British Journal of Mathematical and Statistical Psychology (2011)
© 2011 The British Psychological Society



The British Psychological Society

www.wileyonlinelibrary.com

Angoff's delta method revisited: Improving DIF detection under small samples

David Magis[1,*] and Bruno Facon[2]

[1]University of Liège and KU Leuven, Belgium
[2]Université Lille Nord de France, France

Most methods for detecting differential item functioning (DIF) are suitable when the sample sizes are sufficiently large to validate the null statistical distributions. There is no guarantee, however, that they will still perform adequately when there are few respondents in the focal group or in both the reference and the focal group. Angoff's delta plot is a potentially useful alternative for small-sample DIF investigation, but it suffers from an improper DIF flagging criterion. The purpose of this paper is to improve this classification rule under mild statistical assumptions. This improvement yields a modified delta plot with an adjusted DIF flagging criterion for small samples. A simulation study was conducted to compare the modified delta plot with both the classical delta plot approach and the Mantel–Haenszel method. It is concluded that the modified delta plot is consistently less conservative and more powerful than the usual delta plot, and is also less conservative and more powerful than the Mantel–Haenszel method as long as at least one group of respondents is small.

1. Introduction

This paper focuses on differential item functioning (DIF) among dichotomously scored items. An item is said to function differently (or to exhibit DIF) if respondents from different groups, but with equal proficiency levels, have different probabilities of answering this item correctly. In the past thirty years many methods have been introduced to identify DIF items, for their presence in any performance or assessment test can seriously affect the validity of conclusions drawn from that test (Ackerman, 1992).

The best-known DIF methods are based either on statistical methods and classical test theory concepts or on item response theory (IRT) models. Among the former, the best known are the Mantel–Haenszel (Holland & Thayer, 1988), standardization (Dorans & Kulick, 1986), SIBTEST (Shealy & Stout, 1993) and logistic regression (Swaminathan & Rogers, 1990) approaches. The best known IRT-based methods are

*Correspondence should be addressed to David Magis, Department of Mathematics (B37), University of Liège, Grande Traverse 12, B-4000 Liège, Belgium (e-mail: david.magis@ulg.ac.be).

DOI:10.1111/j.2044-8317.2011.02025.x

Lord's χ^2 test (Lord, 1980), Raju's standardized area approach (Raju, 1988, 1990) and the likelihood-ratio test (Thissen, Steinberg, & Wainer, 1988). All these have become popular primarily because they are suitable for DIF detection, they improve on the earlier methods and they are covered in standard software (especially the IRT-based methods).

All these methods require sufficiently large samples of respondents in both reference group and focal group for their conclusions to be statistically valid. They are built on asymptotic properties of their DIF statistics, and IRT-based methods require accurate fitting of IRT models in addition. When sample sizes are small, the validity of the asymptotic distributions of the DIF statistic is doubtful and the conclusions can be seriously affected. Very little is known about the accuracy of standard DIF detection methods with small samples of respondents. Parshall and Miller (1995) compared the exact and asymptotic versions of the Mantel–Haenszel test to detect DIF under small-sample conditions. This appears to be the only work done in this research area. The main reason is probably that it is most common to have large samples available, for instance in large-scale assessment studies.

However, it is sometimes impossible to collect such large amounts of information. For example, in the clinical psychology and developmental disabilities fields, it is generally very difficult, or even impossible, to obtain samples exceeding 50–100 participants. Since DIF investigation has a definite place in these fields, it is crucial to have reliable methods of item analysis usable with small samples.

One potentially suitable DIF detection method with small samples is the so-called *delta plot method* (Angoff & Ford, 1973). This is a graphical approach, easy to compute and interpret (Angoff, 1982; Camilli & Shepard, 1994; Oosterhof, Atash, & Lassiter, 1984; Osterlind, 1983). However, like many early DIF methods, the delta plot was criticized as potentially leading to misidentification of DIF items. Shepard, Camilli, and Williams (1985) even concluded that the method should simply be retired. Together with the rise of more appropriate methods, this probably explains why the delta plot has disappeared from textbooks on DIF, such as Osterlind and Everson (2009) and Penfield and Camilli (2007). But surprisingly, and in spite of its theoretical drawbacks, the delta plot is still in use across a broad range of DIF topics, such as equating aggregate scores (Karkee & Choi, 2005; Michaelides, 2010), language testing and item translation (Muniz, Hambleton, & Xing, 2001; Robin, Sireci, & Hambleton, 2003; Sireci & Allalouf, 2003), mathematics testing (Abedalaziz, 2010), test adaptation processes (Moon, McLean, & Kaufman, 2003; Sireci, Patsula, & Hambleton, 2005) and intelligence testing in special populations (Facon & Nuchadee, 2010; Van Herwegen, Farran, & Annaz, 2011).

The delta plot does not require advanced computer software, it is conceptually simple, and was highlighted as potentially useful when sample sizes are not large (Muniz *et al*., 2001). For these reasons, the delta plot appears to be a worthy method for DIF studies with small samples of respondents. However, its major drawback is that there is no way yet to select an appropriate DIF flagging criterion. The main purpose of this paper is to propose a method to compute such a DIF flagging criterion. This approach relies on a statistical assumption of normality about the data and does not modify the delta plot itself. As a secondary objective, this paper aims to show that the delta plot with accurate DIF flagging criterion, which is referred to as the *modified delta plot*, is a worthy alternative to standard DIF methods with small samples of respondents, and that it actually outperforms the classical delta method overall.

2. Angoff's delta plot

Also known as the *transformed item difficulties* method, the delta plot compares for each item the proportion of correct responses in each group (Angoff & Ford, 1973). These proportion-correct values are sometimes referred to as the p values (Angoff, 1982). Denote by p_{jg} the proportion-correct for item j in group g, where $g = 0$ for the reference group and $g = 1$ for the focal group.

The proportions are transformed into delta scores in two steps. In the first step, the p_{jg} are transformed into so-called *normalized z scores* z_{jg} (or simply *z scores*), where z_{jg} is the deviate of the standard normal distribution with lower-tail probability $1 - p_{jg}$. For example, the proportion-correct values 0.95 and 0.975 correspond to z scores -1.64 and -1.96, respectively. In this way, easier items get lower z scores than more difficult items. In the second step, for the sake of convenience, all z scores are rendered positive by transforming the z scores into delta (Δ) scores by the linear relationship $\Delta_{jg} = 4z_{jg} + 13$. This transformation results in the delta scores being on a positive scale, ranging from 0 to 26, and with average value 13 and standard deviation 4.

The pairs of delta scores $(\Delta_{j0}, \Delta_{j1})$ can be graphically displayed on a scatter plot, with the reference group on the horizontal axis and the focal group on the vertical axis. This is referred to as a *delta plot*, and the pairs of delta values are called *delta points*. In the absence of DIF, such a delta plot will resemble a narrow ellipse along a major or principal axis, and the correlation between the paired delta scores will be large (0.98 or greater). On the other hand, if there are DIF items the delta plot will clearly depart from this narrow ellipse and thus the DIF items will be easily visible. Flagging items as DIF or non-DIF is done by first computing the perpendicular distance D_j of each item from the major axis of the ellipse. This distance is calculated as follows.

