Psychological Test and Assessment Modeling, Volume 57, 2015 (3), 342-376

Determinants of artificial DIF – a study based on simulated polytomous data

Curt Hagquist[1] *& David Andrich*[2]

Abstract

A general problem in DIF analyses is that some items favouring one group can induce the appearance of DIF in others favouring the other group. Artificial DIF is used as a concept for describing and explaining that kind of DIF which is an artefact of the procedure for identifying DIF, contrasting it to real DIF which is inherent to an item.

The purpose of this paper is to elucidate how real both uniform and non-uniform DIF, *referenced to the expected value curve*, induce artificial DIF, how this DIF impacts on the person parameter estimates and how different factors affect real and artificial DIF, in particular the alignment of person and item locations.

The results show that the same basic principles apply to non-uniform DIF as to uniform DIF, but that the effects on person measurement are less pronounced in non-uniform DIF. Similar to artificial DIF induced by real uniform DIF, the size of artificial DIF is determined by the magnitude of the real non-uniform DIF. In addition, in both uniform and non-uniform DIF, the magnitude of artificial DIF depends on the location of the items relative to the distribution of the persons. In contrast to uniform DIF, the direction of non-uniform real DIF (e.g. favouring one group or the other) is affected by the location of the items relative to the distribution of the persons. The results of the simulation study also confirm that regardless of type of DIF, in the person estimates, artificial DIF never balances out real DIF.

Keywords: differential item functioning, uniform, non-uniform, artificial, Rasch models

[1] *Correspondence concerning this article should be addressed to:* Curt Hagquist, PhD, Centre for Research on Child and Adolescent Mental Health, Karlstad University, SE-651 88 Karlstad, Sweden; email: curt.hagquist@kau.se

[2] The University of Western Australia

Introduction

Independent work on requirements of invariance of comparisons for measurement by the Danish mathematician Georg Rasch (1961) incorporated ideas of Thurstone (1928) and Guttman (1950) into a probabilistic response model in which invariance is an integral property. Rasch's requirements implied that any partition of the data should provide invariant comparisons:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (p.322; Rasch, 1961).

It follows that in order to provide meaningful comparisons of different groups, the comparisons of the stimuli of a measuring instrument have to be invariant, not only along the variable of assessment, but also across the groups to be compared. In this paper instruments refer to tests or questionnaires and therefore the stimuli are referred to as *items*. Because the variable of assessment is inferred from the assessment by the items, it is generally referred to as a *latent* variable. Further references to a variable in this paper are understood to refer to such a variable.

Lack of invariance of the comparisons of item parameters across sample groups is commonly called differential item functioning (DIF). However, DIF may also be used as a generic term to include the lack of the same kind of invariance along the variable. Analysis of DIF in terms of parameter estimates across sample groups has long been used in Rasch model analyses (Andrich & Kline, 1981; Andrich, 1988), although the terminology has changed and new procedures for detecting DIF have been developed.

Among the new procedures for detecting DIF with both Rasch measurement theory and item response theory models, which do not estimate and compare item parameters from different groups, the expected value curve (EVC) of the responses of groups to an item is used. DIF across different groups implies that for the same values of the variable, the EVC of the response to an item for members of the groups are different. If the differences along the variable are homogeneous, then the DIF is referred to as *uniform*; otherwise it is referred to as *non-uniform*. Although DIF can be referenced and studied to multiple groups (e.g. DIF across countries), it has generally been focused on two groups (e.g. DIF across genders). In addition, in some cases one of the groups is considered the standard and dominant group, and the other a focal or minority group. In this paper we focus on two groups of equal standing. Although the data are simulated, to simplify the presentation of the results of the study, we refer to one group as Boys, and the other as Girls.

For items consisting of only two ordered categories, the expected value is the same as the probability of a positive response. In that case the expected value curve (EVC) is known as the item characteristic curve (ICC). In the context of an analysis of multiple items of a

test or questionnaire, and as elaborated further in the ICCs of an item for different groups can be estimated by resolving the item so that an item is created in which persons from only one group respond (Andrich and Hagquist, 2012). For consistency with the literature, we continue to use the term ICC for the expected value curve in the case of dichotomous items. However, in the context of polytomous items, and because of its specific relevance to the particular DIF investigated, the more general term EVC will be used.

