+    irtModel = "2pl", logistic = FALSE, subdivisions = 1000)
```

The corresponding cut-off point for the proposed algorithm is obtained from the 95th quantile of the NCDIF statistics from the generated item parameters.

```
R> cutoffPointEachSZUni <- CutoffIpr(quantiles = 0.95,
+    iprStatistics = twoPlUniNcdif[nullFocal, , drop = FALSE])
```

We will also obtain the cut-off point under the current algorithm. For that, we follow the same steps changing the covariance matrices to be equal for both groups. Since both groups were assumed to have the same distribution, obtaining the asymptotic variance-covariance matrix is possible directly from the ones from each group because they are proportional in this case. The following code obtains the matrices, the replicated parameters, the NCDIF statistics, and the cut-off point under the current approach.

```
R> set.seed(29834328)
R> twoPlUniAseCurrent <- twoPlUniAse
R> twoPlUniAseCurrent[["focal"]] <- twoPlUniAseCurrent[["focal"]][nullFocal]
R> twoPlUniAseCurrent[["reference"]] <- lapply(
+    twoPlUniAseCurrent[["reference"]][nullFocal], "*", kRatio)
R> twoPlUniIprCurrent <- Ipr(itemParameters = itemParametersNull,
+    itemCovariances = twoPlUniAseCurrent, nReplicates = nReplicates)
R> twoPlUniNcdifCurrent <- IprNcdif(itemParameterList = twoPlUniIprCurrent,
+    irtModel = "2pl", logistic = FALSE, subdivisions = 1000)
R> cutoffPointUni <- CutoffIpr(quantiles = 0.95,
+    iprStatistics = matrix(twoPlUniNcdifCurrent, nrow = length(nullFocal)))
```

Power for each alternative is estimated as the proportion of the replicated NCDIF indexes that is greater than the cut-off value. Figures 5 and 6 show how power varies as a function of item difficulty and NCDIF index, respectively. The figures indicate the value for which power

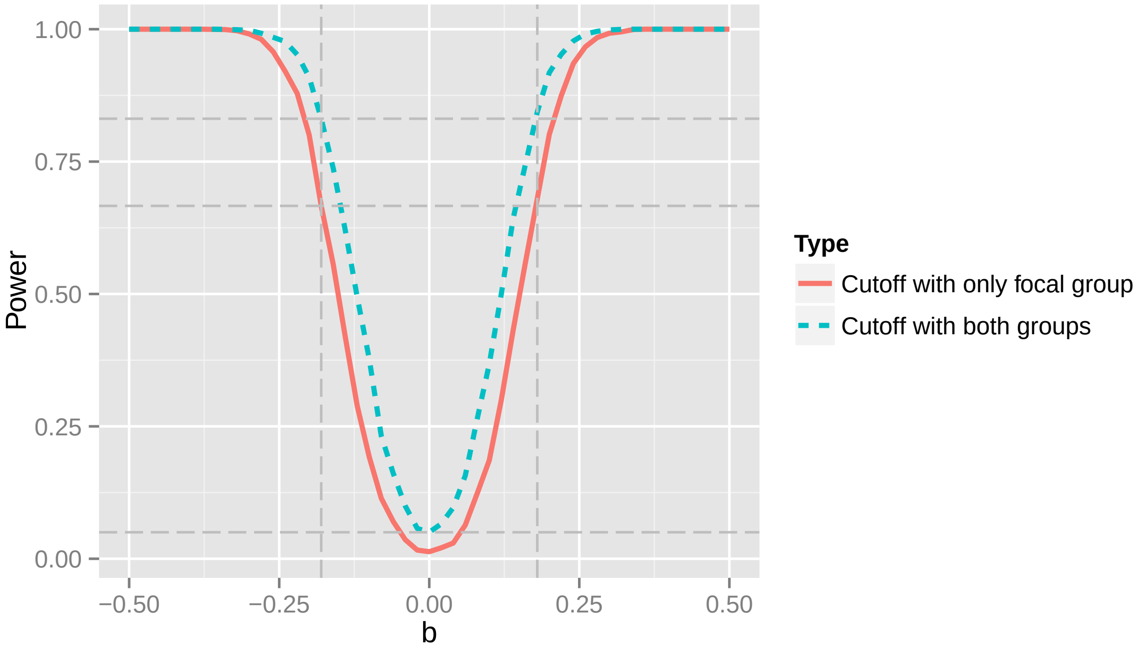

Figure 5: Power curves for uniform DIF as a function of focal item difficulty.