The equation of the major axis of the ellipse is $\Delta_{j1} = a + b\Delta_{j0}$, and the intercept and slope parameters are given by

$$b = \frac{s_1^2 - s_0^2 + \sqrt{(s_1^2 - s_0^2) + 4s_{01}^2}}{2s_{01}} \quad \text{and} \quad a = \bar{x}_1 - b\bar{x}_0, \tag{1}$$

where $\bar{x}_g$, s_g^2 and s_{01} are respectively the sample mean, sample variance and sample covariance of the delta scores Δ_{jg}. If the distributions of proficiency levels are similar in both groups of respondents, the major axis will be very close to the identity line, with zero intercept and unit slope. In case of item impact (i.e., one group of respondents has higher average proficiency than the other), the major axis is parallel to the identity line, but shifted upwards or downwards. Finally, if the dispersion of proficiency levels differs between the two groups, the major axis is rotated with respect to the identity line. Thus, the major axis automatically adjusts for differences in distributions between the two groups.

The perpendicular distance D_j between the delta point $(\Delta_{j0}, \Delta_{j1})$ and the major axis is given by

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

Large perpendicular distances indicate large departures of the delta points from the major axis of the ellipse, which signals the presence of DIF. Items with large positive distances are located above the major axis and are consequently easier for respondents

in the reference group. Conversely, large negative perpendicular distances refer to items which are easier for respondents in the focal group. Finally, very small distances suggest that the item difficulties are similar across groups, which indicates the absence of DIF.

2.1. Strengths

Angoff's delta plot has several assets. It is easy to explain and quite intuitive. The proportion-correct values act as proxy values for the item difficulty estimates. Moreover, it does not require large samples to work and tends to remain quite stable with small samples (Muniz *et al*., 2001). In addition, the delta plot takes the potential differences in group proficiency distributions into account in the computation of the perpendicular distances D_j. As pointed out earlier, the major axis of the ellipse can be close to, shifted from, or rotated around the identity line. These three situations correspond respectively to the cases of equal distributions, unequal average proficiency and unequal dispersions of proficiency. The major axis automatically adjusts for differences in distributions, so that the perpendicular distances are not affected by any such differences between the groups of respondents. Finally, the computation of the delta values, major axis and perpendicular distances is very straightforward and does not require intensive computer implementation. Any standard calculator software is adequate.

2.2. Weaknesses

Since the delta plot method is basically built on the comparison of proportion-correct values, it is designed to identify between-group differences in item difficulties ('uniform' DIF). If items exhibit group differences in item discrimination ('non-uniform' or 'crossing' DIF), then they will probably not be detected by this method. Note that this is also the case with other traditional DIF methods, such as the Mantel–Haenszel method (Holland & Thayer, 1988).

A mathematical issue can occur with the delta plot method when the proportions p_{jg} are exactly equal to 0 (all responses are incorrect) or 1 (all responses are correct). The delta scores then become infinite. Angoff and Ford (1973) recommended constraining the proportions to the range [0.05, 0.95]. In our study, however, we considered the broader range [0.001, 0.999]. The delta scores may then take values between 0.63 and 25.36.

Another drawback of the method, which was pointed out by Cole (1978), Hunter (1975), Lord (1977) and Rudner (1978) and confirmed by Angoff (1993), occurs when the groups of respondents have different average proficiency levels (i.e., in the presence of item impact), and when the item discriminations are not identical across the items (i.e., when some items are more discriminating than others). In this situation, non-DIF items may be incorrectly flagged as DIF, especially those items with large discriminating power. Also, low-discriminating DIF items may go undetected.

Finally, the most important drawback of the delta plot method is the selection of an appropriate classification threshold for the perpendicular distances. It was pointed out above that the classification of items as DIF or non-DIF depends on the size of the perpendicular distances. Deciding how large the threshold distance should be is a difficult task, and several approaches have been suggested (e.g., Merz & Grossen, 1979; Michaelides, 2010; Osterlind, 1983; Rudner, 1977, 1978; Strassberg-Rosenberg & Donlon, 1975). Surprisingly, Angoff himself never advocated any particular threshold value or any method for deciding on one. The most popular threshold value at present

seems to be 1.5 (Facon & Nuchadee, 2010; Muniz *et al*., 2001; Robin *et al*., 2003). All the proposed classification rules suggest, in some way or another, a fixed quantity which does not make reference to a pre-specified significance level. Rather, they tend to act as measures of effect sizes. In particular, the value 1.5 is clearly related to the so-called delta scale, which fixes the thresholds for distinguishing between negligible, moderate, and large DIF effect sizes (Dorans, Schmitt, & Bleistein, 1992; Holland & Thayer, 1985; Jodoin & Gierl, 2001).

Unfortunately, such fixed-threshold rules are potentially a serious source of conservativeness (low Type I errors). Indeed, with large samples or large tests, one may expect all the delta points to lie very close to the major axis, and should regard relatively small departures from that axis as indications of DIF. However, the current 'fixed value' thresholds do not share this view, and therefore it is likely that even the items with large DIF will go undetected.

3. The modified delta plot

We introduce a modification of the delta plot, in an attempt to improve the classification rule by accounting simultaneously for the data and for the pre-specified significance level.

We begin by assuming that, under the null hypothesis (absence of DIF), the delta points $(\Delta_{j0}, \Delta_{j1})$ arise from a bivariate normal distribution,

$$(\Delta_{j0}, \Delta_{j1}) : BVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = (\mu_0, \mu_1)^T \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}, \tag{3}$$

where μ_0 and μ_1 are respectively the means of Δ_{j0} and Δ_{j1}, σ_0^2 and σ_1^2 their respective variances, and σ_{01} is the covariance of the pairs $(\Delta_{j0}, \Delta_{j1})$. The superscript T stands for the transposition function.

This is a somewhat strong assumption, but it actually relies on the delta plot characteristics. Recall that in the absence of DIF, the delta plot resembles a narrow ellipse and the groups' delta scores are highly correlated. The elliptical shape is a major characteristic of the bivariate normal distribution: the contours of the bivariate normal density are ellipsoids centred at $\boldsymbol{\mu}$ and with axes proportional to the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ (Johnson & Wichern, 1998, p. 162). It is easy to show that the parameters (1) of the major axis are obtained under these assumptions. In sum, the bivariate normal distribution of the delta points is a reasonable working assumption in this framework. In practice, however, it is difficult to assess the bivariate normality assumption with small tests (few items), so the impact of deviating from this hypothesis must be further investigated.

