Journal of Statistical Software

July 2014, Volume 59, Code Snippet 1. http://www.jstatsoft.org/

deltaPlotR: An R Package for Differential Item Functioning Analysis with Angoff's Delta Plot

David Magis
University of Liège

Bruno Facon
Université Lille-Nord de France

Abstract

Angoff's delta plot is a straightforward and not computationally intensive method to identify differential item functioning (DIF) among dichotomously scored items. This approach was recently improved by proposing an optimal threshold selection and by considering several item purification processes. Moreover, to support practical DIF analyses with the delta plot and these improvements, the R package **deltaPlotR** was also developed. The purpose of this paper is twofold: to outline the delta plot by describing the original method and its recent improvements in a user-oriented way, and to illustrate the structure and performances of the **deltaPlotR** package. A real data set about language skill assessment is being analyzed as an illustrative example.

Keywords: differential item functioning, Angoff's delta plot, modified delta plot, R package.

1. Introduction

The investigation for differential item functioning (DIF) in psychological and educational assessment has become a broad field of research, both at the theoretical/methodological level and as a practical problem for test administration and evaluation. It exists an abundant literature about DIF, and reviews of existing DIF detection methods are provided by Magis, Béland, Tuerlinckx, and De Boeck (2010), Osterlind and Everson (2009) and Penfield and Camilli (2007), among others.

The increase in computer resources in the last decades also played a role in the expansion of this research field. But as more and more complex DIF detection methods are being developed, mainly to match the increasing complexity of assessment studies, available software for practical DIF analyses remained quite limited. Moreover, most software programs do not incorporate the latest developments and improvements of DIF methods. This issue is not trivial, since it may lead the researcher to favor another DIF method, possibly less powerful

or adequate, even though more flexible methods have been proposed.

Currently there are only a few software packages for DIF detection available. Some well established software packages such as **BILOG-MG** (Zimowski, Muraki, Mislevy, and Bock 1996) and **ConQuest** (Wu, Adams, and Wilson 1997) provide some standard DIF statistics based on item response theory (IRT) models. In terms of specific DIF software, one can mention **DFITPU** (Raju 1995), **DICHODIF** (Rogers, Swaminathan, and Hambleton 1993), **DIFAS** (Penfield 2001), **IRTDIF** (Kim and Cohen 1992), **IRTLRDIF** (Thissen 2001), **SIBTEST** (Li and Stout 1994), and the R packages **difR** (Magis *et al.* 2010; Magis, Béland, and Raiche 2013) and **lordif** (Choi, Gibbons, and Crane 2011).

This paper focuses on DIF investigation of dichotomously scored items across two groups of respondents (further referred to as the *reference* and the *focal* groups). Several standard methods in this framework are: the Mantel-Haenszel method (Holland and Thayer 1988) and logistic regression (Swaminathan and Rogers 1990) as score-based methods, and Lord's chi-square test (Lord 1980) and the likelihood-ratio test (Thissen, Steinberg, and Wainer 1988) as IRT-based methods. However, in this paper a more ancient method is considered, the so-called delta plot method (or shortly *delta plot*), suggested by Angoff (Angoff 1972; Angoff and Ford 1973). It is a simple score-based method that aims at comparing the proportions of correct responses in the reference group and the focal group, by an appropriate non-linear transformation of these proportions. It is a straightforward and not computationally intensive method with an appealing graphical output (the so-called *diagonal plot*). Moreover, unlike most conventional DIF methods, there is no restriction to applying the delta plot with small samples of respondents, either in one of the groups or in both groups (Magis and Facon 2012). Indeed, the delta plot does not rely on sophisticated statistics with known asymptotic distribution, such as e.g., the Mantel-Haenszel method or logistic regression. Its practical usefulness and interest for practitioners can be assessed by recent DIF studies using the delta plot (Abedalaziz 2010; Facon, Magis, and Courbois 2012a; Facon and Nuchadee 2010; Facon, Nuchadee, and Bollengier 2012b; Michaelides 2010; Moon, McLean, and Kaufman 2003; Robin, Sireci, and Hambleton 2003; Sireci and Allalouf 2003; Sireci, Patsula, and Hambleton 2005; Van Herwegen, Farran, and Annaz 2011).

In addition to its recently increasing use, the delta plot was also re-considered from a methodological point of view. The original method suffered from a lack of accurate selection of the DIF detection threshold. Early publications on that method (see e.g., Rudner 1977, 1978) mentioned some fixed value for this detection threshold, but it appeared that such a fixed-threshold approach was most often too conservative in the presence of DIF, and that a data-driven rule to detect DIF items was more appropriate (Magis and Facon 2012). Furthermore, the impact of DIF items on the results of the delta plot has been investigated only recently, by focusing on iterative schemes known as item purification processes in the DIF literature (Candell and Drasgow 1988) and adapting them to the delta plot (Magis and Facon 2013a).

It is also worth mentioning that the delta plot is a *relative DIF* method, in the sense that DIF items are flagged with respect to the set of all items in the test. This approach is similar to another recently introduced DIF method called the outlier DIF approach (Magis and De Boeck 2012). The main asset of these methods is that the identification of DIF items relies on the particular items themselves, while traditional DIF methods make use of fixed detection thresholds arising only from asymptotic statistical distributions. Relative DIF methods were shown to be superior for controlling type I error inflation (Magis and De Boeck 2012, 2014), which also motivates further consideration of the delta plot.

This paper is targeting a double objective. First, the delta plot is briefly sketched together with its recent methodological improvements to provide a clear and complete description of the method, its assets and drawbacks, and rules of application for interested practitioners. Second, to provide a computer software in support for this DIF technique, the paper describes a recently developed R package called **deltaPlotR**, that encompasses all aspects of the delta plot. The main structure of the package is presented with emphasis on the arguments which need to be specified in practical application and output. The delta plot is also illustrated through the analysis of a real example on English skill assessment. Some R code and output is included and discussed for further understanding of the delta plot and the software.