In the dichotomous Rasch model, which has only an item location parameter, the ICCs of all items are parallel. Therefore a difference between ICCs for an item for two groups implies only a difference between the locations of the ICCs. In that case it is possible for responses to fit the ICC for each group statistically even though the locations of the item for the two groups are different. Because the ICCs are parallel, if the responses fit the ICC the differences between the groups are homogeneous across the variable and therefore the DIF is uniform and can be quantified by the difference in the locations for the item between the two groups. On the other hand, it is possible that even after resolving an item and estimating a separate location for the item for each group, the responses do not fit the parallel ICCs. In that case the differences are not homogeneous across the variable and the DIF is non-uniform. Thus, in contrast to uniform DIF, items with non-uniform DIF cannot be resolved while retaining statistical fit to the model.

Real and artificial DIF

The emphasis in the study of DIF has been on uniform DIF with dichotomous items especially in the context of educational assessment. Over the years a number of different procedures have been used for detecting DIF with dichotomous items of which the Mantel-Haenszel (MH) procedure seems to be the most popular one (Holland & Thayer, 1988; Holland & Wainer, 1993; Osterlind & Everson, 2009). A general problem observed in various DIF analyses, including the MH method, is that some items favouring one group can induce the appearance of DIF in others favouring the other group when in fact no DIF is present. Referring to outcomes based on simulated data, Tennant & Pallant (2007) reported that: "Although not significant, all items showed some level of DIF, indicating how the presence of two items favouring males forced other items to favor females" (p. 1083). They also noted that " …the compensatory DIF may lead the analyst to presume that DIF is cancelling out, but clearly there is a significant impact on individual estimates, and some impact on group estimates" (Tennant & Pallant, 2007, p. 1083). This confusion was compellingly described, but not explained, by Wang & Su (2004) for the MH procedure: "…the M-H-1 tends to declare mistakenly those DIF-free items as DIF items favouring the focal group and to declare mistakenly those DIF items as DIF-free items under these conditions. This tendency makes the total DIF yielded via M-H-1 more or less balanced between groups, when in fact all or most of the DIF items are simulated to favour the reference group" (p. 141).

Summarizing the state of the art of DIF Osterlind & Everson (2009) reported similar observations: ''Sometimes, for reasons unknown, calculations of a DIF detection strategy may suggest DIF, where none truly exists'' (p. 21).

Recently these kinds of observations have been explained. Andrich & Hagquist (2012) used the Rasch model for dichotomous responses as the theoretical basis for the MH procedure for the detection of DIF to introduce the concept of *artificial* DIF: artificial DIF is an artefact of the procedure for identifying it. Contrasting artificial DIF with real DIF which is inherent to an item, Andrich & Hagquist state: "The reason that the MH procedure, as analysed through the Rasch model, generates artificial DIF can be traced to the substitution of the estimates of the person locations for unknown values. In the Rasch model, the person total score is a sufficient statistic for the person parameter estimate, and therefore grouping persons by total scores is equivalent to grouping persons according to their estimates." (Andrich & Hagquist, 2012, p. 413).

Distinguishing between real and artificial DIF is decisive for understanding and interpreting correctly outcomes from DIF analyses. Misidentifying artificial DIF items as having real DIF can give the mistaken impression that the effect of DIF among items where some items favour one group and others favour the other group cancels out in comparing mean estimates of groups. Failure to distinguish between real and artificial DIF may also violate the requirement of invariant properties of an instrument, e.g. by removing or resolving artificial DIF items as if they were real DIF items.

Andrich & Hagquist (2012) showed algebraically that real DIF in one item is distributed as artificial DIF across all other items, that the magnitude of artificial DIF is a function of the number of items with real DIF, the direction of the DIF, and the total number of items. They also showed that if an item is resolved to an item for each group, then it no longer induces artificial DIF in other items. As a result, Andrich & Hagquist also showed that the magnitude and direction of DIF in an item could be estimated by resolving the item and providing a new unique item for each group. In addition, they showed that the logical process for identifying all items with real DIF required a sequential process of resolving items, beginning with the item which showed the greatest initial DIF. The process continued till no further DIF, at the level of the power of detection, was present, leaving as identified a set of items with no DIF, a "pure" set. Of course, because the DIF identified is relative to the whole set of items analysed, the process presumes that only a minority of items have real DIF.