One can derive the statistical distribution of the perpendicular distances D_j under the null hypothesis of absence of DIF. First, any linear combination of the components of a bivariate normal distribution is itself univariate normally distributed: if $\mathbf{X} \sim BVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\mathbf{C} = (c_1, c_2)^T$ is a bivariate vector, and if ε is a real value, then $\mathbf{C}^T\mathbf{X} + \varepsilon \sim N(\mathbf{C}^{\mathrm{T}}\boldsymbol{\mu} + \varepsilon, \mathbf{C}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{C})$ (Johnson & Wichern, 1998, p. 165). Second, the perpendicular distance D_j can be written as $D_j = \mathbf{C}^T\boldsymbol{\Delta}_j + \varepsilon$, using (2) and with

$$\boldsymbol{\Delta}_j = (\Delta_{j0}, \Delta_{j1})^T, \quad \mathbf{C} = \left(\frac{b}{\sqrt{b^2+1}}, \frac{-1}{\sqrt{b^2+1}} \right)^T, \quad \varepsilon = \frac{a}{\sqrt{b^2+1}}. \tag{4}$$

It turns out that, under the null hypothesis, $D_j \sim N(\mu_D, \sigma_D^2)$, where

$$\mu_D = \frac{b\mu_0 + a - \mu_1}{\sqrt{b^2+1}} \quad \text{and} \quad \sigma_D^2 = \frac{b^2\sigma_0^2 - 2b\sigma_{01} + \sigma_1^2}{b^2+1}. \tag{5}$$

The mean μ_D and variance σ_D^2 of D_j are obtained by combining (3) and (4). Thus, the standardized perpendicular distance

$$\delta_j = \frac{D_j - \mu_D}{\sigma_D} \tag{6}$$

has a standard normal distribution. For any fixed value of the significance level α, the perpendicular distance D_j will be judged as abnormally large or abnormally small if $|\delta_j| > Q_z(1-\alpha/2)$, where Q_z stands for the standard normal deviate. Equivalently, item j will be flagged as DIF if

$$|D_j| > \mu_D + Q_z\Big(1-\frac{\alpha}{2}\Big)\sigma_D = \frac{b\mu_0 + a - \mu_1}{\sqrt{b^2+1}} + Q_z\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{b^2\sigma_0^2 - 2b\sigma_{01} + \sigma_1^2}{b^2+1}}. \tag{7}$$

The mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown and must be estimated from the data. Let

$$\bar{\mathbf{x}} = (\bar{x}_0,\ \bar{x}_1)^T \text{ and } \mathbf{S} = \begin{pmatrix} s_0^2 & s_{01} \\ s_{01} & s_1^2 \end{pmatrix} \tag{8}$$

be the usual sample estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The modified delta plot can then be sketched as follows:

(a) Compute the perpendicular distances D_j as with the classical Angoff delta plot.
(b) Compute the modified classification threshold

$$T_\alpha = \frac{b\bar{x}_0 + a - \bar{x}_1}{\sqrt{b^2+1}} + Q_z\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{b^2 s_0^2 - 2b s_{01} + s_1^2}{b^2+1}}, \tag{9}$$

where a and b are the intercept and slope parameters of the major axis and α is the significance level; or simply

$$T_\alpha = Q_z\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{b^2 s_0^2 - 2b s_{01} + s_1^2}{b^2+1}}, \tag{10}$$

by definition of intercept parameter a given in (1);
(c) Flag item j as DIF if $|D_j| > T_\alpha$ and as non-DIF otherwise.

The modified delta plot retains the essence of the original proposition by Angoff and Ford (1973) while adjusting the classification threshold by taking into account: the location of the ellipse (through b), its shape (through s_0^2, s_{01} and s_1^2) and the significance level α. Smaller significance levels will increase the threshold T_α and hence reduce the

mistaken flagging of non-DIF items (Type I errors). Moreover, as the correlation between the paired delta scores decreases, the covariance s_{01} becomes increasingly smaller than the variances s_0^2 and s_1^2, and consequently the threshold T_α increases. This is consistent with the fact that with lower correlations, the delta plot produces a wider ellipse. Recall that DIF will occur for items that exhibit a large perpendicular distance relative to the other items. If the delta points are all close to the major axis, the correlation between the delta scores will be high, and relatively small perpendicular distances will be enough to identify DIF items. Similarly, when the delta points are more dispersed around the major axis, the delta score correlation will be smaller, and larger perpendicular distances will then be necessary to separate DIF from non-DIF items. The direct consequence is an increase in the T_α value.

4. Simulation study

The modified delta plot should ideally be less conservative than the original method, yet the Type I error rate should remain under control. Moreover, the loss of conservativeness should not happen at the cost of decreased power. In addition, it is worth comparing the modified delta plot with some standard approaches such as the Mantel–Haenszel method. The main emphasis is put on small samples of respondents, though larger samples are also considered for a complete overview of the methods' efficiencies.

4.1. Design

When no DIF items were generated, the design factors were: (a) the size of the reference and focal groups; (b) the difference in average proficiency level across groups; (c) the length of the test (the number of items); and (d) the distributional assumptions of the item difficulties. When DIF items incorporated into the tests, two additional factors were manipulated: (e) DIF size and (f) the percentage of DIF items.

As explained above, the design of the simulation study should include item impact and different discrimination levels across the items. Modelling uniform DIF is achieved by differences in item difficulties between the groups of respondents, but the discrimination level should be identical in both groups of respondents. To satisfy these requirements, a closed form of the two-parameter logistic (2PL) model was used:

$$P_{ijg} = P\left(X_{ijg} = 1|\theta_{ig}, \alpha_j, \beta_{jg}\right) = \frac{\exp\left[\alpha_j\left(\theta_{ig} - \beta_{jg}\right)\right]}{1 + \exp\left[\alpha_j\left(\theta_{ig} - \beta_{jg}\right)\right]}, \tag{11}$$

where X_{ijg} is the response to item j by respondent i in group g (0 for the reference group and 1 for the focal group), θ_{ig} is the proficiency level of respondent i in group g, α_j is the discrimination level of item j, and β_{jg} is the item difficulty of item j in group g. This model allows different levels of discrimination across the items, each item having, however, the same discriminative power in both groups of respondents. Only differences in item difficulties (i.e., uniform DIF) can be considered. With this model, it is expected that the classical delta plot will be inefficient in detecting DIF items, and especially in the presence of item impact (as discussed in section 2.2).

The item difficulty parameters β_{j0} in the reference group were drawn from either a $N(0,1)$ distribution or a uniform $U([-2, 2])$ distribution, and the item discrimination parameters α_j arose from a log-normal distribution with mean 0 and variance 0.1225.

These distributions were chosen to be consistent with previous DIF studies (e.g. Donoghue & Allen, 1993; Zwick, Donoghue, & Grima, 1993; Penfield, 2003). The uniform distribution for item difficulties is rather unusual but was chosen to study the robustness of the modified delta plot against violations of the assumption of bivariate normality. Indeed, the bivariate normality assumption also makes the marginal distributions of the delta scores in each group normal. By generating item difficulties from a uniform distribution, the corresponding delta scores are no longer normally distributed and thus the bivariate normality assumption is violated.

Moreover, if item j was not DIF, then its difficulty was equal across groups: $\beta_{j0} = \beta_{j1}$. Otherwise, the item difficulty in the focal group β_{j1} was obtained by adding the signed value of the DIF size δ_j to the item difficulty of the reference group: $\beta_{j1} = \beta_{j0} + \delta_j$. Two DIF sizes were considered: $\delta_j = 0.4$ and $\delta_j = 0.8$. The value 0.4 is frequently considered in simulation studies (Clauser, Mazor, & Hambleton, 1993; Donoghue, Holland, & Thayer, 1993). However, the larger value 0.8 was also considered in order to quantify the increase in power when the DIF size is larger than usual.