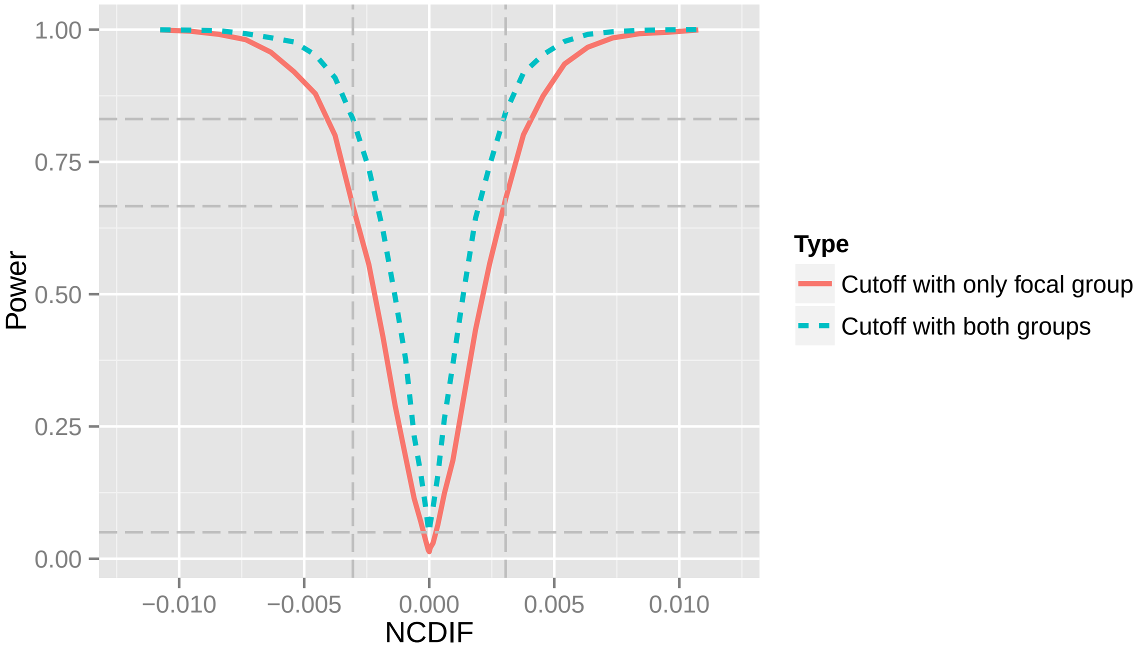

Figure 6: Power curves for uniform DIF as a function of NCDIF.

is approximately 0.8 (0.831). It may be seen that low type I error rates as those reported by Oshima *et al.* (2006) are expected for the current IPR algorithm because the actual type I error is well below the nominal value. It may also be seen that power is affected by this. Power with the current algorithm only reaches 0.6663 for the same item difficulty, about 0.16 less than with the proposed algorithm.

The code to obtain the power curves for nonuniform DIF is similar to the one presented for uniform DIF; thus, it is not shown. Figures 7 and 8 show how power varies as a function of item discrimination and NCDIF index, respectively. The figures indicate the value for which power is 0.8177 with the proposed modification, power with the current algorithm only reaches 0.633. Furthermore, although the power curves are not symmetrical with respect to the discrimination value, they are as a function of the NCDIF index value. It is also apparent that power varies for uniform and nonuniform DIF; power is about 0.8 for a value of 0.0031 where DIF is uniform, while to achieve the same power for nonuniform DIF (with equal sample sizes, no impact and item difficulty equal to both groups ability mean), the NCDIF index needs to reach only 0.0025.