Andrich & Hagquist (2014) generalised the concept of artificial DIF to polytomous items using the polytomous Rasch model (PRM), a generalisation of the dichotomous Rasch model. However, instead of using the MH procedure to identify items with DIF, which may be real or artificial, they used a two way analysis of variance (ANOVA) of person-item residuals classified by class intervals on the variable and by groups. This procedure has the advantage that it immediately, and simultaneously, provides evidence of general fit of the responses of each item to the model across the variable together with evidence for both uniform and non-uniform DIF. Because the expected values of responses used to calculate residuals from the observed responses are obtained using parameter estimates from the data for each person and item, then as in the MH procedure, the procedure will induce artificial DIF. Andrich and Hagquist demonstrated that the same principles hold for polytomous responses using the ANOVA method for detecting DIF as with the MH procedure for dichotomous items. In particular, resolving items sequentially, beginning with the item with the apparent greatest DIF, successively eliminates artificial DIF in-

duced by that item in the remaining items, thus permitting the identification and quantification of items with real DIF. Andrich and Hagquist focussed on uniform DIF with polytomous items.

Uniform and non-uniform DIF

This paper is concerned with studies of the impact of *both* uniform and non-uniform real DIF in inducing both uniform and non-uniform artificial DIF when assessed using the PRM and the method of ANOVA of residuals and quantified using the resolution of items which show DIF. However, non-uniform DIF for a polytomous item referenced to the EVC in the PRM is more complex than for dichotomous items and is briefly summarized here.

As shown formally later in the paper, an item with $m+1$ categories has m threshold parameters at which the probability of a response in two adjacent categories is identical. In the dichotomous Rasch model where $m+1=2$, there is only one threshold (the item location), and in proficiency assessment, this threshold is commonly referred to as a *difficulty*. Then the probability of success at the threshold as a function of the value on the variable is simply the ICC.

In the dichotomous Rasch model there is a manifest dichotomous response between the categories located at the threshold which is graphically represented by a probability curve for the item, i.e. the ICC. In the PRM there is a *latent* dichotomous response between pairs of adjacent categories located at each threshold which is modelled by the dichotomous Rasch model (Andrich, 1978). In this case, the modelled latent dichotomous response is referred to as a *threshold probability curve*. As with the ICCs in the dichotomous Rasch model, the threshold probability curves *within* and *between* items in the PRM are also all parallel. However, in contrast to the dichotomous Rasch model, the EVCs for the PRM do not have to be parallel. The slope of the EVC of a polytomous item is a function of the distances between its thresholds – the closer the thresholds, the steeper the slope. Therefore it is possible for different items in the PRM to have non-parallel EVCs yet still fit the model.

As a result of this same feature, there are three different kinds of DIF with polytomous responses modelled by the PRM. First, on resolving an item by groups, it can have parallel EVCs located at different points on the variable for different groups in which the responses fit the model for each group. In this case the DIF is uniform and it is the one studied in Andrich and Hagquist (2014). Because the responses fit the model, it can be inferred that the *latent* threshold probability curves are parallel as required by the PRM. Thus this case is analogous to that of uniform DIF with dichotomous items.

Second, on resolving an item by groups, it may have non-parallel EVCs located at the same or different points on the variable for different groups in which the responses again fit the model for each group. In this case, in relation to the EVC, the DIF is non-uniform. Because the responses fit the model, it can be inferred that the latent threshold probability curves are parallel as required by the PRM. Thus this case of non-uniform DIF is analogous to uniform DIF with dichotomous items but has no counterpart with non-

uniform DIF in dichotomous items. This is the case studied in this paper – that is, on the resolution of an item for different groups, the responses fit the model and the non-uniform DIF is a function of the relative locations of the thresholds for each group.