2. The delta plot and its improvement

This section sketches the original delta plot and its recent improvements. Further detailed information can be found in e.g., Angoff and Ford (1973), Holland and Wainer (1993) and Magis and Facon (2012).

The basic idea behind the delta plot is the comparison of proportions of correct responses per item and in each group of respondents. Consider a test made of J dichotomously scored items and for any item j ($j = 1, \ldots, J$) set p_{jk} as the proportion of correct responses in group k (where for convenience $k = 0$ for the reference group and $k = 1$ for the focal group). These proportions are then first transformed into standard normal deviates: $z_{jk} = q_z(1 - p_{jk})$ with $q_z(.)$ standing for the quantile of the standard normal distribution. Secondly, the normal deviates are finally transformed into delta scores with the following linear relationship: $\Delta_{jk} = 4z_{jk} + 13$. The latter transformation ensures delta scores are centered on the value 13 and range between zero and 26, if the proportions p_{jk} are in [0.00058, 0.99942] (Angoff and Ford 1973).

Each item j therefore has a pair of delta scores (Δ_{j0}, Δ_{j1}), referred to as the delta point. These delta points can be displayed in a scatter plot, called the *diagonal plot*, with the delta scores of the reference group on the X axis and the delta scores of the focal group on the Y axis. The plot usually takes the form of an elliptical cloud of delta points, from lower left corner to upper right corner of the scatter plot. As mentioned by Angoff and Ford (1973, p. 97), correlations between delta scores of .98 or .99 are expected with comparable groups of respondents. Differences in average group abilities will lead the delta points to lie either below or above the identity line, depending on the direction of the mean group difference. However, items that substantially depart from this diagonal plot (that is, from the main axis of this ellipsoid) can be flagged as DIF (Angoff and Ford 1973). Note that lower correlations between the delta scores are expected in presence of DIF items, but even in absence of DIF, practical situations with low correlations can occur. Although such cases have not been studied so far, the main framework of delta plot remains applicable, as described below.

In practice, assessment of DIF is performed in three simple steps: (a) the major axis of the ellipsoid is determined by computing its intercept and slope; (b) the perpendicular distances of all delta points to the major axis are computed; (c) items with a perpendicular distance larger than an acceptable threshold value are flagged as DIF.

More precisely, the major axis of the ellipsoid has the following equation:

$$\Delta_{j1} = a + b\,\Delta_{j0}, \tag{1}$$

with

$$b = \frac{{s_1}^2 - {s_0}^2 + \sqrt{({s_1}^2 - {s_0}^2)^2 + 4\,{s_{01}}^2}}{2\,s_{01}} \quad \text{and} \quad a = \overline{x}_1 - b\,\overline{x}_0, \tag{2}$$

and $\overline{x}_0$, $\overline{x}_1$, ${s_0}^2$, ${s_1}^2$ and s_{01} are respectively the sample means, sample variances, and sample covariance of the delta scores.

Now, for any item j, the perpendicular distance D_j between the major axis given in Equation 1 and the delta point $(\Delta_{j0}, \Delta_{j1})$ is computed as follows:

$$D_j = \frac{b\,\Delta_{j0} + a - \Delta_{j1}}{\sqrt{b^2 + 1}}. \tag{3}$$

The item is flagged as DIF as soon as $|D_j|$ is larger than some selected detection threshold. Most often, the fixed threshold 1.5 is considered (Baghi and Ferrara 1989; Facon and Nuchadee 2010; Facon *et al.* 2012b; Muniz, Hambleton, and Xing 2001; Osterlind 1983; Robin *et al.* 2003; Sireci *et al.* 2005; Van Herwegen *et al.* 2011). Other authors, such as Merz and Grossen (1979), Michaelides (2010) and Rudner (1977, 1978), proposed different rules-of-thumb to get a suitable detection threshold.

The delta plot is a simple and straightforward method that does not require intensive computing power and resources. Moreover, it can handle missing responses as they are not included in the computation of the proportions p_{jk}. The method is also useful with small samples of respondents, as it does not rely on asymptotic assumptions or statistical distributions. Its main drawback, however, is the selection of an appropriate DIF detection threshold. Instead of a fixed detection threshold, Magis and Facon (2012) proposed to derive it by using a normality assumption on the delta points (this is further referred to as the *normal approximation*). More precisely, they assumed the sample of delta points arises from a bivariate normal distribution. Starting from this hypothesis, they derived the following threshold T_α:

$$T_\alpha = z_{1-\alpha/2} \sqrt{\frac{b^2\,{s_0}^2 - b\,{s_{01}}^2 + {s_1}^2}{b^2 + 1}}, \tag{4}$$

which depends now on the significance level α, the slope of the major axis b and the sample variances ${s_0}^2$, ${s_1}^2$ and covariance s_{01}. Thus, instead of an arbitrarily fixed threshold, T_α depends on the set of delta scores and the shape of the delta points (through the sample variances and covariance and the slope of the major axis) and on the significance level. This threshold was shown to outperform the standard fixed-threshold approach, by avoiding conservatism and by increasing the power to detect DIF items.

Nevertheless, in the presence of DIF items, it may be expected that the test and item scores, and consequently the proportions of correct responses, will be influenced in some way. This may affect the conclusions of the delta plot, even when replacing the fixed threshold by its normal approximation given by Equation 4. Most often, for test-score based approaches such as the Mantel-Haenszel method (Holland and Thayer 1988), an iterative process is performed to reduce the impact of DIF items onto the total test scores. This process is called item purification (Candell and Drasgow 1988). It consists of an iterative re-application of the method by discarding items currently flagged as DIF from the computation of test scores. This iterative process aims at sequentially reducing the impact of DIF on the test results, and getting therefore "purified" results.

This approach was recently adapted to the delta plot (Magis and Facon 2013a). In this framework, the item purification process (IPP) can be sketched as follows:

(a) Re-compute the intercept and slope parameters of the major axis by discarding items that are currently flagged as DIF.

(b) Re-compute the perpendicular distances of all items by plugging-in the updated intercept and slope parameters.