4.2. Power calculation

This section presents how to calculate power for particular item difficulties for both the reference and the focal group using the 1PL model. The difficulties shown in Table 4 are used, and the null hypothesis assumed is that item difficulties are the ones from the focal group; additionally, equal standard normal distributions for the abilities from both groups with unequal sample sizes are assumed. A similar procedure may be used for other IRT models.

The following code sets the sample size conditions for both groups, the distribution parameters and the number of IPR replications that will be used. First, we set the global variables,

```
R> nFocal <- 500
R> nReference <- 1500
```

The variances of item difficulties are obtained for each group according to the item parameters for the focal group and their respective sample size and common distribution.

```
R> nullParameters <- list()
R> nullParameters[["focal"]] <- raschParameters[["focal"]]
R> nullParameters[["reference"]] <- raschParameters[["focal"]]
R> nullAse <- list()
R> nullAse[["focal"]] <- AseIrt(itemParameters = nullParameters[["focal"]],
+    distribution = "norm", sampleSize = nFocal,
+    distributionParameters = distriParam, irtModel = "1pl",
+    logistic = FALSE)
R> nullAse[["reference"]] <- AseIrt(itemParameters =
+    nullParameters[["reference"]], distribution = "norm",
+    sampleSize = nReference, distributionParameters = distriParam,
+    irtModel = "1pl", logistic = FALSE)
```

With these variances for the parameter estimates, the cut-off point for each item using the IPR approach is obtained for the null hypothesis of each item.

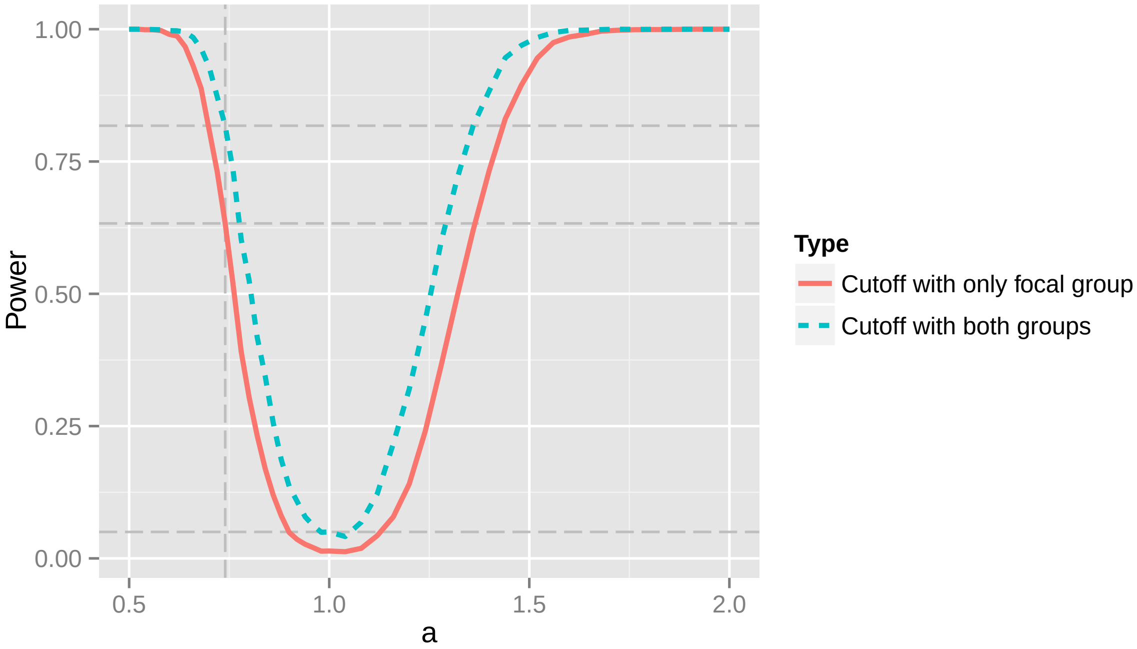

Figure 7: Power curves for nonuniform DIF as a function of focal item discrimination.