The number of test items took three possible: values 20 (short tests), 40 (moderate tests) and 60 (large tests). When DIF was present, the percentage of DIF items in the test was set to either 5%, 10% or 20%. The latter value is rather large and should reveal the presence of Type I error inflation.

The sizes of the focal and reference groups were set to 50, 100 or 500 units, but the focal group never had more respondents than the reference group. This yielded six situations, three with equal group sizes 50 (50), 100 (100) and 500 (500), and three with different group sizes 100 (50), 500 (50) and 500 (100). This arrangement was meant to reflect the common practical situation in with the focal group is smaller than the reference group.

In the reference group, the proficiency levels of the respondents θ_{i0} were drawn from the $N(0,1)$ distribution. In the focal group, the distribution of proficiencies θ_{i1} was the $N(\gamma, 1)$ distribution. The parameter γ coded for the difference in average proficiency. Two levels of γ were chosen: either $\gamma = 0$, in which case the groups' average levels were equal, or $\gamma = -1$, such that the focal group has lower average proficiency than the reference group. The latter case permitted introduction of item impact into the generated data sets (see Penfield, 2001).

4.2. DIF methods

Three DIF detection methods were considered. The first was the classical delta plot, with classification threshold of 1.5. This value is not the only one to have been considered in the literature, but today it is quite commonly used. In addition, this large threshold would obviously imply significant conservativeness, which was expected to result from the simulations.

The second method was the modified delta plot described above, with the classification threshold T_α given by (9).

The third method was the Mantel–Haenszel DIF method. This was chosen for three reasons: (a) like the delta plot, it is a non-IRT method; (b) also like the delta plot, it focuses primarily on uniform DIF; and (c) it is known to be one of the best-performing non-IRT DIF methods (Wainer, 2010) when the samples of respondents are sufficiently large. Very little is known about the performance of this method with small samples, apart from its general properties of asymptotic validity with categorical data (Agresti, 1996). In this study, the Mantel–Haenszel χ^2 statistic was computed for each item, and items

with statistically significant χ^2 values were flagged as DIF. The continuity correction (Holland & Thayer, 1988) was applied to each statistic.

Item purification (Candell & Drasgow, 1988) was not performed and one may therefore expect some Type I error inflation in the presence of large percentages of DIF items (Clauser *et al*., 1993). This process was apparently never applied to the delta plot, although it is technically feasible: at each iteration, items flagged as DIF should be removed from the computation of the means, variances and covariance of the delta score. In this way, the intercept and slope parameters of the major would iteratively become unaffected by items flagged as DIF. This process is worthy of further study but is outside the scope of the present paper. To maintain comparability of results, item purification was applied to none of the methods.

4.3. Data generation

In the absence of DIF, this design yielded 72 different settings (six pairs of sample sizes, three test lengths, two average proficiency levels in the focal group and two distributions of item difficulties). When DIF was added to the data, 432 different settings were considered (three percentages of DIF and two DIF sizes in addition to the previous factors). For each of these different situations, 1,000 data sets were randomly generated. Binary responses were drawn as follows: the response X_{ijg} of respondent i from group g to item j was drawn from a Bernoulli distribution with success probability P_{ijg} given by (11). The three methods for DIF investigation were applied to each data set. The significance level was fixed to 5%.

4.4. Type I error and power

The Type I error rates were computed as the empirical proportion of items incorrectly flagged as DIF. Power was computed as the empirical proportion of DIF items correctly flagged as DIF. Power was not computed, of course, when DIF was absent from the data.

In order to identify which factors actually did affect the Type I error rates and power values, the following approach was adopted. First, the Type I error (separately in the presence and absence of DIF items) and the power were transformed onto a more continuous scale by the logit transformation. Second, an analysis of variance was carried out, with either the Type I error rate or the power as the dependent variable, and the design factors as covariates. The analysis of variance was carried out with all possible interaction terms between all covariates. This saturated model was reduced by removing all non-significant interaction terms and all factors which explained less than 1% of the variability of the dependent variable. With this approach, one expects to identify the most important factors and to remove the non-contributory and non-significant factors.

5. Results

Before analysing the results of the simulation study, it is worth mentioning that at most four factors are retained: the sizes of the groups of respondents, the DIF methods, and the DIF size and the percentage of DIF items (when DIF is present). The number of items, the difference in average proficiency levels, and the distributions of item difficulty levels either were not statistically significant or did not explain sufficient variability. These factors are therefore disregarded in the following discussion. One important conclusion

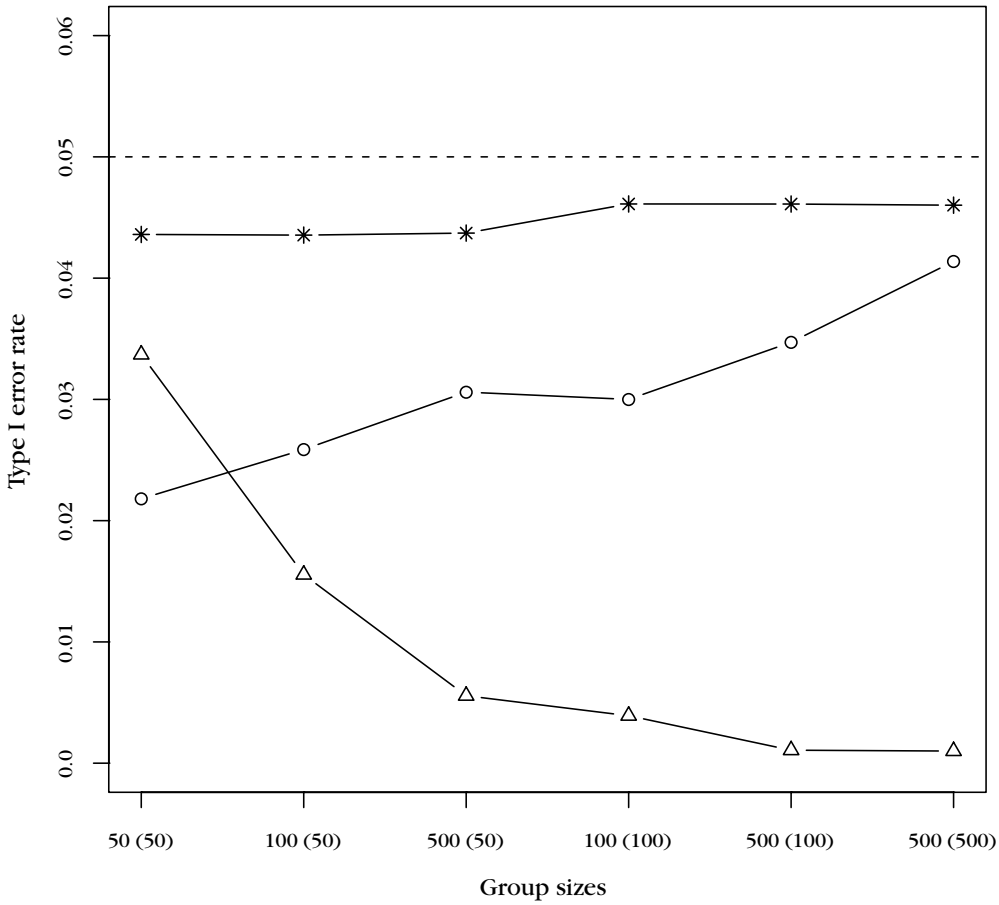

Figure 1. Type I error rates in the absence of DIF, for various sizes of reference and focal groups. The focal group size is displayed in parentheses. The three methods are: the classical delta plot method (triangles), the modified delta plot method (stars) and the Mantel–Haenszel method (circles).

at this point is that the modified delta plot is not affected by data sets that are clearly violating the assumption of bivariate normality of the delta scores.