(c) Possibly re-compute the DIF selection threshold (see below).

(d) Test for DIF by using the newly computed perpendicular distances and detection threshold. Stop if the items currently flagged as DIF correspond exactly to those identified at the previous step. Otherwise go to Step (a).

The main difference between this process and the usual IPP with other test-score based DIF methods, is that the detection threshold can also be updated, especially the normal approximation of Equation 4 as it depends on the set of delta points (through the sample estimates and the slope of the major axis). Three scenarios were considered by Magis and Facon (2013a). In the first scenario (called IPP1), the detection threshold is kept constant through the iterative steps and equal to the value obtained after the first run of the delta plot. In the second scenario (IPP2), only the slope parameter b is being updated in the normal approximation, the sample estimates of variance and covariance being kept equal to their original values (obtained with the full set of items). Finally, the third scenario (IPP3) makes a full update of all possible parameters after each step, that is, the slope parameter b and the sample variances ${s_0}^2$, ${s_1}^2$ and covariance s_{01}.

Although several item purification approaches are allowed, Magis and Facon (2013a) highlighted that none of them was clearly outperforming the delta plot with normal approximation of the threshold, without purification. This was explained by the fact that making use of the modified threshold T_α is a major improvement of the delta plot and no further purification process is required to significantly improve the DIF detection results. This study, however, was limited to the context of small samples of respondents, and further investigation would obviously become necessary with large samples. Note that, though being an iterative process, the routine discussed above does not require intensive computations though the calculations would become difficult to perform by hand.

3. The deltaPlotR package

To support practical DIF analysis with the delta plot, and to allow the interested practitioner to get access to the recently suggested improvements to the method, an R (R Core Team 2014) package called **deltaPlotR** (Magis and Facon 2013b) was created. It encompasses all aspects of the delta plot approach described in the previous section, allows for flexibility in data input, and provides both descriptive and summary statistics and graphical output. This section briefly describes its functionality, while a practical illustration (with R code) is provided in the next section.

3.1. Structure

The **deltaPlotR** package is built around a central function, called `deltaPlot`. This function requires some input data in a specific format and several operating functions that define,

among others, the type of item purification and the detection threshold. All calculations are performed internally and returned as object with a user-friendly display. This output can also be saved into a text file for further use. Note that the package also includes a graphical function, called `diagPlot`, which permits to draw the diagonal plot and control for graphical options.

3.2. The `deltaPlot` function

The `deltaPlot` function has several mandatory and optional arguments that are listed in Table 1 and detailed hereafter.

The input data must be provided through the `data` argument and can have any of the following two specific formats. The first allowed format is a matrix with one row per respondent and $J + 1$ columns, where J is the number of test items. Item responses are coded as zeros and ones, and missing responses are allowed but must be coded as `NA`. One of the columns must code for the group membership, with two different values (either numeric or character or as a factor) for the reference group and the focal group. All columns may have names, and the names of the item columns will be used as item names. The group membership column can also be characterized by a column name. For this format, the `type` argument must be set to `"response"` (which is its default value).

The second data format is a $J \times 2$ matrix, with one row per item and the first and second column coding for the reference and the focal groups respectively. In this case, the matrix must hold either the proportions of correct responses or the delta scores for each item and each group of respondents. The `type` argument permits to distinguish between proportions-correct and delta scores, by setting its value to `"prop"` or `"delta"` respectively. This input format is particularly useful when the original data are not available anymore and one has only access to summary values such as the proportions of correct responses or delta scores.

If the data are passed through the full data matrix, the identification of the group membership variable is performed by the `group` argument. It takes as value, either the column number (starting from one for the first column to $J + 1$ for the last column) or the column name if any. The column number can always be provided even if the data matrix has column names. Moreover, the reference and focal groups are distinguished by means of the `focal.name` argument, which takes the (numeric or character) value flagging for focal group response patterns. Note that if the data input is not of the `"response"` type, these two arguments are useless.

Once the data are loaded and ready to be analyzed, several working options must be set. Note that all options have default values so the package can run without further specifications. First, the detection threshold must be specified through the `thr` argument. By default, it takes the value `"norm"`, so that the threshold is derived by using the normal approximation of Equation 4. However, any positive numeric value can be specified, for instance the commonly used 1.5 value. For the normal approximation, the significance level must also be provided through the `alpha` argument. Its default value is 5% and is of course ignored if a numeric value was provided to `thr`.

Item purification can be performed by setting the `purify` argument to `TRUE`, however this is not the case by default. In case of item purification, two further options can be set by the user: the maximal number of iterations allowed for the process before stopping, determined by the `maxIter` argument (with default value 10), and the type of IPP, set by the `purType`

Argument	Role	Value	Default	Ignored if
`data`	Specifies the data input object.	data matrix	`NA`	`NA`
`type`	Specifies the type of input `data`.	`"response"`, `"prop"` or `"delta"`	`"response"`	`NA`
`group`	Group membership locator.	column name or number	`NA`	`"type"` is not `"response"`
`focal.name`	Focal group identifier.	numeric or character	`NA`	`"type"` is not `"response"`
`thr`	Specifies the DIF flagging threshold.	numeric or `"norm"`	`"norm"`	`NA`
`purify`	Should item purification be performed?	logical	`FALSE`	`NA`
`purType`	Specifies the type of IPP.	`"IPP1"`, `"IPP2"` or `"IPP3"`	`"IPP1"`	`purify` is `FALSE`
`maxIter`	Sets the maximal number of IPP iterations.	integer	10	`purify` is `FALSE`
`alpha`	Fixes the significance level.	numeric in $]0, 1[$	0.05	`thr` is not `"norm"`
`extreme`	Specifies the constraint for proportions-correct.	`"constraint"` or `"add"`	`"constraint"`	`type` is not `"response"`
`const.range`	Sets the constraint interval for proportions.	vector of 2 numeric values	`c(0.001, 0.999)`	`type` is not `"response"` or `extreme` is `"add"`
`nrAdd`	Sets the n parameter.	integer	1	`type` is not `"response"` or `extreme` is `"constraint"`
`save.output`	Should output be saved?	logical	`FALSE`	`NA`
`output`	Options for saving output.	name and path	`c("out", "default")`	`save.output` is `FALSE`

Table 1: Arguments, roles, values and default values of the `deltaPlot` function.

argument. The three possible values of `purType` naturally refer to the three aforementioned processes, namely `"IPP1"`, `"IPP2"` and `"IPP3"`. The default process is `"IPP1"`.