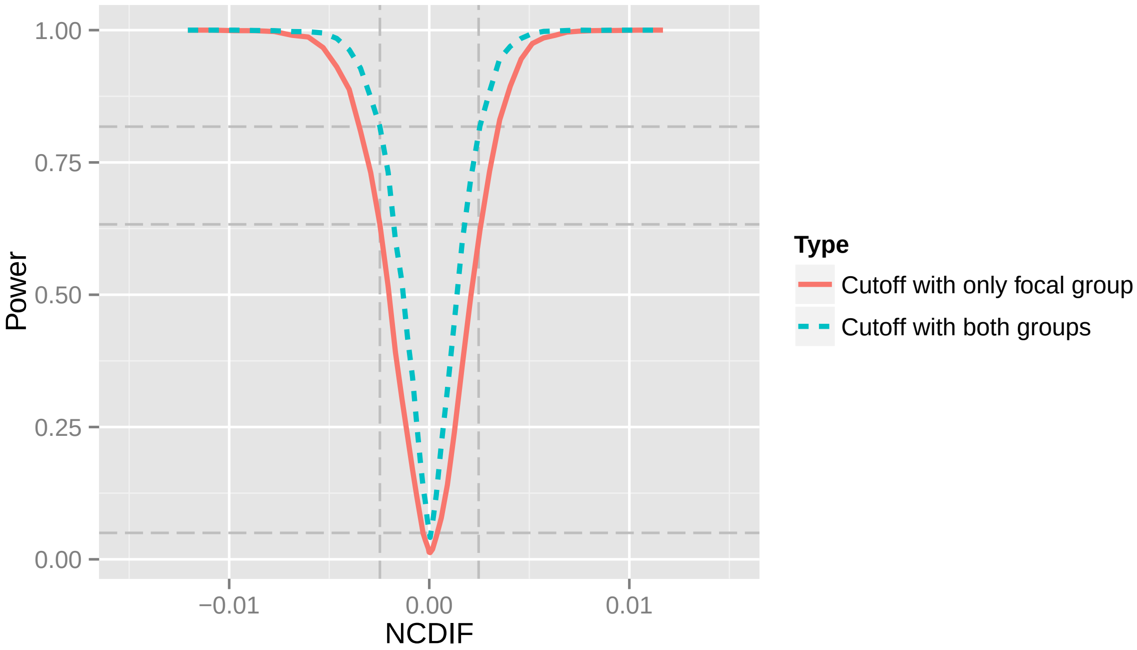

Figure 8: Power curves for nonuniform DIF as a function of NCDIF.

Item	Reference b	Focal b	True NCDIF	Cut-off	Power
1	−3	−3.0	0.00000	0.00084	0.05
2	−3	−2.7	0.00051	0.00091	0.31
3	−3	−2.4	0.00273	0.00116	0.88
4	−3	−2.2	0.00582	0.00121	0.99
5	0	0.0	0.00000	0.00199	0.05
6	0	0.3	0.00841	0.00200	0.98
7	0	0.6	0.03228	0.00191	1.00
8	0	0.8	0.05503	0.00194	1.00

Table 6: NCDIF power calculated for particular item difficulties under the 1PL model using the IPR approach.

```
R> set.seed(29834528)
R> cutoffPoints <- CutoffIpr(quantiles = 0.95,
+    itemParameters = nullParameters, itemCovariances = nullAse,
+    irtModel = "1pl", logistic = FALSE, statistic = "ncdif",
+    nReplicates = nReplicates)
```

Next, the respective variances of item difficulties under the alternative hypothesis for each item are obtained, and the Monte Carlo chains of NCDIF indexes under the specified differences are obtained through the IPR approach.

