Educational and Psychological
Measurement
73(2) 293–311
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164412451903
epm.sagepub.com
SAGE

Item Purification Does Not Always Improve DIF Detection: A Counterexample With Angoff's Delta Plot

David Magis[1,2] and Bruno Facon[3]

Abstract

Item purification is an iterative process that is often advocated as improving the identification of items affected by differential item functioning (DIF). With test-score-based DIF detection methods, item purification iteratively removes the items currently flagged as DIF from the test scores to get purified sets of items, unaffected by DIF. The purpose of this article is to highlight that item purification is not always useful and that a single run of the DIF method may return equally suitable results. Angoff's Delta plot is considered as a counterexample DIF method, with a recent improvement to the derivation of the classification threshold. Several possible item purification processes may be defined with this method, and all of them are compared through a simulation study and a real data set analysis. It appears that none of these purification processes clearly improves the Delta plot performance. A tentative explanation is drawn from the conceptual difference between the modified Delta plot and the other traditional DIF methods.

Keywords

differential item functioning, delta plot, item purification, Type I error, power

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

Corresponding Author:
David Magis, Department of Mathematics (B37), University of Liège, Grande Traverse 12, B-4000 Liège, Belgium.
Email: david.magis@ulg.ac.be

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

Introduction

The identification of differential item functioning (DIF) has received increasing attention in recent decades. Early methods were mostly based on test scores, such as Angoff's Delta plot (Angoff, 1972), Mantel–Haenszel (Holland & Thayer, 1988), and standardization (Dorans & Kulick, 1986) methods. Item response theory (IRT)–based methods were also described by Lord (1980); Thissen, Steinberg, and Wainer (1988); and Raju (1990), among others. Together with collecting larger data sets and the advent of powerful computers and software, more sophisticated DIF detection methods have been proposed later, based, for example, on hierarchical modeling (Kamata, 2001; Swanson, Clauser, Case, Nungester, & Featherman, 2002), on mixture models (Cohen & Bolt, 2005; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010), or on generalized linear mixed models (De Boeck, 2008).

Most of these techniques require large or very large data sets to ensure the convergence of optimization algorithms, the accuracy of parameter estimates, or the asymptotic validity of the DIF statistics. Unfortunately, it is not always possible to obtain or collect large data sets by giving the test to a large number of respondents. In the intellectual disabilities research field, for example, the birth prevalence of many syndromes is so low (e.g., 1 per 4,000 births for males with Fragile X syndrome or 1 per 7,500 births for individuals with Williams Syndrome; Lightbody & Reiss, 2009; Mervis & John, 2010) that samples of 50 participants are often considered as rather large (see, e.g., Facon, Magis, & Belmont, 2011; Facon & Nuchadee, 2010). Under this circumstance (i.e., small sample sizes), even the traditional DIF methods, yet the least demanding in terms of sample sizes, may return spurious results because of the lack of a valid statistical asymptotic framework (Fidalgo, Ferreres, & Muniz, 2004; Muniz, Hambleton, & Xing, 2001).

From the standpoint of sample sizes, Angoff's Delta plot method (Angoff, 1972; Angoff & Ford, 1973) is probably the most convenient. Indeed, besides the fact that it is straightforward and easy to compute, it does not necessarily require large samples of respondents. Moreover, the Delta plot has recently been improved (Magis & Facon, 2012) to return more accurate DIF detection thresholds under a statistical framework. This modified Delta plot was shown to be very efficient in terms of Type I error control and power to detect DIF items. It clearly outperforms Angoff's original Delta plot and is even more accurate than the Mantel–Haenszel method when applied to small samples of respondents (Magis & Facon, 2012). Although promising, these results call for further research to optimize the accuracy of the modified Delta plot. For example, the influence of item purification technique on the DIF detection rate is unknown and, thus, remains to be studied.

Item purification is the most common approach to lessen the impact of DIF items on the accuracy of the results of the DIF methods (Candell & Drasgow, 1988). Purification is an iterative application of the DIF method in which all items currently flagged as DIF are removed from the matching criterion. For instance, item purification is applied to test-score-based methods by removing items flagged as DIF from the computation of the test scores used as proxies for ability levels. With IRT-based

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

methods, item purification acts rather when item parameters in both groups of respondents are being rescaled, usually onto the reference group scale. At each step of the purification process, rescaling is made by removing all items currently flagged as DIF. Usually, taking DIF items into account when computing the matching criterion yields an inflation of the Type I error rate, that is, more non-DIF items are incorrectly flagged as DIF (Clauser, Mazor, & Hambleton, 1993). On the contrary, item purification tends to prevent Type I error inflation, returning values much closer to the nominal significance level, but also tends to increase the power to detect true DIF items.

Although the practical usefulness of item purification was demonstrated by several studies (Candell & Drasgow, 1988; Clauser et al., 1993; Fidalgo, Mellenbergh, & Muniz, 2000; Wang & Su, 2004), thus giving support for advocating a systematic application of the process, there is no theoretical argument to ensure that it is always improving the DIF detection results. The purpose of this article is to actually highlight a counterexample against the necessity of item purification. This example arises from Angoff's Delta plot method, and more precisely its recently modified version, further referred to as the *modified* Delta plot.

Two issues must be pointed out at this step. First, the modified Delta plot is most suitable for small samples, as it behaves better or similarly to the Mantel–Haenszel method (Magis & Facon, 2012). While item purification may be applied with any method and any sample size, in practice the following study will be based on small samples only, in order to keep consistency with previous simulation studies and results. Also, not any other DIF method than the modified Delta plot and its purified version will be compared in this article. The second issue comes from the fact that item purification is not uniquely defined with the Delta plot. As will be illustrated later on, three different purification approaches can be applied, leading to three different classification thresholds with various implications on the DIF detection results. So, unlike the traditional DIF methods, it is also necessary to determine which purification process performs best and whether it outperforms the (not purified) Delta plot.