Finally, the delta plot analysis cannot be technically run when, for at least one item and one group of respondents, the proportion of correct responses is either exactly zero or one. This is because in this case the z score, and consequently the delta score, takes an infinite value. To overcome this issue, two approaches are possible. First, the range of proportions of correct responses can be shrunk to an interval close to $[0, 1]$, for instance $[0.001; 0.999]$. That is, any proportion of correct responses outside this range is converted into 0.001 or 0.999 according to its original value. The second constraining option consists in virtually adding n correct and n incorrect responses, so that, for instance, a proportion of zero is converted into $n/(n_k + 2n)$ for instance (where n_k stands for the number of respondents in group k). Setting $n = 1$ yields the so-called Laplace rule (Jaynes 2003), but in practice any integer value can be specified.

In the **deltaPlotR** package, the `extreme` argument permits to select between these two approaches. Its allowable values are `"constraint"` (the default value) to restrict extreme proportions to some interval, and `"add"` to add n correct and n incorrect responses. In the first case, the constraining interval is specified by the `const.range` argument, with default value `c(0.001, 0.999)` (that is, the $[0.001; 0.999]$ interval). In the second case, the parameter$n$ is specified through the `nrAdd` argument, with default value one to apply the Laplace rule. Note that these arguments are useful only if the data are provided as a full binary matrix. Otherwise, either the proportions of correct responses need to be away from 0 and 1 or the delta scores must have finite values, and the `constraint`, `const.range` and `nrAdd` arguments are ignored.

3.3. Output

The output of the `deltaPlot` function is returned as a list of class '`deltaPlot`' that can be used for further analyses or examination. This output will by default be printed on the screen in an optimized fashion, through the S3 `print` method for '`deltaPlot`' objects, to facilitate its interpretation.

Several summary statistics are returned:

1. the proportion of correct responses (if the input data did not contain the delta scores only),
2. the delta scores for each group of respondents and the perpendicular distances from the major axis,
3. the parameters (intercept and slope) of the major axis,
4. the type of detection threshold and its value,
5. the items that were flagged as DIF (if any).

If item purification was performed, some additional output information is printed: the number of iterations required to ensure convergence (completed by a warning message if convergence was not reached), and the initial and final values of the perpendicular distances and the major axis parameters. Note that all intermediate steps can also be printed, by setting the `only.final` argument of the S3 `print` method for '`delatPlot`' objects to `FALSE`. This is

Argument	Role	Value	Default
`x`	Sets the output from `deltaPlot`.	object of class 'deltaPlot'	`NA`
`pch`	Sets the type of delta point.	integer	2
`pch.mult`	Sets the symbol for superposed delta points.	integer	17
`axis.draw`	Should major axis be drawn?	logical	`TRUE`
`thr.draw`	Should detection thresholds be drawn?	logical	`FALSE`
`dif.draw`	Sets options for DIF identification.	vector of two numeric values	`c(1, 3)`
`print.corr`	Should delta scores correlation be printed?	logical	`FALSE`
`xlim`, `ylim`, `xlab`, `ylab`	Usual axis limits and labels options.	vectors of two numeric values	`NULL`
`save.plot`	Should the plot be saved?	logical	`FALSE`
`save.options`	Sets options for saving the plot.	name, path and and extension	`c("out", "default", "pdf")`

Table 2: Arguments, roles and values of the `diagPlot` function.

helpful for tracking the successive changes in DIF detection across each step of the purification process.

Finally, this output can be locally saved into a text file, by specifying the target folder to save the output and the name of the output file. To save the output in a text file, the `save.output` argument of the `deltaPlotR` function must be set to `TRUE`, its default value being `FALSE` and thus preventing from saving the output. Moreover, the output file name and path are specified through the `output` argument, as a two-component vector with the name and the file path as character components. By default, `output` takes the value `c("out","default")`, meaning that the output will be saved in the default folder with the name *out.txt*.

3.4. The `diagPlot` function

The output of the `deltaPlot` function can also be displayed in a graphical format, by creating the diagonal plot and optionally saving it to a JPEG or a PDF file. This can be done with the function called `diagPlot` which is also included in the `deltaPlotR` package. Table 2 lists the different arguments of this `diagPlot` function.

The `diagPlot` function uses the output of `deltaPlot` as input data through the argument `x` and controls for several graphical parameters. First, the delta points are displayed with a symbol that can be specified through the `pch` argument. Its default value of 2 makes use of triangles to display the points. If several items are located onto exactly the same delta point, the `pch.mult` argument allows for a different drawing of this delta point. By default, it is fixed to 17, that is, a full black triangle.

Two types of lines can be drawn: the major axis of the ellipsoid and the upper and lower DIF detection thresholds. The major axis is drawn with a solid line if the `axis.draw` argument is set to `TRUE`, and the DIF thresholds (as dashed lines) by setting the `thr.draw` to `TRUE`. By

default, only the major axis is drawn.

Items that are flagged as DIF can be graphically identified by superposing an additional type of point symbol. Both the symbol type and size can be specified through the `dif.draw` argument, by a two-component vector with the numbers of the point type and the point size, respectively. By default, this argument takes the value `c(1, 3)`, meaning that a three-times sized circle is drawn to surround the delta points of the items flagged as DIF.

The last two available options of `diagPlot` are the following. First, the correlation between the delta scores can be printed in the upper left corner of the plot, by specifying the logical argument `print.corr` to `TRUE`. By default, this correlation is not printed. Second, the axis limits and labels can be specified through the usual `xlim`, `ylim`, `xlab` and `ylab` arguments. By default, their `NULL` value yields an automatic selection of axis limits, and the axis labels are fixed to *Reference group* and *Focal group* respectively.