```
R> altAse <- list()
R> altAse[["focal"]] <- AseIrt(itemParameters = raschParameters[["focal"]],
+    distribution = "norm", sampleSize = nFocal,
+    distributionParameters = distriParam, irtModel = "1pl",
+    logistic = FALSE)
R> altAse[["reference"]] <- AseIrt(itemParameters =
+    raschParameters[["reference"]], distribution = "norm",
+    sampleSize = nReference, distributionParameters = distriParam,
+    irtModel = "1pl", logistic = FALSE)
R> altIPR <- Ipr(itemParameters = raschParameters, itemCovariances = altAse,
+    nReplicates = nReplicates)
R> altNcdif <- IprNcdif(itemParameterList = altIPR, irtModel = "1pl",
+    logistic = FALSE)
```

Table 6 shows the calculated power for each item given the sample sizes and ability distributions. The column "Cut-off" presents the IPR cut-off points with the modified algorithm and 3000 replications. The column "Power" contains the calculated power to detect the true NCDIF with sample sizes of 500 and 1500 for the focal and reference groups, respectively.

5. Final remarks

The differential functioning of items and tests (DFIT) framework has been proposed as a parametric IRT alternative for DIF detection (Raju *et al.* 1995). Recent work has focused on the development of a Monte Carlo approach to obtain appropriate cut-off points for the

NCDIF index in this framework (Oshima *et al.* 2006; Raju *et al.* 2009), on the type I error and power of this approach (Clark and LaHuis 2012), and on the establishment of effect sizes for that index (Wright 2011; Wright and Oshima 2015).

This paper presented the package **DFIT** which implements the framework for R. The package is capable of obtaining the DIF and DTF indices from the framework for the main unidimensional IRT models. It is able to calculate Monte Carlo cut-off points based on the IPR approach for dichotomous and polytomous models as currently available software given the same input from the user; that is, item parameters for both groups and variance-covariance matrices for item estimates for one or both groups.

It should be noted that the package is recommended based on its capabilities and on its accuracy with regards to the theoretical framework. Comparisons with the program **DFIT8** (Oshima *et al.* 2009) and with the SAS (SAS Institute Inc. 2013) macro **DIFCUT** (Nanda, Oshima, and Gagné 2006), which implement the current approach, have been conducted by taking published item parameters (Raju *et al.* 1995, 2009; Oshima *et al.* 1998, 2009; Wright and Oshima 2015) and verifying against the results for all three DFIT indices; the same testing procedure was employed for Raju's area measures (Raju 1988), the Mantel-Haenszel statistic (Wright 2011), and the implementation of asymptotic variance-covariance matrices calculations (Li and Lissitz 2004); additionally, the simulation procedure relies on the package **mvtnorm** (Genz, Bretz, Miwa, Mi, and Hothorn 2016), whose accuracy was studied by Mi, Miwa, and Hothorn (2009). Although the accuracy for the IPR algorithms follows from the previous, cut-off points for NCDIF were also compared with published results; differences were consistent with differences between repeated runs of the algorithm such as those shown in Table 4.

The package **DFIT** is also able to use either version of the IPR approach without empirical estimates of the variance-covariance matrices by using the asymptotic ones for dichotomous models. Additionally, the package is flexible and allows obtaining the cut-off points under the null hypothesis and sampling conditions specified by the user. Finally, the package complements the DFIT framework by making it the first DIF approach for which power may be calculated a priori, and thus be used in the planning of DIF studies, rather than post-hoc under limited simulation conditions as has been the case until now.

There are several features yet to be implemented in the package **DFIT** from the eponymous framework. Future work on the package will focus on acquiring item parameter estimates and their covariances from different estimation packages to be used for the calculations; calculating the asymptotic variance-covariance matrices for polytomous models; implementing Differential Bundle statistics; calculating cut-off points at the test level (DTF); allowing each item on a set to be modeled by a different IRT model; and implementing the indices for multivariate IRT models.