Delta Plot and Modified Delta Plot

This section starts with a brief outline of the original Delta plot and its modified version.

Delta Plot

Consider a test of J items to be administered to two groups of respondents, the *reference group* and the *focal group*. These groups are further referred to with subscripts 0 and 1, respectively. The first step of the Delta plot is to compute the proportions of correct responses (also sometimes called p-values) for each item in each group of respondents. Let p_{j0} and p_{j1} ($j = 1, \ldots, J$) be the proportion of correct responses to item j in the reference group and the focal group, respectively. These proportions are then transformed onto a more continuous scale using standard normal (z) scores:

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

$z_{jt} = \Phi^{-1}(1 - p_{jt})$ ($t = 0, 1$), where Φ is the density function of the standard normal distribution. In this way small and large proportions (e.g., 0.05 and 0.95) are transformed into, respectively, large and small z scores (here 1.64 and -1.64). Finally, z scores are transformed into Delta (Δ) scores by the linear relationship $\Delta_{jt} = 4z_{jt} + 13$ to avoid any negative values.

Each item is characterized by a pair of Delta scores (Δ_{j0}, Δ_{j1}), called the *Delta point*, and all items can be simultaneously displayed on a scatter plot by using the Delta points as axis coordinates. This scatter plot is called the *Delta plot*. Traditionally, the reference group is displayed on the x axis and the focal group on the y axis. In the absence of DIF, the Delta plot looks like an ellipsoid, while DIF items clearly depart from this elliptical shape. This provides the basis for developing the following DIF statistic.

First, the principal (or major) axis of the ellipsoid is computed as

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

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

where $\bar{x}_t$ and s_t^2 ($t = 0, 1$) are, respectively, the mean and the variance of the Delta scores Δ_{jt} in group t, and s_{01} is the covariance between Delta scores. Equation (1) characterizes the major axis from the lower left to upper right corners of the Delta plot. The DIF statistic is then obtained as the perpendicular distance D_j between the Delta point (Δ_{j0}, Δ_{j1}) and the major axis:

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

The distance D_j is positive, equal to zero, or negative, depending on whether the Delta point is located below, on, or above the major axis. A large perpendicular distance (in absolute value) indicates an important departure of the item from the major axis, that is, from the majority of the item set, which is evidence of DIF. Muniz et al. (2001) recommend flagging as DIF any item j whose $|D_j| > 1.5$ (see also Robin, Sireci, & Hambleton, 2003). Other authors have suggested different, yet fixed, threshold values (see, e.g., Michaelides, 2010; Osterlind, 1983; Rudner, 1977, 1978).

Modified Delta Plot

The main drawback of the original Delta plot is its use of a fixed threshold that is chosen more or less arbitrarily. Because the perpendicular distances are computed relative to the locations of all Delta points together (through the axis parameters a and b, depending themselves on means, variances, and covariance of the Delta scores), it may be assumed that the DIF threshold also depends on such information. Moreover, fixed thresholds have no relationship at all to the prespecified significance

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

level one usually fixes as an acceptable rate for incorrect identification of non-DIF items as DIF. For those reasons, Magis and Facon (2012) proposed the following modification.

On starts by assuming that the Delta points (Δ_{j0}, Δ_{j1}) arise from a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$(\Delta_{j0},\ \Delta_{j1}) \sim \text{BVN}(\boldsymbol{\mu},\ \boldsymbol{\Sigma}) \quad \text{with} \quad \boldsymbol{\mu} = (\mu_0,\ \mu_1)^{\text{T}} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}. \tag{4}$$

There are arguments in favor of this bivariate normality assumption. For one, the shape of the Delta plot (in the absence of DIF) is typical of such joint distributions. Using then the properties of this distribution, it happens that the perpendicular distances D_j follow a (univariate) normal distribution with mean and variance given, respectively, by

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

In sum, item j can now be flagged as DIF if

$$|D_j| > \frac{b\mu_0 + a - \mu_1}{\sqrt{b^2+1}} + \Phi^{-1}(1-\alpha/2)\sqrt{\frac{b^2\sigma_0^2 - 2b\sigma_{01} + \sigma_1^2}{b^2+1}}, \tag{6}$$

where α is the significance level (usually fixed to 5%). Replacing the population parameters by their sample estimates, and by definition of the intercept parameter a, it turns out that the modified Delta plot flags item j if $|D_j| > T_\alpha$ with

$$T_\alpha = \Phi^{-1}(1-\alpha/2)\sqrt{\frac{b^2 s_0^2 - 2b s_{01} + s_1^2}{b^2+1}}. \tag{7}$$

The modified threshold T_α depends now on the significance level α and the shape of the Delta plot (through parameters s_0, s_1, and s_{01}) but not on its location. When the Delta scores are highly correlated (say .98 or .99), all Delta points lie close to the major axis and the threshold T_α becomes smaller. At the extreme case of a perfect correlation, T_α equals zero as b reduces to s_1/s_0. By contrast, when the correlation decreases (i.e., when s_{01} becomes much smaller than s_0/s_1), then the Delta points are more dispersed around the major axis and T_α becomes larger.

This modified Delta plot has several advantages over the traditional fixed-threshold approach (Magis & Facon, 2012). First, it returns empirical Type I error rates much closer to the nominal significance level, while the traditional Delta plot is highly conservative. This is true when DIF is either present or absent. Moreover, the modified Delta plot is far more powerful in detecting true DIF items than the original Delta plot. In addition, it was shown that the modified Delta plot competes well against the Mantel–Haenszel DIF method, in terms of Type I error and power, as long as the samples of respondents are not too large (no more than 100 respondents).

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

Because Mantel–Haenszel is an asymptotic method, it clearly outperforms the Delta plot with large groups of respondents, but it is less accurate than the Delta plot for smaller samples.