5.1. Type I error without DIF

The Type I error rate in the absence of DIF was best explained by the DIF method, the group sizes and the first-order interaction between them. This model had an R^2 of 0.981, while the R^2 of the fully saturated model was 0.999. The Type I error rates are displayed in Figure 1. They are ordered by group sizes, first by ascending size of the focal group, then by ascending size of the reference group. The classical delta plot, the modified delta plot and the Mantel–Haenszel method are shown as triangles, stars and circles, respectively.

The modified delta plot performed better than the other two methods regardless of the sample sizes. Indeed, the Type I error rates were nearly constant across group sizes,

and all closer to the nominal significance level than those of the other two methods. On the other hand, as anticipated, the classical delta plot was highly conservative. The Type I error rate was about 0.035 with small samples, but it quickly decreased as the size of the reference group and/or the focal group increased. The Mantel–Haenszel method displayed the opposite trend: the Type I error rate was very small with small samples (even smaller than with the classical delta plot), but increased quickly as sample size increased. With large samples, the Type I error rates were nearly equal for the Mantel–Haenszel and modified delta plot methods. Note, however, that the Mantel–Haenszel method returned consistently smaller Type I error rates than the modified delta plot. In addition, with samples of at least 100 respondents in each group, the Mantel–Haenszel Type I error rates were similar to those reported in previous studies (e.g. Penfield, 2001).

5.2. Type I error in the presence of DIF

When DIF is present, the Type I error rate is best explained with four factors: DIF method, sizes of the groups, DIF size and percentage of DIF. Moreover, all first-order (pairwise) interactions were significant. The Type I errors rates evolved differently across DIF methods, depending either on the group sizes, the size of the DIF effect or the percentage of DIF items. This model yielded an R^2 of 0.956, compared to 0.998 for the saturated model. The Type I errors are displayed in Figure 2. The left-hand panels show Type I error rates when the DIF size is 0.4, while the right-hand panels refer to DIF size 0.8. The top, middle and bottom panels correspond respectively to 5%, 10% and 20% of DIF items.

Independently of the DIF size, the relationship between the Type I error rates and group sizes differs for each method. With the classical delta plot and the Mantel–Haenszel method, one sees trends similar to those for the absence of DIF: the Type I error quickly decreases with group sizes for the classical delta plot, while it increases for the Mantel–Haenszel method. Note that in the case of DIF 0.8, 10% (or more) of DIF items and 500 respondents in each group, the Mantel–Haenszel Type I error rate exceeded the nominal significance level, up to 0.15 when 20% of items are DIF. This extreme situation is a clear indicator that Type I error rate inflation can occur with the Mantel–Haenszel method (recall that we did not carry out item purification).

The picture is somewhat different with the modified delta plot. The Type I error rate remains relatively constant across group sizes, except with large groups, where it decreases somewhat dramatically. Note also that the rates were consistently lower in this context than when DIF was absent.

Moreover, the Type I error rate for the modified delta plot decreased when DIF size was increasing, while remaining quite constant for the classical delta plot. With the Mantel–Haenszel method, Type I error rates are somewhat unaffected by an increase in DIF size, except when the samples are large since a clear Type I error inflation is observed. It was also possible to notice these trends when the percentage of DIF items increased.

5.3. Power

The power values can be discussed with exactly the same model as that for the Type I error rates when DIF is present. This model returned an R^2 of 0.963, while the saturated model yielded an R^2 of 0.998. The power values are displayed in Figure 3, in the same way as the Type I errors in Figure 2.