References

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Cervantes VH (2012). "On Using the Item Parameter Replication (IPR) Approach for Power Calculation of the Noncompensatory Differential Item Functioning (NCDIF) Index." In

C Arce, G Seoane (eds.), *V European Congress of Methodology – Book of Abstracts*, pp. 206–207. Universidade de Santiago de Compostela, Santiago de Compostela.

Cervantes VH (2017). ***DFIT**: An R Package for the Differential Functioning of Items and Tests Framework*. Instituto Colombiano para la Evaluacion de la Educacion [ICFES], Bogotá, Colombia. R package version 1.0-3, URL https://CRAN.R-project.org/package=DFIT.

Chalmers RP (2012). "**mirt**: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software*, **48**(6), 1–29. doi:10.18637/jss.v048.i06.

Choi SW, Gibbons LE, Crane PK (2011). "**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*, **39**(8), 1–30. doi:10.18637/jss.v039.i08.

Clark PC, LaHuis DM (2012). "An Examination of Power and Type I Errors for Two Differential Item Functioning Indices Using the Graded Response Model." *Organizational Research Methods*, **15**(2), 229–246. doi:10.1177/1094428111403815.

Cohen AS, Kim SH, Baker FB (1993). "Detection of Differential Item Functioning in the Graded Response Model." *Applied Psychological Measurement*, **17**(4), 335–350. doi:10.1177/014662169301700402.

Frick H, Strobl C, Leisch F, Zeileis A (2012). "Flexible Rasch Mixture Models with Package **psychomix**." *Journal of Statistical Software*, **48**(7), 1–25. doi:10.18637/jss.v048.i07.

Frick H, Strobl C, Zeileis A (2015). "Rasch Mixture Models for DIF Detection: A Comparison of Old and New Score Specifications." *Educational and Psychological Measurement*, **75**(2), 208–234. doi:10.1177/0013164414536183.

Genz A, Bretz F, Miwa T, Mi X, Hothorn T (2016). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 1.0-5, URL https://CRAN.R-project.org/package=mvtnorm.

Holland PW, Thayer DT (2009). "Differential Item Performance and the Mantel-Haenszel Procedure." In H Wainer, HI Braun (eds.), *Test Validity*, pp. 129–145. Routledge, New York. Reprinted. Original work published in 1988.

Komboz B, Strobl C, Zeileis A (2017). "Tree-Based Global Model Tests for Polytomous Rasch Models." *Educational and Psychological Measurement*. doi:10.1177/0013164416664394. Forthcoming.

Li YH, Lissitz RW (2004). "Applications of the Analytical Derived Asymptotic Standard Errors of Item Response Theory Item Parameter Estimates." *Journal of Educational Measurement*, **41**(2), 85–117. doi:10.1111/j.1745-3984.2004.tb01109.x.

Magis D, Beland S, Tuerlinckx F, De Boeck P (2010). "A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning." *Behavior Research Methods*, **42**(3), 847–862. doi:10.3758/brm.42.3.847.

Mair P, Hatzinger R (2007). "Extended Rasch Modeling: The **eRm** Package for the Application of IRT Models in R." *Journal of Statistical Software*, **20**(9), 1–20. doi:10.18637/jss.v020.i09.

Mair P, Hatzinger R, Maier MJ (2016). ***eRm**: Extended Rasch Modeling.* R package version 0.15-7, URL https://CRAN.R-project.org/package=eRm.

Mi X, Miwa T, Hothorn T (2009). "**mvtnorm**: New Numerical Algorithm for Multivariate Normal Probabilities." *The R Journal*, **1**(1), 37–39.

Nanda AO, Oshima TC, Gagné P (2006). "**DIFCUT**: A SAS/IML Program for Conducting Significance Tests for Differential Functioning of Items and Tests (DFIT)." *Applied Psychological Measurement*, **30**(2), 150–151. doi:10.1177/0146621605280971.