Item Purification

Although the modified Delta plot is a clear improvement over the original method, the influence of DIF items on the computation of the threshold T_α is still unknown. Indeed, as *all* items are involved in the computation of both the DIF statistics (3) and the threshold T_α, the presence of DIF items may seriously affect the classification rule. The computation of the perpendicular distances may return spurious values and jeopardize the correct identification of DIF items. This issue relates to the recently introduced idea of viewing DIF items as outliers with respect to the other (non-DIF) items (Magis & De Boeck, 2012).

An item purification process for the modified Delta plot, in the spirit of Candell and Drasgow (1988), could be developed as follows:

(a) Perform the DIF analysis using the modified Delta plot. If none of the items are flagged as DIF, stop. Otherwise go to Step (b).
(b) Compute the means $\bar{x}_0^*$ and $\bar{x}_1^*$, the variances s_0^* and s_1^* and the covariance s_{01}^* of Delta scores of items that are not currently flagged as DIF.
(c) Compute the parameters a^* and b^* of the major axis using the sample estimates from Step (b).
(d) Compute the perpendicular distances D_j^* by plugging-in a^* and b^* into Equation (3) (instead of a and b).
(e) Select an updated DIF detection threshold T_α^* (see below).
(f) If the set of items flagged as DIF (using D_j^* and T_α^*) is identical to the set from the previous iteration, stop. Otherwise rerun Steps (b) to (f) with the latest set of items not flagged as DIF in Step (b).

The aim of the process is clear. At each iteration, the items currently flagged as DIF are withdrawn from the computation of the sample estimates of location and scale of the Delta plot. These estimates are to some extent becoming free of the influence of DIF items and thus better reflect the true location and shape of the ellipsoid of DIF-free items. Similarly, the updated perpendicular distances D_j^* are now computed with respect to the ''purified'' major axis with parameters a^* and b^*. Note that DIF items are withdrawn only from the calculations of Step (b), but the distances D_j^* are still computed for all items in the test.

With traditional DIF methods, this process is complete as such, since the DIF threshold arises from the null statistical distribution and does not depend on the data. With the modified Delta plot, however, the threshold T_α depends directly on sample estimates and the slope of the major axis, which are iteratively adjusted at Steps (b) and (c) of the purification process. The question is, then, how can T_α be iteratively

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

updated to get T_α^*? A first solution is to keep T_α fixed to its initial value, obtained at Step (a), throughout the process. This approach is essentially similar to the usual item purification process, but with the drawback of being influenced itself by the possible presence of DIF. Alternatively, T_α could be fully updated at each iteration, by plugging in s_0^*, s_1^*, s_{01}^*, and b^*. This is intuitively equivalent to processing similarly for the perpendicular distances and the detection threshold. However, this implies that only items not currently flagged as DIF are considered for computing the updated threshold T_α^*. Obviously at each step of the process the sample variances and covariance might decrease dramatically as items with large perpendicular distances are being removed from their computation, yielding the risk of flagging more and more items throughout the iterations. Hence, a fully updated threshold might yield inflated Type I error rates as undesirable side effect. An intermediate solution could therefore be to update T_α^* iteratively, but only by plugging in the updated slope parameter b^*, and keeping the original sample estimates of the ellipsoid. This ensures that the distribution of all Delta points is considered. And since b^* is less influenced by DIF items, the global detection rule could be improved. As a matter of fact, one may expect this approach to return efficiency (in terms of Type I error and power) intermediate between the fixed T_α approach and the fully updated T_α^* approach.

In sum, three item purification processes can be considered. They are schematically equivalent, differing only by the type of DIF detection threshold in use at each step. The three possible thresholds are the following:

$$T_\alpha^* = \Phi^{-1}(1-\alpha/2)\sqrt{\frac{b^2 s_0^2 - 2bs_{01} + s_1^2}{b^2+1}} \equiv T_\alpha \quad \text{(IPP1)}, \tag{8}$$

$$T_\alpha^* = \Phi^{-1}(1-\alpha/2)\sqrt{\frac{b^{*2} s_0^2 - 2b^* s_{01} + s_1^2}{b^{*2}+1}} \quad \text{(IPP2)}, \tag{9}$$

and

$$T_\alpha^* = \Phi^{-1}(1-\alpha/2)\sqrt{\frac{b^{*2} s_0^{*2} - 2b^* s_{01}^* + s_1^{*2}}{b^{*2}+1}} \quad \text{(IPP3)}. \tag{10}$$

The processes are further referred to as IPP (item purification process), each of them with its corresponding number: IPP1 with the initial threshold, IPP2 with the ''partial'' update (b^* only), and IPP3 with the full update. Comparing their efficiency altogether, and with respect to the standard Delta method (without purification), is the topic of the next section.

Simulation Study

The simulation study is a fully crossed factorial design and is broadly similar to that in Magis and Facon (2012). At most six design factors were manipulated: (a) the sizes

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

of the groups of respondents, (b) the number of items in the test (test length), (c) the between-groups difference in average ability level, (d) the proportion of DIF items, (e) the size of the DIF effect, and (f) the DIF methods.

The reference group and the focal group sizes could take the values 50, 100, and 200. The focal group was never larger than the reference group, thus leading to six pairs of sizes: equal sizes 50 (50), 100 (100), 200 (200), and unequal sizes 100 (50), 200 (50), 200 (100). Larger samples were not considered because one may expect the modified Delta plot would be most efficient with small to moderate samples. In the reference group the ability levels θ_{i0} were drawn from a standard normal $N(0, 1)$ distribution, while in the reference group these abilities θ_{i1} were drawn from a normal distribution $N(\gamma, 1)$. The parameter γ reflects the difference in average ability between the groups of respondents. Two values of γ were considered: $\gamma = 0$ (i.e., no difference in average ability) and $\gamma = -1$ (i.e., the focal group has lower average ability).