Finally, the diagonal plot can be saved as a figure, by specifying its name, its location folder and its extension type. Requesting for saving the plot is performed by the argument `save.plot` that has to be set to `TRUE` (default is `FALSE`, that is, the plot is not saved). In case of saving the plot, the three saving arguments (name, location, extension) must be provided, in this order, as character values to the `save.options` argument. Note that the extension can be either `"pdf"` or `"jpeg"`, so only PDF and JPEG files are produced. By default, the argument `save.options` takes the value `c("out", "default", "pdf")`, meaning that the plot is saved as *out.pdf* file in the default folder.

3.5. Availability

The **deltaPlotR** package can be downloaded freely from the Comprehensible R Archive Network (CRAN) at `http://CRAN.R-project.org/package=deltaPlotR`. Its currently available version is 1.4.

4. An example

We shall illustrate the usefulness and flexibility of the **deltaPlotR** package by analyzing a real data set about language skill assessment, using two different settings of the delta plot and with both text and figure outputs.

The selected data set comes from the TCALS-II questionnaire, a language skill assessment test for French speaking Canadian students from the province Quebec (Laurier, Froio, Pearo, and Fournier 1998). This test was developed to assess the level of English, as a second language, of Canadian French-speaking students prior to entering into college education in Quebec province. The TCALS-II questionnaire consists of 85 items, merged into five subcategories (such as listening comprehension and reading) and was first administered in 1998. Further information about this questionnaire, as well as results from dimensionality and fidelity analyzes, can be found in Laurier *et al.* (1998) and Raiche (2002).

In this application, the DIF analysis will be conducted between two different years of administration in the College of Outaouais (Gatineau, Quebec, Canada), the year 1998 (treated as the reference group) and the year 2000 (treated as the focal group). A total of 1,373 respondents were recorded in 1998 and 1,547 respondents were recorded in 2000. Moreover, in order to limit the amount of output in this illustration, the analysis will be limited to the first 33

items of the TCALS-II questionnaire, i.e., listening comprehension of sentences, dialogues and short texts.

The data set to be used as input for the R function `deltaPlot` contains the full binary responses, without missing values, and an additional column with group membership, coded as "1998" and "2000" to refer to the years of administration. Item names are simply coded as "Item1", "Item2" and so on. This yields a $2,920 \times 34$ matrix, saved into the `Data TCALS 1998-2000` text file. The data set is available as supplementary material from the journal web page.

In a first step, the working directory is defined by specifying its path. This working directory should contain the input data set and will contain all output files that are produced with the functions of the **deltaPlotR** package. For instance, setting the working directory as the current folder, the following R code can be used:

```
R> path <- "."
R> setwd(path)
```

After reading the data set into R and loading the **deltaPlotR** package, the DIF analysis can begin. In the first illustrative analysis, the delta plot is run without item purification, and the detection threshold is determined by using the normal approximation. The output is not saved at this step. The full R code is given below.

```
R> Data <- read.table("Data-TCALS-1998-2000.txt", header = TRUE)
R> library("deltaPlotR")
R> res <- deltaPlot(data = Data, type = "response", group = "Year",
+    focal.name = "2000", thr = "norm")
R> res
```

The output of this DIF analysis is displayed below, exactly as it appears on the R console. The first part of this output mentions the main settings of the analysis: no item purification and extreme proportions constrained to $[0.001; 0.999]$, though it was unnecessary here. The 33 DIF statistics (i.e., the perpendicular distances) are then listed, together with the proportions of correct responses and the delta scores in each group of respondents. One can notice that item 18 is pinned with a `***` symbol right next to its DIF statistic, which indicates at a glance that it was flagged as DIF. This is confirmed below with the last part of the output, returning the parameters (intercept and slope) of the major axis, the detection threshold as computed from the normal approximation (and related significance level), and the items that were flagged as DIF. A final sentence mentions that the output was not captured and saved into a text file.

```
Detection of Differential Item Functioning using Angoff's delta method
  without item purification

Extreme proportions adjusted by constraining to [0.001; 0.999]

Statistics:

       Prop.Ref Prop.Foc Delta.Ref Delta.Foc Dist.
```

```
Item1   0.9395   0.9005   6.7960    7.8635   -0.2255
Item2   0.9272   0.8933   7.1800    8.0220   -0.0884
Item3   0.9563   0.9263   6.1629    7.2046   -0.1549
Item4   0.9031   0.8552   7.8016    8.7639   -0.2286
Item5   0.8514   0.8410   8.8298    9.0060    0.2748
Item6   0.8602   0.8358   8.6758    9.0905    0.1093
Item7   0.8507   0.8177   8.8424    9.3733    0.0090
Item8   0.7240   0.6852  10.6214   11.0709   -0.0746
Item9   0.6919   0.6761  10.9949   11.1722    0.0982
Item10  0.6242   0.5953  11.7341   12.0347   -0.0538
Item11  0.5506   0.5003  12.4911   12.9968   -0.2683
Item12  0.5870   0.5307  12.1203   12.6918   -0.2874
Item13  0.8798   0.8552   8.3035    8.7639    0.1053
Item14  0.9680   0.9438   5.5939    6.6513   -0.1205
Item15  0.8653   0.8190   8.5830    9.3537   -0.1490
Item16  0.8806   0.8630   8.2890    8.6251    0.1993
Item17  0.8616   0.8106   8.6495    9.4795   -0.1987
Item18  0.8492   0.8487   8.8674    8.8758    0.3970 ***
Item19  0.8383   0.8016   9.0499    9.6113   -0.0307
Item20  0.8361   0.8009   9.0854    9.6206   -0.0140
Item21  0.6985   0.6606  10.9200   11.3432   -0.0793
Item22  0.7626   0.7227  10.1417   10.6366   -0.0696
Item23  0.7189   0.6626  10.6821   11.3220   -0.2217
Item24  0.5878   0.5495  12.1128   12.5029   -0.1513
Item25  0.3744   0.3704  14.2813   14.3232   -0.0674
Item26  0.7247   0.7059  10.6127   10.8344    0.0961
Item27  0.5222   0.5217  12.7772   12.7828    0.0818
Item28  0.8026   0.7880   9.5959    9.8023    0.1901
Item29  0.8252   0.7951   9.2585    9.7032    0.0396
Item30  0.7087   0.7111  10.8020   10.7741    0.2671
Item31  0.6227   0.6361  11.7494   11.6081    0.2749
Item32  0.6540   0.6606  11.4150   11.3432    0.2501
Item33  0.8121   0.7873   9.4575    9.8112    0.0913

Code: '***' if item is flagged as DIF

Parameters of the major axis:

      a      b
 1.5042 0.8913

Detection threshold: 0.3533 (significance level: 5%)

Items detected as DIF items:

 Item18
```

```
Output was not captured!
```