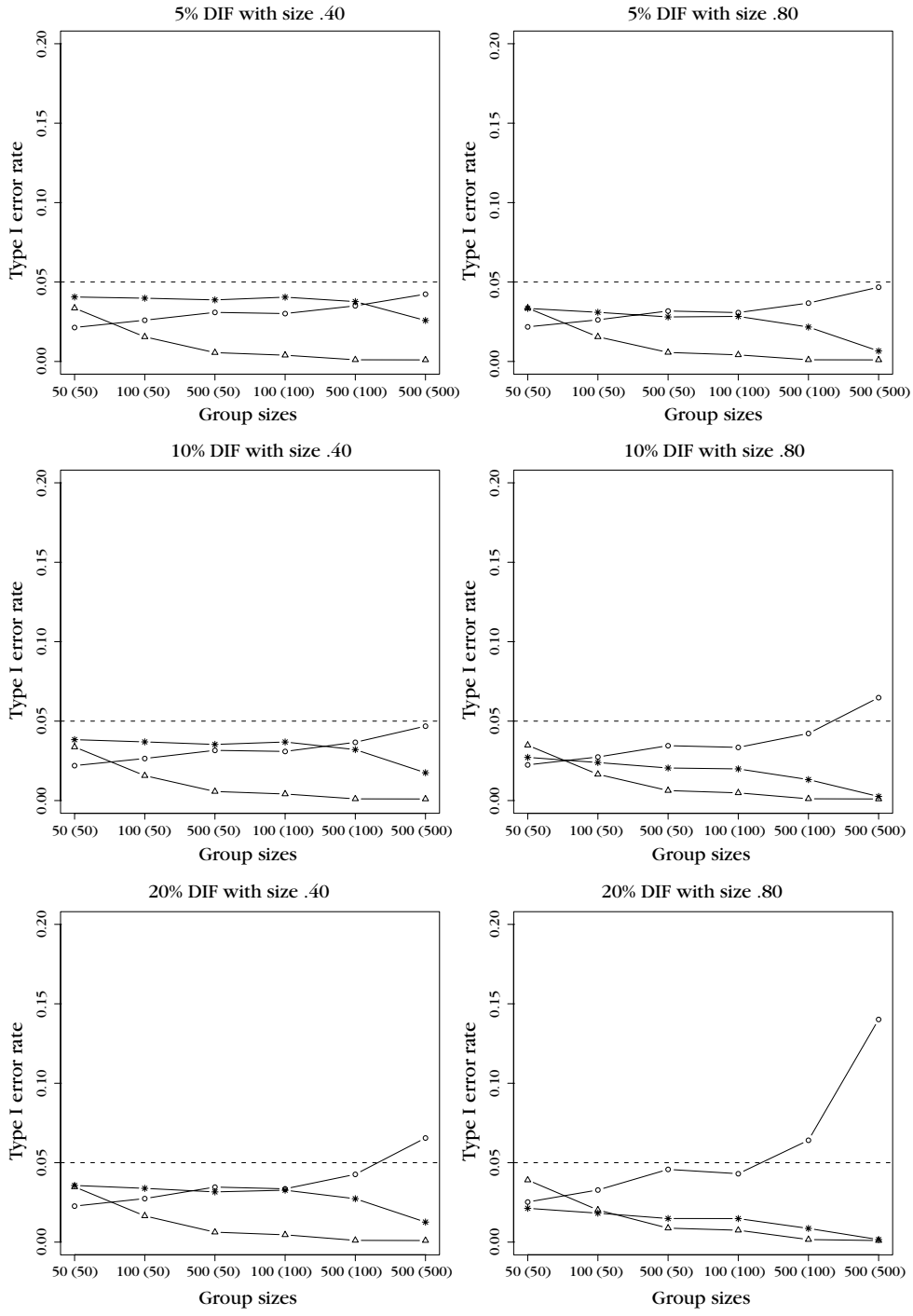

Figure 2. Type I error rates in the presence of DIF, for various sizes of reference and focal groups, with DIF size 0.4 (left) and 0.8 (right) and DIF percentage 5% (top), 10% (middle) and 20% (bottom). The focal group size is displayed in parentheses. The methods are the classical delta plot (triangles), the modified delta plot (stars) and the Mantel–Haenszel method (circles).

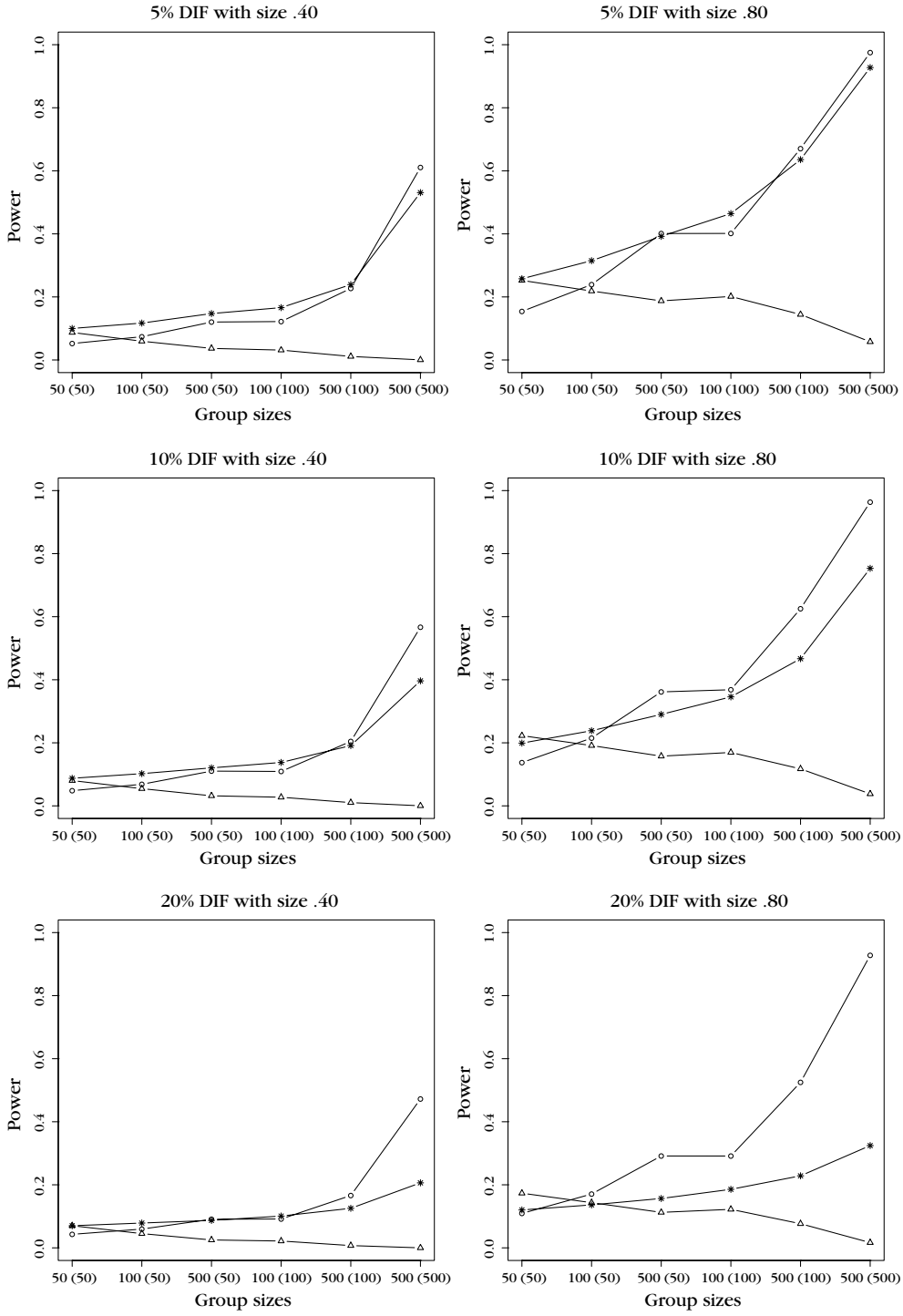

Figure 3. Power values in the presence of DIF, for various sizes of reference and focal groups, with DIF size 0.4 (left) and 0.8 (right) and DIF percentage 5% (top), 10% (middle) and 20% (bottom). The focal group size is displayed in parentheses. The methods are the classical delta plot (triangles), the modified delta plot (stars) and the Mantel–Haenszel method (circles).

As group sizes increased, power also increased for the Mantel–Haenszel method and the modified delta plot. This was expected since larger samples permit better identification of DIF items. However, for the classical delta plot the increase in group sizes did not affect power and with the larger DIF effect it even seemed to decrease with group size. This can be explained by the fact that larger groups usually yield smaller perpendicular distances, and a large classification threshold yields gross under-detection of DIF.

Moreover, as expected, power increased consistently for each method and for each group size as the DIF effect increased. In this case, DIF items departed more clearly from non-DIF items, and their detection was thus easier. Also, power decreased as the percentage of DIF items increased. This is more marked for the modified delta plot than for the other two methods.

Interestingly, power was larger for the modified delta plot than for Mantel–Haenszel as long as the groups were not too large, typically with size 50 or 100. When both groups had 50 respondents, the Mantel–Haenszel method was the less powerful method. With larger groups, however, Mantel–Haenszel outperformed the two delta plot methods in terms of power. This was more marked when the DIF effect was large. Finally, the gap between the modified delta plot and the Mantel–Haenszel method increases when the percentage of DIF items increases. The former gets an important reduction of power, while the latter displays rather stable power values even with large percentages and sizes of DIF.

5.4. Conclusions

The classical delta plot has demonstrated its limitations in identifying uniform DIF, even with small samples. The method is clearly conservative, with Type I error rates grossly underestimating the nominal significance level, and power generally small and unresponsive to sample size. In this regard, the conclusions are similar to those of Shepard *et al*. (1985). However, it is worth mentioning that with very small samples (50 respondents in each group), this approach performs equally well as the Mantel–Haenszel method. This is probably because the latter is designed for large samples.