The underlying IRT model for generating item responses was a closed-form of the two-parameter logistic (2PL) model:

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

where X_{ijg} is the response of subject i from group g ($g = 0, 1$) to item j ($j = 1, \ldots, J$); a_j is the item discrimination level (equal in both groups of respondents) and b_{jg} is the difficulty level of item j in group g (thus possibly different across groups).

The item discrimination levels a_j were drawn from the log-normal distribution with mean 0 and standard deviation 0.1225 (Donoghue & Allen, 1993; Penfield, 2003; Zwick, Donoghue, & Grima, 1993). Item difficulty levels in the reference group b_{j0} were drawn from the standard normal distribution $N(0, 1)$. If item j was not DIF, the difficulty level in the focal group b_{j1} was set equal to that of the reference group: $b_{j1} = b_{j0}$. Otherwise, the DIF size δ was added to the item difficulty: $b_{j1} = b_{j0} + \delta$. This DIF size was set equal for all DIF items present in the data set. This is a somewhat unrealistic assumption, but it permits quantitative comparisons of the absolute effects of DIF across the different methods. Moreover, this approach has already been considered in previous studies (Magis & De Boeck, 2012; Magis & Facon, 2012).

Three test lengths were considered: 20 items, 40 items, and 60 items. For each test length, four percentages of DIF items were investigated: 0% (i.e., no DIF), 5%, 10%, and 20%. The last value is rather extreme, but it helps in assessing the impact of a large proportion of DIF items on the efficiency of the various DIF methods. Finally, two DIF sizes δ were taken into account, always in favor of the reference group: $\delta = 0.4$ and $\delta = 0.8$. The former is frequently considered in DIF studies (Clauser et al., 1993; Donoghue, Holland, & Thayer, 1993) and the latter is included to quantify the impact of larger DIF sizes. Items selected to function differently were randomly drawn among the generated items, in order not to restrict to specific trends in

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

DIF (for instance, very difficult and discriminating DIF items) but rather to focus on a large panel of items possibly affected by DIF.

Four DIF methods were compared: the modified Delta plot without item purification and the same method with each of the three aforementioned IPPs. The original Delta plot was not included because it is clearly outperformed by the modified version (Magis & Facon, 2012). Moreover, since the study was focused on small samples and its main goal was to assess the usefulness of item purification within this framework, the introduction of any other standard DIF method (such as Mantel–Haenszel) was unnecessary.

In summary, the simulation study had 36 settings in the absence of DIF (six combinations of group size, three test lengths, two average ability levels) and 216 settings in the presence of DIF (three percentages of DIF and two DIF sizes in addition to the three aforementioned factors), thus yielding a total of 252 settings. For each setting, 1,000 data sets were generated by drawing at random, for each item and each respondent, the item response from a Bernoulli distribution with success probability given by (11). At most 10 iterations were allowed for the three IPPs and the significance level was fixed at 5%.

The output of the DIF analyses was summarized by computing the empirical Type I error rates (the proportions of non-DIF items that were incorrectly flagged as DIF) and the power values (the proportion of true DIF items that were correctly flagged as DIF) for all four methods. To identify which design factors explained most of the variability of these summary statistics, the Type I error rates and the power values were first transformed to more continuous variables by the logistic transformation. An analysis of variance was then carried out to detect the most important factors and to discard those that explained very little or no variability (less than 1%). Type I errors and power were eventually graphically displayed for each DIF method according to the identified important factors only.

The data generation, the fitting of the DIF methods, and the analysis of their output was performed with the R software version 2.13.1 (R Development Core Team, 2011). The R code can be obtained from the first author.

Results

The results are presented in three parts: Type I error in the absence of DIF, Type I error in the presence of DIF, and power.

Type I Error in the Absence of DIF

In the absence of DIF, the Type I error is best explained by two main factors, the DIF method and the test length, and a pairwise interaction between them. Neither the difference in average abilities nor the group sizes are relevant factors, as neither explains even 1% of the variance. The curves of Type I errors per method and per test length are displayed in Figure 1.

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

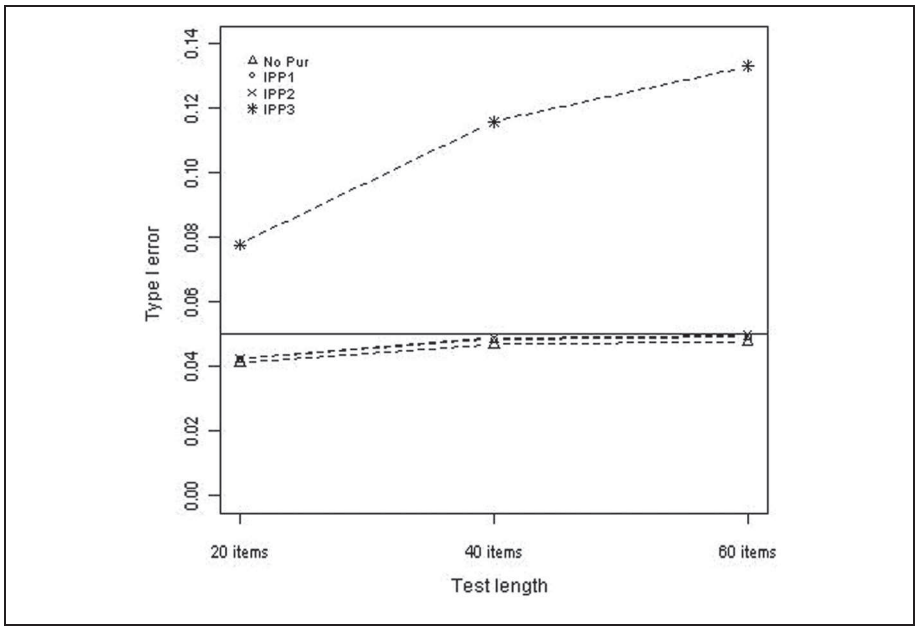

Figure 1. Empirical Type I error rates in absence of DIF.
Note. "No Pur" refers to the modified Delta plot without purification. The horizontal line represents the nominal significance level of 5%.