In a second analysis, and because one item exhibited DIF, the delta plot is run but now with item purification. The second type of purification (IPP2) was selected. Moreover, it was also requested to save the output into a text file, called `out.txt` and to be located in the working directory set by `path`. The corresponding R code for performing this analysis is displayed below.

```
R> res2 <- deltaPlot(data = Data, type = "response", group = "Year",
+    focal.name = "2000", thr = "norm", purify = TRUE, purType = "IPP2",
+    save.output = TRUE, output = c("out", path))
```

The output of this second analysis is partly displayed below. Only a selected part of the output is shown, corresponding to the sections of the output that differ from the previous analysis without purification.

```
Detection of Differential Item Functioning using Angoff's Delta method
  with item purification

Convergence reached after 2 iterations

Threshold adjusted iteratively using normal approximation
 and 5% significance level
 (only slope parameter updated [IPP2])

[SKIPPED OUTPUT]

Parameters of the major axis (first and last iterations only):

           a      b
First 1.5042 0.8913
Last  1.5713 0.8861

First and last detection thresholds: 0.3533 and 0.3536
 (significance level: 5%)

Items detected as DIF items:

 Item18

Output was captured and saved into file
 './out.txt'
```

First, the description of the analysis is modified by mentioning the IPP2 process. Second, an indication about the convergence of this process is displayed: only two iterations were necessary in this case (note that the first run of the delta plot is taken as the first iteration). Moreover, after displaying the DIF statistics, the major axis parameters and the detection

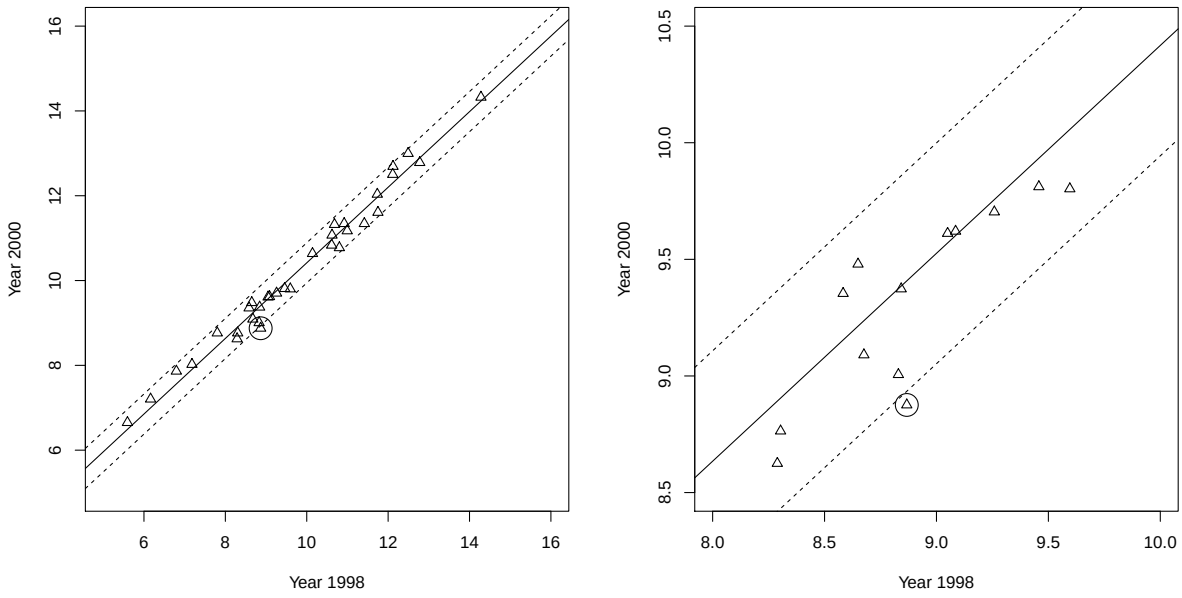

Figure 1: Graphical display of the output of the first DIF analysis.

threshold of the first and last iterations are printed. Finally, the items flagged as DIF are listed, and a note that the output was saved in the required file and folder is added.

As explained earlier, the `deltaPlot` function can also be used when the data are merged as a two-column dataset with either the proportions of correct responses per item or the delta scores. To illustrate these approaches, both the proportions of correct responses and the delta scores were computed from the 33 selected TCALS-II items. The data transformation and the first delta plot analysis (i.e., without purification) was reproduced with each set of values using the following R code:

```
R> Data_prop <- sapply(split(Data[, -1], Data$Year), colMeans)
R> Data_delta <- 4 * qnorm(1 - Data_prop) + 13
R> deltaPlot(data = Data_prop, type = "prop", thr = "norm")
R> deltaPlot(data = Data_delta, type = "delta", thr = "norm")
```

With respect to the first analysis using the full data set, the `type` argument has been updated according to the type of data input (i.e., either proportions or delta scores) and the `group` and `focal.name` arguments are not specified anymore (since the reference and focal groups are clearly identified as the first and second column of the data files, respectively). As expected, the two output results are identical to the one displayed above and are consequently not reproduced here.