The modified delta plot, by contrast, constitutes a clear and valuable improvement. It is the least conservative method in the absence of DIF. The Type I error rate is unaffected by test length, difference in average proficiency levels or distributions of item difficulties. The modified delta method is therefore rather robust against the violations of the bivariate normality assumption of the delta scores.

The Mantel–Haenszel method yields largest power values with large sample sizes. However, this might only be an artefact because the Type I error rate also tends to be larger with large samples for this method. From Figure 2 this is particularly the case when the proportion of DIF items is very large. With small groups of respondents, the asymptotic validity of the Mantel–Haenszel chi-squared statistic is clearly doubtful, and this has an impact on both the Type I error rate and the power of the method. With less than 100 respondents in at least one of the two groups, the modified delta method performs best, but the Mantel–Haenszel approach is the best alternative with larger groups of respondents. This confirms the superiority of Mantel–Haenszel when the samples are sufficiently large, but gives a rationale for the modified delta plot under small samples. In sum, the potential violation of the bivariate normality assumption for the modified delta plot, which is impossible to assess with small samples, seems less

problematic than small-sample violation of the asymptotic Mantel–Haenszel chi-squared distribution for DIF detection.

It is also worth mentioning that the presence of item impact and different discrimination levels does not affect the modified delta plot. In the presence of DIF, impact is not a significant factor for explaining Type I error and power; and in the absence of DIF, the Type I error rate of the modified delta plot remains unchanged when impact is present. This is therefore a clear and valuable improvement of the delta method.

In sum, the modified delta plot is a worthy alternative to classical DIF methods when the samples are small. It is designed to automatically adjust the classification threshold to the data. Also, the classical DIF methods are most often asymptotic methods, and most practical studies make use of the standard, large-sample forms of the classical methods. The modified delta plot is certainly an asset in this context.

One possible criticism of the modified delta plot is that, in the presence of DIF, the Type I error rate decreases quickly with increasing group sizes. On the other hand, power increases at the same time, so that with larger groups of respondents, more DIF items are detected as such, while non-DIF items are less and less incorrectly flagged as DIF. The gap between DIF and non-DIF items is more marked with larger groups, and the modified delta plot adjusts conveniently to yield an almost perfect classification of the items. For the Mantel–Haenszel method, in contrast, the increase in power is counterbalanced by an increase in Type I errors. That is, more items are flagged as DIF overall, and among them the number of non-DIF items also increases.

Another possible explanation of this decrease in Type I error is that the presence of DIF items provokes an overestimation of the overall dispersion of the delta points. That is, DIF items act as outliers and have a non-negligible impact on the sample estimates (8) of the mean vector and covariance matrix. It turns out that the classification threshold (7) may be overestimated, yielding both lower Type I error and lower power than expected. This is particularly the case when both the sample sizes and the DIF size are large. Replacing sample estimates by robust (in the statistical sense) alternatives could overcome this issue and return less conservative Type I errors and increased power, as was recently suggested in the DIF framework (Facon, Magis, Nuchadee, & De Boeck, in press; Magis, & De Boeck, in press).

Finally, the classical delta plot, which returns low Type I error rates and low power, behaves exactly as claimed by Rudner (1978) and Angoff (1982) when the item discriminative power differs from one item to the next. That is, the classical delta plot is not suitable in this context and fails to correctly identify DIF items. The modified delta plot does not suffer from this drawback.

6. Discussion

We have proposed a modification of the delta plot in order to improve its classification rule. The modification avoids some known problems of the classical method (conservativeness, low power) and provides a more theory-based platform for the derivation of the classification threshold. The modified method is less conservative and more powerful than its original form. Moreover, it is more powerful than the Mantel–Haenszel method as long as at least one of the groups is small. However, the Mantel–Haenszel method remains most appropriate for larger groups, yielding comparable Type I error rates and more power than the modified delta plot. It is, however, noticeable that with large

proportions of DIF items, the Mantel–Haenszel method suffers from Type I error inflation, which also might artificially lead to an increased power.

The modified delta plot considered DIF items as items either easier or more difficult for the focal group, without any preference. In this regard, the absolute value of the perpendicular distance was taken into account in the classification rule. If however, the practitioner has some insight into the trend of the DIF effect – for instance, if the DIF items are more difficult for the focal group – then the classification rule could be written as $\delta_j > Q_z\,(1-\alpha)$ (using (6) and the fact that positive distances D_j refer to easier items in favour of the reference group). This approach would concentrate on asymmetric DIF only and could give even better results than those of the present study (since only asymmetric DIF was generated).

Other distances than the perpendicular distance could also be considered. In line with several proposals (Merz and Grossen, 1979; Rudner, Getson, & Knight, 1980), the residual (i.e., vertical) distance between the delta points and the major axis could be used as an indicator of DIF. The vertical distance is equal to $\sqrt{b^2+1}$ times the perpendicular distance, and it is straightforward to derive the corresponding classification threshold T_α using the vertical distances instead of the perpendicular ones.

Another technical problem of interest is the investigation of items whose proportions of correct responses are nearly zero or one. Arbitrarily constraining the range of such proportions is the easiest way to overcome the issue, but other approaches could be considered. For instance, the 'add two successes and two failures' method (Agresti & Coull, 1998) has proved to be efficient in estimating proportions nearly equal to zero or one. Maybe this would also be suitable for the delta plot with such items. Also, as pointed out in Section 4.2, the benefits of applying item purification to the delta plot should be further investigated.

Developing or improving statistical methods is one thing, being sure of their utility is quite another. We have argued that delta plot could, with some improvements, be useful in many situations, such as the fields of clinical psychology and developmental disabilities. In our opinion, the potential contributions of this statistical tool are threefold.

First, it can enable us to address the question of equivalency of item functioning for clinical versus typical populations. Data on this subject are very scarce, as if common tacit consent has been given to the conclusion, once and for all and without any proof, that the functioning of items is the same for these clinical groups as for the general population. This implicit position is quite paradoxical in view of the wide use of tests in the field for diagnosis, therapeutic or educational choices, and placement decisions. The same goes for research studies comparing typical and clinical populations because test scores are ubiquitous as independent, control and outcome measures. As a matter of fact, it is quite common to read research articles on dyslexia, autism, cerebral palsy or intellectual disability in which a great variety of psychometric tests are used without any examination of the functioning of the items they contain (Magis, & De Boeck, in press).

Second, the fact that the modified delta plot can detect items that display differential difficulty across groups of small to moderate size may well prove useful for improving the accuracy of matching groups in comparative studies. It can allow us, in particular, to go beyond the simple equivalence of mean raw scores and to assert that groups are well and truly matched on the psychological trait in question. From this standpoint, the modified delta plot can provide evidence that groups are matched *both* on the raw score and on the item scores – or, as may sometimes be the case, provide evidence to the contrary, that the groups are not well matched.