As a general trend, the Type I errors increase with the test length, and the increase is much faster for the IPP3 method than for all other methods. Moreover, the modified Delta plot and both IPP1 and IPP2 approaches yield closely similar Type I errors, although the modified Delta plot is slightly more conservative. It is clear, however, that even without purification, the modified Delta plot controls very well for Type I error inflation (conservativeness). On the other hand, the IPP3 method is clearly inappropriate in this context, as its Type I error rate is much larger than the nominal significance level. This method flags far more items as DIF than expected.

Type I Errors in the Presence of DIF

When DIF is included in the data, the type I errors are best explained by four factors: the DIF method, the group sizes, the percentage of DIF items, and the size of DIF. Neither the difference in average ability nor the test length is a relevant factor. The curves of Type I errors are displayed in Figure 2.

As group sizes increase, the Type I error rates tend to decrease slightly for all but the IPP3 method, whose Type 1 error rate is roughly constant across group size. Moreover, as when there is no DIF, the IPP3 method returns Type I errors much larger than the nominal significance level. The three other methods, on the other

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

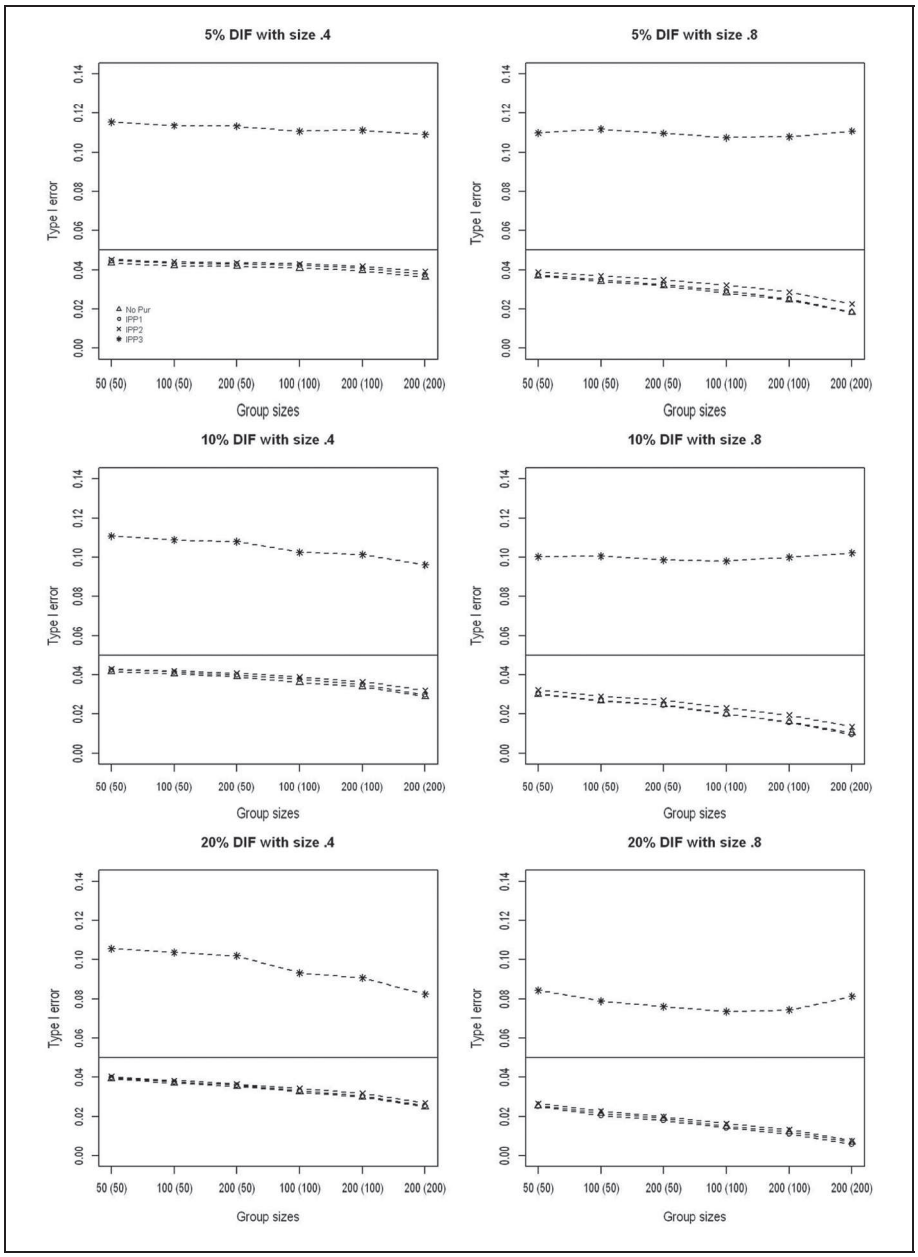

Figure 2. Empirical Type I error rates in presence of DIF, as a function of reference (focal) group sizes, percentages of DIF 5% (top), 10% (middle), and 20% (bottom) and DIF sizes .4 (left) and .8 (right).

Note. "No Pur" refers to the modified Delta plot without purification. The horizontal line represents the nominal significance level of 5%.

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

hand, still return closely similar values, though the IPP2 method displays slightly larger Type I errors than the other two.

As already noticed (Magis & Facon, 2012), the modified Delta plot is slightly conservative overall but with decreasing Type I errors when both the percentage of DIF and the DIF size increases. Surprisingly, both IPP1 and IPP2 follow the same trend and cannot therefore overcome this conservative trend.

Power

The power values were best explained using the factors that best explained the Type 1 error rates. Consequently, the display of the power curves in Figure 3 is highly similar to Figure 2.