To conclude this illustration, the diagonal plot for this data set was created with the `diagPlot` function. Both the major axis and the detection thresholds were displayed, and the X and Y labels were updated to fit the current analysis. The diagonal plot was eventually saved as a PDF file, with the name `figure.pdf` and stored in the working directory set by `path`. However, since all delta points tend to lie closely to the major axis, and because the band made by the detection thresholds around the major axis is quite narrow, a second diagonal plot was created by selecting the range of delta scores to $[8; 10]$ for the year 1998 and to

[8.5; 10.5] for the year 2000. This choice was made after visual inspection of the first diagonal plot, and in order to keep the item flagged as DIF into the figure while improving the visibility of this specific area of the plot. The full R code for these two plots is displayed hereafter.

```
R> diagPlot(res, axis.draw = TRUE, thr.draw = TRUE, xlab = "Year 1998",
+    ylab = "Year 2000", save.plot = TRUE, save.options = c("figure",
+    path, "pdf"))
R> diagPlot(res, axis.draw = TRUE, thr.draw = TRUE, xlab = "Year 1998",
+    ylab = "Year 2000", xlim = c(8, 10), ylim = c(8.5, 10.5))
```

The two diagonal plots are represented in Figure 1, the full plot being displayed on the left panel and the selected part of this plot on the right panel. It is noticeable that the item flagged as DIF lies outside the range of detection thresholds from the major axis, as expected. Moreover, by default the `diagPlot` function surrounds the item symbol by a big black circle to help in identifying those items exhibiting DIF. Note that the type and the size of this flagging symbol can be modified by specifying the appropriate argument of the `diagPlot` function.

5. Concluding remarks

This paper briefly presented the R package **deltaPlotR** that was developed to support dichotomous DIF analysis using the delta plot method. It integrates the latest methodological developments about this method and can handle full binary response data sets, or summarized data by means of the simple proportions of correct responses or delta scores. Several flexible options for threshold selection and computation, item purification, extreme proportion adjustment, and output saving are available. The diagonal plot can also be easily constructed, optimized, and saved as a figure file with standard graphical options.

The delta plot is a simple and easy-to-use DIF method, and the recent improvements about DIF threshold selection look promising. However, this approach could be less appealing than other standard DIF methods in some practical situations. For instance, low correlations between the delta scores might seriously impact the results of DIF identification, since delta scores would be somewhat dispersed and impact therefore the perpendicular distances. The usefulness of the delta plot might also be undermined if all items exhibit the same DIF size, possibly very large, since no clear departure from the major axis will be observed for some subset of items. Further research is necessary to determine the limits of applicability of this approach, and to evaluate it with respect to other traditional DIF methods in such situations.

The **deltaPlotR** package has two main assets. First, it incorporates many options for maximum flexibility and usefulness, including some very recent developments and improvements of the delta plot. Second, it is designed to easily integrate new developments from future or ongoing research. Its main drawback, however, stands in its limitation in the R software which is not commonly used by psychologists. It is expected nevertheless that future applications will handle an efficient interaction between some end-user interface and the R software.

Acknowledgments

The authors wish to thank two anonymous reviewers for their helpful comments. This research was supported by a post-doctoral grant “Chargé de recherches” from the National Funds for

Scientific Research (FNRS, Belgium), the Research Fund of KU Leuven (GOA/15/003), the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06), and the National Research Agency/Agence Nationale de la Recherche (France), ANR, LANG & HANDICAPS, Project no. ANR-09-ENFT-019. Correspondence should be sent to: David Magis.

References

Abedalaziz N (2010). "A Gender-Related Differential Item Functioning of Mathematics Test Items." *International Journal of Educational and Psychological Assessment*, **5**(1), 101–116.

Angoff WH (1972). "A Technique for the Investigation of Cultural Differences." Paper presented at the Annual Meeting of the American Psychological Association, Honolulu.

Angoff WH, Ford SF (1973). "Item-Race Interaction on a Test of Scholastic Aptitude." *Journal of Educational Measurement*, **10**(2), 95–106.

Baghi H, Ferrara S (1989). "A Comparison of IRT, Delta Plot, and Mantel-Haenszel Techniques for Detecting Differential Item Functioning Across Subpopulations in the Maryland Test of Citizenship Skills." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Candell GL, Drasgow F (1988). "An Iterative Procedure for Linking Metrics and Assessing Item Bias in Item Response Theory." *Applied Psychological Measurement*, **12**(3), 253–260.

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. URL http://www.jstatsoft.org/v39/i08/.

Facon B, Magis D, Courbois Y (2012a). "On the Difficulty of Relational Concepts among Participants with Down Syndrome." *Research in Developmental Disabilities*, **33**(1), 60–68.

Facon B, Nuchadee ML (2010). "An Item Analysis of Raven's Colored Progressive Matrices among Participants with Down Syndrome." *Research in Developmental Disabilities*, **31**(1), 243–249.

Facon B, Nuchadee ML, Bollengier T (2012b). "A Qualitative Analysis of General Receptive Vocabulary of Adolescents with Down Syndrome." *American Journal on Intellectual and Developmental Disabilities*, **117**(3), 243–259.

Holland PW, Thayer DT (1988). "Differential Item Performance and the Mantel-Haenszel Procedure." In H Wainer, H Braun (eds.), *Test Validity*, pp. 129–145. Lawrence Erlbaum Associates, Hillsdale.

Holland PW, Wainer H (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale.

Jaynes ET (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Kim SH, Cohen AS (1992). "**IRTDIF**: A Computer Program for IRT Differential Item Functioning Analysis." *Applied Psychological Measurement*, **16**(2), 158.

Laurier M, Froio L, Pearo C, Fournier M (1998). "Test de Classement d'Anglais Langue Seconde au Collégial." *Technical report*, College de Maisonneuve, Montreal, QC.

Li HH, Stout W (1994). ***SIBTEST**: A Fortran V Program for Computing the Simultaneous Item Bias DIF Statistics*. Department of Statistics, University of Illinois, Urbana-Champaign.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale.