Third, the current modification of Angoff's item analysis approach could be used fruitfully as a post-hoc statistical tool for better examination of the effects of independent variables on dependent variables. Very often these effects are considered only in terms of mean differences. For example, neuropsychological profiles of members of the most studied clinical groups in intellectual disabilities research (e.g., Down's syndrome, Williams syndrome, Fragile X, Prader–Willi) are almost exclusively established by analysing between-group mean differences across behavioural, cognitive or linguistic domains (e.g., Bellugi, Lichtenberger, Jones, & Lai, 2000; Järvinen-Pasley *et al*., 2008; Mervis & Johns, 2010; Schneider, Hagerman, & Hessl, 2009; Silverman, 2007). Why should we not refine the analyses by carrying out post-hoc item response analyses to gather high-definition information about the effects of the variables under study? In this respect, the modified delta plot could yield rich results.

The statistical identification of potentially DIF items is only the first step of a larger process. A second, validation step should be performed next, for instance by using appropriate measures of effect size. The modified delta plot is a clear improvement on the standard delta plot approach, and as such it is worth considering not only with small samples, but also as a straightforward tool for preliminary DIF investigation. The appropriate setting of an effect size measure for the modified delta plot would obviously be an asset and may help in discussing relevance of the item construct.

Acknowledgements

We thank the Editor and two anonymous reviewers for their helpful comments, and Prof. John M. Belmont for suggestions on an earlier draft of this paper. The study was funded by a Research Associate stipend from the National Funds for Scientific Research (FNRS, Belgium), the Research Funds of the KU Leuven, Belgium, and a research grant from the French National Research Agency (Agence Nationale de la Recherche, France) ANR, LANG & HANDICAPS, Projet no. ANR-09-ENFT-019.

References

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *International Journal of Educational and Psychological Assessment*, *5*, 101–116.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

Agresti, A. (1996). *Introduction to categorical data analysis*. New York: Wiley.

Agresti, A., & Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician*, *52*, 119–126.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Beck (Ed.), *Handbook of methods for detecting item bias* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95–106. doi:10.1111/j.1745-3984.1973.tb00787.x

Bellugi, U., Lichtenberger, L., Jones, W., & Lai, Z. (2000). I. The neurocognitive profile of Williams syndrome: A complex pattern of strengths and weaknesses. *Journal of Cognitive Neuroscience*, *12*, 7–29.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253–260. doi:10.1177/014662168801200304

Cole, N. S. (1978, March). *Approaches to examining bias in achievement test items*. Paper presented at the annual meeting of the American Personnel and Guidance Association, Washington, DC.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*, 269–279. doi:10.1207/s15324818ame0604_2

Donoghue, J. R., & Allen, N. L. (1993). Thin vs. thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics*, *18*, 131–154. doi:10.3102/10769986018002131

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte-Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368. doi:10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*, 309–319. doi:10.1111/j.1745-3984.1992.tb00379.x

Facon, B., Magis, D., Nuchadee, M.-L., & De Boeck, P. (in press). Do Raven's Colored Progressive Matrices function in the same way in typical and clinical populations? Insight from the intellectual disability field. *Intelligence*.

Facon, B., & Nuchadee, M.-L. (2010). An item analysis of Raven's Colored Progressive Matrices among participants with Down syndrome. *Research in Developmental Disabilities*, *31*, 243–249. doi:10.1016/j.ridd.2009.09.011

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty*. *Research Report RR-85–43*. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunter, J. E. (1975, December). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items*. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Järvinen-Pasley, A., Bellugi, U., Reilly, J., Mills, D. L., Galaburda, A., Reiss, A. L., & Korenberg, J. R. (2008). Defining the social phenotype in Williams syndrome: A model for linking gene, the brain, and behavior. *Development and Psychopathology*, *20*, 1–35. doi:10.1017/S0954579408000011

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329–349. doi:10.1207/S15324818AME1404_2

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Karkee, T., & Choi, S. (2005, April). *Impact of eliminating anchor items flagged from statistical criteria on test score classifications in common item equating*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In N. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Magis, D., & De Boeck, P. (in press). A robust outlier approach to prevent Type I error inflation in DIF. *Educational and Psychological Measurement*.

Magis, D., & De Boeck, P. (in press). Identification of differential item functioning in multiple-groups settings: A multivariate outlier detection approach. *Multivariate Behavioral Research*.

Mervis, C. B., & John, A. E. (2010). Cognitive and behavioral characteristics of children with Williams syndrome: Implications for intervention approaches. *American Journal of Medical Genetics Part C*, *154*, 229–248. doi:10.1002/ajmg.c.30263

Merz, W. R., & Grossen, N. E. (1979). *An empirical investigation of six methods for examining test item bias*. Unpublished research report. Sacramento: California State University.

Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Applied Psychological Measurement*, *34*, 365–369. doi:10.1177/0146621609359626

Moon, S., McLean, J. E., & Kaufman, A. S. (2003). A cross-cultural validation of the sequential simultaneous theory of intelligence in children. *School Psychology International*, 24, 449–461. doi:10.1177/01430343030244007

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*, 115–135. doi:10.1207/S15327574IJT0102_2

Oosterhof, A. C., Atash, M. N., & Lassiter, K. L. (1984). Facilitating identification of item bias through use of delta plots. *Educational and Psychological Measurement*, *44*, 619–627. doi:10.1177/0013164484443009

Osterlind, S. J. (1983). *Test item bias*. Newbury Park, CA: Sage.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement*, *32*, 302–316. doi:10.1111/j.1745-3984.1995.tb00469.x

Penfield, R.D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*, 235–259. doi:10.1207/S15324818AME1403_3

Penfield, R.D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research*, *49*, 231–243.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharray (Eds.), *Handbook of Statistics 26: Psychometrics* (pp. 125–167). Amsterdam: Elsevier.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197–207. doi:10.1177/014662169001400208

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*, 1–20. doi:10.1207/S15327574IJT0301_1

Rudner, L. M. (1977, June). *Efforts toward the development of unbiased selection and assessment instruments*. Paper presented at the third International Symposium on Educational Testing, Leyden, The Netherlands.

Rudner, L. M. (1978). Using standard tests with the hearing impaired: The problems of item bias. *Volta Review*, *80*, 31–40.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, *17*, 1–10. doi:10.1111/j.1745-3984.1980.tb00810.x

Schneider, A., Hagerman, R., & Hessl, D. (2009). Fragile X syndrome – From genes to cognition. *Developmental Disabilities Research Reviews*, *15*, 333–342. doi:10.1002/ddrr.80

Shealy, R. T., & Stout, W. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159–194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*, 77–105. doi:10.1111/j.1745-3984.1985.tb01050.x

Silverman, W. (2007). Down syndrome: Cognitive phenotype. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 228–236. doi:10.1002/mrdd.20156

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*, 148–166. doi:10.1191/0265532203lt249o

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Hillsdale, NJ: Lawrence Erlbaum Associates.

Strassberg-Rosenberg, B., & Donlon, T. F. (1975, June). *Context influences on sex differences in performance and aptitude tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–170). Hillsdale, NJ: Lawrence Erlbaum Associates.

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*, 93–99. doi:10.1016/j.ridd.2010.09.005

Wainer, H. (2010). Fourteen conversations about three things. *Journal of Educational and Behavioral Statistics*, *35*, 5–25. doi:10.3102/1076998609355124

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–251. doi:10.1111/j.1745-3984.1993.tb00425.x

Received 17 Jan 2011; revised version received 07 July 2011