First, obviously straightforward conclusions are drawn: independently of the DIF method, the power increases with group sizes and with the size of DIF, while it decreases with the percentage of DIF. Moreover, IPP3 returns consistently larger values than the three other methods. This has to be counterbalanced, however, by the fact that the Type I errors for IPP3 are also consistently larger. All things being equal, this is another indication that the IPP3 method is inappropriate as a purification criterion for identifying DIF items.

When the DIF is small (.4), IPP1 and IPP2 yield power similar to that of the modified Delta plot. The gap is larger, however, with larger DIF (.8): IPP2 tends to be slightly more powerful, followed by IPP1 and the modified Delta plot. This gap tends also to increase when the percentage of DIF items increases.

It should be noticed that the power values are low overall. Only the situation of large samples (two times 200) and large DIF sizes (.8) yield convenient power. This is not surprising since this study focused on small samples and rather small DIF sizes. Moreover, it was shown (see, e.g., Magis & Facon, 2012) that the Mantel–Haenszel is not very powerful with such a design. Rather, this method and the modified Delta plot returned similar power values. The lack of power is therefore due to insufficient amount of information because of insufficiently large samples of respondents. Larger samples and/or larger DIF sizes would obviously return larger power, and maybe that item purification processes would even improve the results. However, the accuracy of the modified Delta plot may be doubtful with larger samples, as highlighted by Magis and Facon (2012).

Discussion

The IPP3 method is clearly an inappropriate purification method. It returns Type I errors larger than the significance level, and consequently larger power values. Thus, as it was initially hypothesized (see section ‘‘Item Purification’’), the method flags too many items as DIF. The two other purification processes, however, perform similarly in terms of Type I error and power, with a very slight superiority for IPP2 (slightly less conservative and larger power with large DIF). Nevertheless, neither of

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

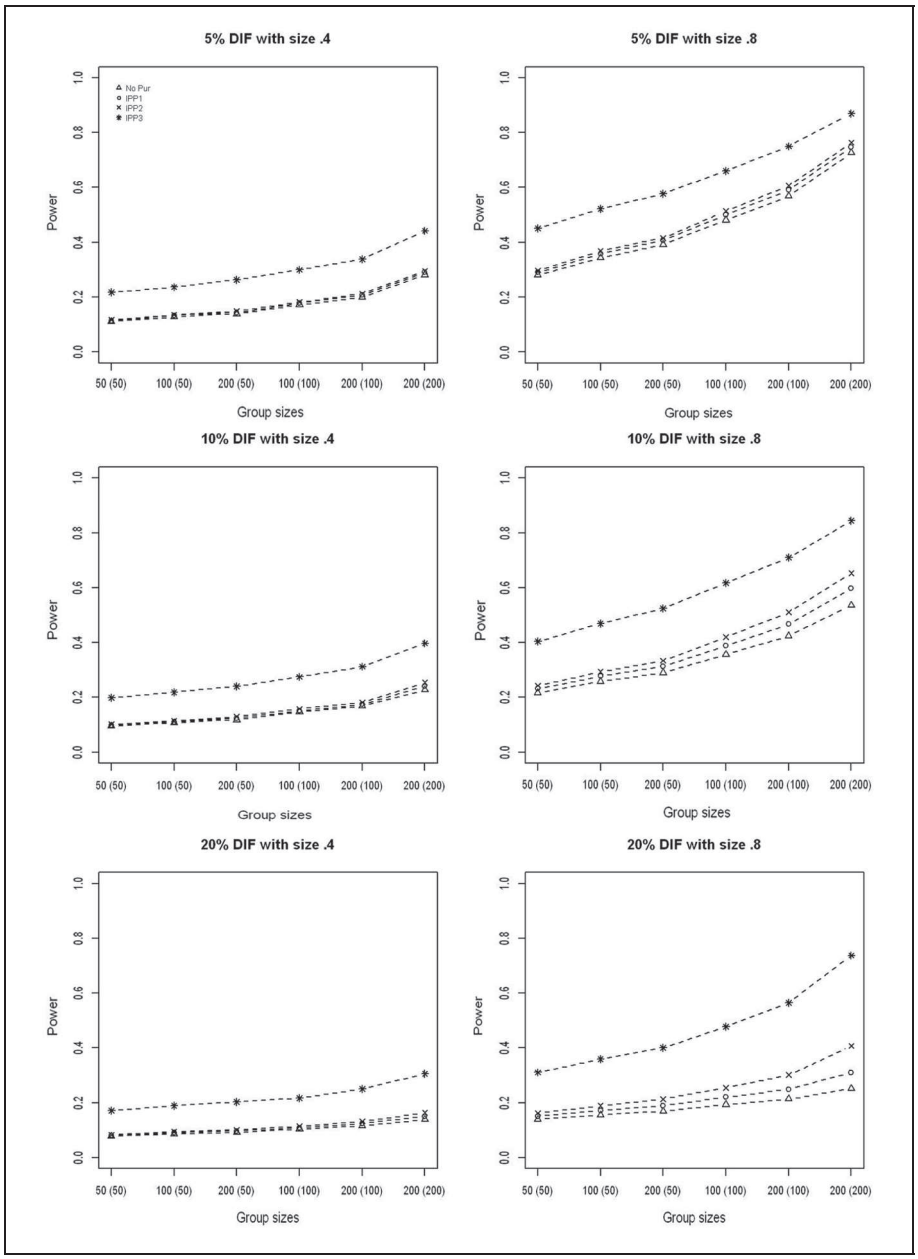

Figure 3. Empirical power values, displayed as a function of reference (focal) group sizes, percentages of DIF 5% (top), 10% (middle), and 20% (bottom) and DIF sizes .4 (left) and .8 (right).

Note. "No Pur" refers to the modified Delta plot without purification.