Oshima TC, Kushubar S, Scott JC, Raju NS (2009). ***DFIT8** for Windows User's Manual: Differential Functioning of Items and Tests.* Saint Paul. Assessment Systems Corporation.

Oshima TC, Morris SB (2008). "Raju's Differential Functioning of Items and Tests (DFIT)." *Educational Measurement: Issues and Practice*, **27**(3), 43–50. doi:10.1111/j.1745-3992.2008.00127.x.

Oshima TC, Raju NS, Flowers CP (1997). "Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of Items and Tests." *Journal of Educational Measurement*, **34**(3), 253–272. doi:10.1111/j.1745-3984.1997.tb00518.x.

Oshima TC, Raju NS, Flowers CP, Slinde JA (1998). "Differential Bundle Functioning Using the DFIT Framework: Procedures for Identifying Possible Sources of Differential Functioning." *Applied Measurement in Education*, **11**(4), 353–369. doi:10.1207/s15324818ame1104_4.

Oshima TC, Raju NS, Nanda AO (2006). "A New Method for Assessing the Statistical Significance in the Differential Functioning of Items and Tests (DFIT) Framework." *Journal of Educational Measurement*, **43**(1), 1–17. doi:10.1111/j.1745-3984.2006.00001.x.

Preinerstorfer D (2016). ***mRm**: An R Package for Conditional Maximum Likelihood Estimation in Mixed Rasch Models.* R package version 1.1.6, URL http://CRAN.R-project.org/package=mRm.

Raju NS (1988). "The Area Between Two Item Characteristic Curves." *Psychometrika*, **53**(4), 495–502. doi:10.1007/bf02294403.

Raju NS, Fortmann-Johnson KA, Kim W, Morris SB, Nering ML, Oshima TC (2009). "The Item Parameter Replication Method for Detecting Differential Functioning in the Polytomous DFIT Framework." *Applied Psychological Measurement*, **33**(2), 133–147. doi:10.1177/0146621608319514.

Raju NS, Van der Linden W, Fleer PF (1995). "IRT-Based Internal Measures of Differential Functioning of Items and Tests." *Applied Psychological Measurement*, **19**(4), 353–368. doi:10.1177/014662169501900405.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rizopoulos D (2006). "**ltm**: An R Package for Latent Variable Modelling and Item Response Theory Analyses." *Journal of Statistical Software*, **17**(5), 1–25. doi:10.18637/jss.v017.i05.

Roussos L, Schnipke D, Pashley P (1999). "A Generalized Formula for the Mantel-Haenszel Differential Item Functioning Parameter." *Journal of Educational and Behavioral Statistics*, **24**(3), 293–322. doi:10.3102/10769986024003293.

SAS Institute Inc (2013). *The SAS System, Version 9.4*. SAS Institute Inc., Cary. URL http://www.sas.com/.

Strobl C, Kopf J, Zeileis A (2015). "Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model." *Psychometrika*, **80**(2), 289–316. doi:10.1007/s11336-013-9388-3.

Wright KD (2011). "Improvements for Differential Functioning of Items and Tests (DFIT): Investigating the Addition of Reporting an Effect Size Measure and Power." Unpublished Doctoral Dissertation, Georgia State University.

Wright KD, Oshima TC (2015). "An Effect Size Measure for Raju's Differential Functioning of Items and Tests." *Educational and Psychological Measurement*, **75**(2), 338–358. doi:10.1177/0013164414532944.

Affiliation:

Víctor H. Cervantes
Subdirección de Estadística
Instituto Colombiano para la Evaluación de la Educación – ICFES
Bogotá, Colombia
E-mail: vhcervantesb@unal.edu.co
and
Department of Psychological Sciences
Purdue University
West Lafayette, IN, United States of America
E-mail: cervantv@purdue.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2017, Volume 76, Issue 5 *Submitted:* 2013-07-22
doi:10.18637/jss.v076.i05 *Accepted:* 2015-11-06