Magis D, Béland S, Raiche G (2013). ***difR**: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF) in Psychometrics*. R package version 5.0, URL http://www.CRAN.R-project.org/package=difR.

Magis D, Béland 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.

Magis D, De Boeck P (2012). "A Robust Outlier Approach to Prevent Type I Error Inflation in DIF." *Educational and Psychological Measurement*, **72**(2), 291–311.

Magis D, De Boeck P (2014). "Type I Error Inflation in DIF Identification with Mantel-Haenszel: An Explanation and a Solution." *Educational and Psychological Measurement*, **74**(4), 713–728.

Magis D, Facon B (2012). "Angoff's Delta Method Revisited: Improving the DIF Detection Under Small Samples." *British Journal of Mathematical and Statistical Psychology*, **65**(2), 302–321.

Magis D, Facon B (2013a). "Item Purification Does not Always Improve DIF Detection: A Counter-Example with Angoff's Delta Plot." *Educational and Psychological Measurement*, **73**(2), 293–311.

Magis D, Facon B (2013b). ***deltaPlotR**: Identification of Dichotomous Differential Item Functioning (DIF) using Angoff's Delta Plot Method*. R package version 1.3, URL http://www.CRAN.R-project.org/package=deltaPlotR.

Merz WR, Grossen NE (1979). "An Empirical Investigation of Six Methods for Examining Test Item Bias." *Unpublished research report*, California State University, Sacramento, CA.

Michaelides MP (2010). "Sensitivity of Equated Aggregate Scores to the Treatment of Misbehaving Common Items." *Applied Psychological Measurement*, **34**(5), 365–369.

Moon S, McLean JE, Kaufman AS (2003). "A Cross-Cultural Validation of the Sequential Simultaneous Theory of Intelligence in Children." *School Psychology International*, **24**(4), 449–461.

Muniz J, Hambleton R, Xing D (2001). "Small Sample Studies to Detect Flaws in Item Translations." *International Journal of Testing*, **1**(2), 115–135.

Osterlind SJ (1983). *Test Item Bias*. Sage, New York.

Osterlind SJ, Everson HT (2009). *Differential Item Functioning*. 2nd edition. Sage, Thousand Oaks, CA.

Penfield RD (2001). "**DIFAS**: Differential Item Functioning Analysis System." *Applied Psychological Measurement*, **29**(2), 150–151.

Penfield RD, Camilli G (2007). "Differential Item Functioning and Item Bias." In CR Rao, S Sinharray (eds.), *Handbook of Statistics: Psychometrics*, volume 26, pp. 125–167. Elsevier, Amsterdam, The Netherlands.

Raiche G (2002). "Le Dépistage du Sous-Classement aux Tests de Classement en Anglais, Langue Seconde, au Collégial [The Detection of Under Classification to the Collegial English as a Second Language Placement Tests]." *Technical report*, College de l'Outaouais, Gatineau, QC.

Raju NS (1995). ***DFITPU**: A Fortran Program for Calculating DIF/DTF*. Georgia Institute of Technology, Atlanta, GA.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robin F, Sireci SG, Hambleton RK (2003). "Evaluating the Equivalence of Different Language Versions of a Credentialing Exam." *International Journal of Testing*, **3**(1), 1–20.

Rogers HJ, Swaminathan H, Hambleton RK (1993). ***DICHODIF**: A FORTRAN Program for DIF Analysis of Dichotomously Scored Item Response Data*. University of Massachusetts, Amherst, MA.

Rudner LM (1977). "Efforts Toward the Development of Unbiased Selection and Assessment Instruments." Paper presented at the 3rd International Symposium on Educational Testing, Leyden, The Netherlands.

Rudner LM (1978). "Using Standard Tests with the Hearing Impaired: The Problems of Item Bias." *Volta Review*, **80**, 31–40.

Sireci SG, Allalouf A (2003). "Appraising Item Equivalence Across Multiple Languages and Cultures." *Language Testing*, **20**(2), 148–166.

Sireci SG, Patsula L, Hambleton RK (2005). "Statistical Methods for Identifying Flawed Items in the Test Adaptation Process." In RK Hambleton, PF Merenda, CD Spielberger (eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, pp. 93–115. Lawrence Erlbaum Associates, Hillsdale.

Swaminathan H, Rogers HJ (1990). "Detecting Differential Item Functioning using Logistic Regression Procedures." *Journal of Educational Measurement*, **27**(4), 361–370.

Thissen D (2001). ***IRTLRDIF** v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC.

Thissen D, Steinberg L, Wainer H (1988). "Use of Item Response Theory in the Study of Group Difference in Trace Lines." In H Wainer, HH Braun (eds.), *Test Validity*, pp. 147–170. Lawrence Erlbaum Associates, Hillsdale.

Van Herwegen J, Farran E, Annaz D (2011). "Item and Error Analysis on Raven's Colored Progressive Matrices in Williams Syndrome." *Research in Developmental Disabilities*, **32**(1), 93–99.

Wu ML, Adams RJ, Wilson MR (1997). ***ConQuest**: Multi-Aspect Test Software*. Australian Council for Educational Research, Camberwell, Australia.

Zimowski MF, Muraki E, Mislevy RJ, Bock RD (1996). ***BILOG-MG**: Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Scientific Software International, Chicago.

Affiliation:

David Magis
Department of Education (B32)
University of Liège
Boulevard du Rectorat 5, B-4000 Liège, Belgium
E-mail: david.magis@ulg.ac.be
and
Research Group of Quantitative Psychology
KU Leuven
Tiensestraat 102, B-3000 Leuven, Belgium
E-mail: david.magis@ppw.kuleuven.be

Bruno Facon
Univ Lille Nord de France
UDL3, URECA
F-59653 Villeneuve d'Ascq, France
E-mail: bruno.facon@univ-lille3.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 59, Code Snippet 1 *Submitted:* 2012-12-10
July 2014 *Accepted:* 2014-04-15