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

these purification methods clearly outperformed the modified Delta plot. In sum, unlike the traditional methods, for which item purification constitutes a clear improvement of the DIF detection results, with the modified Delta plot this improvement is barely visible and hence far less convincing. It is worth mentioning it as many studies advocated for a systematic use of item purification of the DIF detection results.

Application

The modified Delta plot method and the three item purification processes are now illustrated by analyzing the so-called Boehm data set described in Facon, Magis, and Courbois (2012). The data were collected by administering an adaptation of the French-language versions of the Boehm tests of basic concepts. This adaptation, designed to test preschool and kindergarten children, consists in 72 items administered as follows: Each respondent was proposed four to six drawings and had to select the one corresponding to a concept spoken by the examiner. The entire pool of items was administered to three groups of respondents. The first group comprised typical participants aged about 5 years who had never been referred for a psychological assessment, the second group comprised participants with Down syndrome, and the third group comprised participants with intellectual disability of undifferentiated etiology. Forty-seven participants were selected from each the three groups such that each triad had exactly the same total score. In the present analysis, only the first two groups are considered. The reference group was set as the group of typical children and the focal group as the group of participants with Down syndrome.

With the modified Delta plot without item purification, three items (12, 36, and 40) were flagged as DIF. The threshold T_α was equal to 2.005. With IPP1, three iterations were necessary for the process to stop, T_α was also equal to 2.005, and Item 9 was flagged as DIF in addition to 12, 36, and 40. With IPP2, three iterations were also necessary. The final threshold T_α^* was equal to 1.867 and the same items were flagged as DIF as those by IPP1. Finally, six iterations were necessary with IPP3. The final threshold equaled 1.389, and 11 items were flagged as DIF in total. Table 1 summarizes these DIF items, together with their perpendicular distances (after the last step of the purification process) for the four methods.

These results confirm that both IPP1 and IPP2 return similar results to the nonpurified Delta plot, while IPP3 tends to flag more items than the other three methods. This can also be seen from the number of iterations required: six for IPP3 versus less than four for the other approaches.

For completeness, the Delta plots of the four methods are displayed in Figure 4. Delta points are drawn by means of empty triangles, and full dark triangles indicate that several items are located exactly on the same Delta point. Items flagged as DIF are encircled. Also, the solid line represents the major axis of the ellipsoid, while the two dashed lines are the upper and lower bounds for flagging items as DIF. Clearly, all but the lower right panel (corresponding to IPP3) yield highly similar results, as

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

Table 1. Perpendicular Distances of the Items Flagged as DIF by at Least One Method.

Item	Delta Plot	IPP1	IPP2	IPP3
9	−1.890	−2.062[a]	−2.062[b]	−1.958[c]
11	1.960	1.857	1.857	1.897[c]
12	−2.836[a]	−3.054[a]	−3.054[b]	−2.908[c]
17	−1.412	−1.581	−1.581	−1.480[c]
24	1.732	1.494	1.494	1.658[c]
33	1.720	1.560	1.560	1.653[c]
36	−2.582[a]	−2.754[a]	−2.754[b]	−2.650[c]
40	−2.228[a]	−2.326[a]	−2.326[b]	−2.290[c]
60	1.777	1.636	1.636	1.711[c]
65	1.568	1.431	1.431	1.503[c]
72	−1.803	−1.849	−1.849	−1.860[c]

Note. DIF = differential item functioning; IPP = item purification process. "Delta plot" refers to the modified Delta plot without item purification.

[a]Items flagged as DIF with respect to threshold T_α equal to 2.005 ($\alpha = 0.05$) without purification or under IPP1.

[b]Items flagged as DIF with respect to threshold T_α equal to 1.867 ($\alpha = 0.05$) under IPP2.

[c]Items flagged as DIF with respect to threshold T_α equal to 1.389 ($\alpha = 0.05$) under IPP3.

expected, while the two bounds for IPP3 are much closer to the major axis, yielding overidentification of items as functioning differently.

Summary

The purpose of this article was to explore the results of adapting item purification for use with the modified Delta plot. Unlike with the traditional DIF methods, there are several ways of implementing purification with Delta plot by adjusting the detection threshold. The present simulation studies showed that a full adjustment of this threshold (the IPP3 method) is not suitable because it strongly inflates the Type I error rates. The two other processes, IPP1 (keeping the threshold constant) and IPP2 (updating the slope parameter only), return similar results and do not constitute a clear improvement of the modified Delta plot (without purification).

In sum, item purification with the modified Delta plot in small-sample studies constitutes a remarkable counterexample of its usefulness with respect to other traditional DIF methods. Indeed, it has been often advocated that item purification constitutes a clear improvement for the DIF methods, by limiting Type I error inflation and increasing power. This study demonstrated that item purification is not improving the results of all DIF methods.

One possible reason for not improving the Delta plot by item purification is that the detection threshold depends on the sample of Delta scores itself, while with traditional DIF methods the threshold is determined directly from an underlying statistical distribution. Since the threshold is determined itself by the sample Delta scores, it can

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

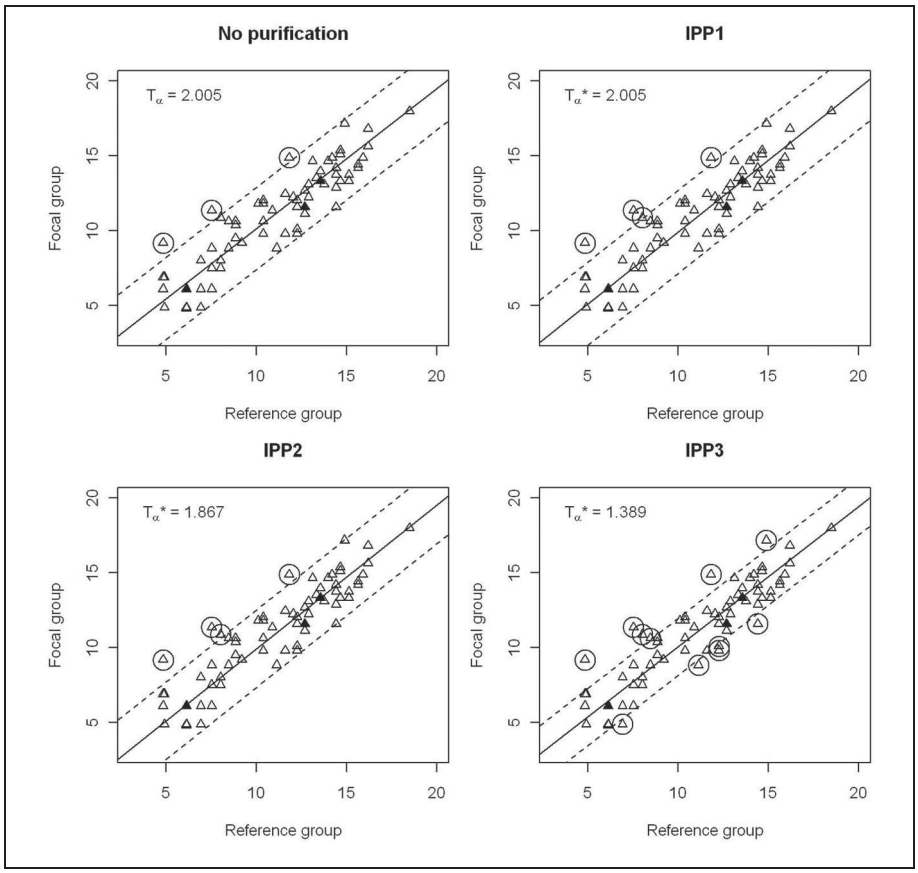

Figure 4. Delta plots from the Boehm data set, using the modified Delta plot without purification (upper left), with IPP1 (upper right), with IPP2 (lower left), and with IPP3 (lower right).

Note. Full black triangles refer to multiple items having the same Delta point as coordinates. Items flagged as DIF are rounded by a circle. Full and dashed lines refer to, respectively, the major axis and the bounds for DIF item detection.

adjust itself to situations for which Type I inflation would possibly occur with traditional DIF methods. For instance, it may happen that Delta scores are much more dispersed around the major axis, which would lead to larger DIF statistics and inflated Type I errors with traditional test score methods (such as Mantel–Haenszel or traditional Delta plot). With the modified Delta plot, however, the increased variability in test scores is taken into account in the classification threshold (7) (through the sample variances and covariance), so that it can adjust and keep stable Type I error rates at no cost of power. Because this adjustment was shown to be most optimal (Magis &

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

Facon, 2012), one was able to predict that the DIF results can hardly be expected to be significantly improved by item purification.

Another possible reason of the failure of item purification in improving the DIF results is that this process is often primarily used to control for Type I error inflation. By an iterative purification of the set of anchor items, one is able to better distinguish between the DIF and non-DIF items, thus lessening the proportion of false identifications (i.e., Type I error) and increasing the proportions of correct identifications of DIF items (i.e., power). The modified Delta plot, however, was shown to have stable and not inflated Type I error rates. With regard to this, it is coherent that item purification does not improve both the Type I error control and the power.

It is worth mentioning that, though the most popular method, item purification is not the only one that permits control of Type I error inflation. Considering DIF items as outliers, Magis and De Boeck (2011, 2012) proposed to lessen the impact of DIF items by replacing sample estimates by statistically robust estimators. Basically this yields a one-step process, similar to the standard application of DIF methods, but wherein the effect of large outlying (DIF) measures is significantly decreased. However, the present study shows once again that Type I error is under good control using the modified Delta plot, so there is no guarantee that introducing robust estimators instead of the sample values will much improve Type I error rate or power. Instead, both approaches are similar in the sense that DIF items are flagged with respect to the test items, by using sample estimates (Delta scores in this study) to determine the classification threshold. This point deserves further investigation.

This study is restricted to small samples because it is the optimal framework for the Delta plot method. However, it would be interesting to perform the same kind of study with larger samples, for which it is known the Delta plot is generally less suited. Item purification might be more useful in that setting.

Acknowledgments

We would like to thank George A. Marcoulides, editor, and two anonymous referees for their comments and suggestions. We are also grateful to John M. Belmont for his helpful reading of the article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a postdoctoral grant ''Chargé de Recherches'' of the National Fund for Scientific Research/Fonds National pour la Recherche Scientifique (Belgium), the Research Funds of KU Leuven, and the National Research Agency/

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

Agence Nationale de la Recherche (France). ANR, LANG, & HANDICAPS, Project no. ANR-09-ENFT-019.

References

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-106.

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.

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.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.

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.

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* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

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.

Facon, B., Magis, D., & Belmont, J. M. (2011). Beyond matching on the mean in developmental disabilities research. *Research in Developmental Disabilities, 32*, 2134-2147.

Facon, B., Magis, D., & Courbois, Y. (2012). On the difficulty of relational concepts among participants with Down syndrome. *Research in Developmental Disabilities, 33*, 60-68.

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.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*, 925-936.

Fidalgo, A. M., Mellenbergh, G. J., & Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research, 5*, 43-53.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*, 432-457.

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.

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Lightbody, A., & Reiss, A. (2009). Gene, brain, and behavior relationships in fragile X syndrome: Evidence from neuroimaging studies. *Developmental Disabilities Research Reviews, 15*, 343-352.

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

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research, 46*, 733-755.

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent Type I error inflation in DIF. *Educational and Psychological Measurement, 72*, 291-311.

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*, 302-321.

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, 154C*, 229-248.

Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Applied Psychological Measurement, 34*, 365-369.

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

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

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.

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

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

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, Leiden, Netherlands.

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

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*, 53-75.

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.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*, 113-144.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

Downloaded from epm.sagepub.com at Universite de Liege on March 1, 2013