Third, on resolving an item by groups, it may have parallel or non-parallel EVCs located at the same or different points on the variable for different groups, despite the resolution because of perceived DIF, the responses still do not fit the model for each group. Because the responses do not fit the model, it can be inferred that the latent threshold probability curves are *not* parallel as required by the PRM. This case is analogous to that of non-uniform DIF in dichotomous items which inevitably implies misfit to the dichotomous Rasch model. We are not concerned with this case in the paper.

In summary, the purpose of this paper is to elucidate how real both uniform and non-uniform DIF, *referenced to the EVC*, induce artificial DIF, how this DIF impacts on the person parameter estimates and how different factors affect real and artificial DIF, in particular the alignment of person and item locations. All examples involve polytomous items, which on resolution of an item by the group, and irrespective of whether the DIF is uniform or non-uniform DIF, show fit to the model. That is, in the case of non-uniform DIF, it concerns the second case above. The importance of this case is analogous to that with uniform DIF discussed in Andrich and Hagquist (2014). Although on resolution of the items by groups the responses may fit the PRM, the parameters of an item are no longer invariant with respect to the two groups. Thus there is a trade-off between the requirement of model fit and the requirement of the invariance of item parameter estimates between the groups. On resolution of the items which improves the fit, and depending on the magnitude of the real DIF in the items, the relative person estimates in the two groups will be changed. Whether it is more valid to resolve an item and obtain fit than to not resolve and retaining invariant estimates for the item parameters in obtaining person estimates for comparison, depends on each particular context, including the nature of the definition of the variable (Andrich and Hagquist, 2014).

The first set of research questions address how artificial DIF is affected by the proportion of real DIF items, the magnitude of the real DIF in the items, the direction of the real DIF items, whether favouring the same or opposite groups and the distribution of the persons along the variable relative to the location of the items. The second set of research questions examine the effects on artificial DIF if real DIF items are resolved, the effects on real DIF if artificial DIF items are resolved, and on the person estimates in the cases that items with real DIF or the artificial DIF are resolved. All studies involve simulated data.

Methods

The polytomous Rasch model and the DIF-analysis

The Rasch model for items with ordered response categories which follows from Rasch's requirement of invariance (Rasch, 1961; Andersen, 1977, Andrich, 1978; Wright and Masters, 1982), where dichotomous response categories are a special case, can be expressed in the form

$$\Pr\{X_{ni} = x\} = \exp(x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_{ki}) / \gamma_{ni}\,, \tag{1}$$

where $X_{ni} = x \in \{0,1,...m_i\}$ are the scores associated with the $m_i + 1$ successive categories of item i, τ_{ki} are m_i thresholds defining the $m_i + 1$ successive categories on the continuum, $\tau_{ki} \equiv 0, \sum_{k=}^{m_i} \tau_{ki} = 0$, β_n and δ_i are the respective location parameters of person n and item i on the same continuum, generally referred to as a variable, and γ_{ni} is a normalising factor. Clearly, with a single location parameter of the persons the model is unidimensional. This reflects the requirement of data in formal measurement.

The concept of latent threshold probability curves was referred to in the Introduction. For completeness, we note that the probability of a response with score x relative to a score of $x-1$ or x is, on simplification, given by

$$\begin{aligned} \frac{\Pr\{X_{ni} = x\}}{\Pr\{X_{ni} = x-1\} + \Pr\{X_{ni} = x\}} &= \Pr\{X_{ni} = x \mid X_{ni} = x-1 or\ x\} \\ &= \exp(\beta_n - \delta_i) - \tau_{xi})) / \gamma_{ni} \\ &= \exp(\beta_n - (\delta_i + \tau_{xi})) / \gamma_{ni} \\ &= \exp(\beta_n - \delta_{ix}) / \gamma_{ni}, \end{aligned} \tag{2}$$

where δ_{ix} is the threshold $x, x = 1,2,...m_i$ referenced to the same origin as that of all items rather than to δ_i and where $\gamma_{ni} = 1 + \exp(\beta_n - \delta_{ix})$. Eq. (2) is the dichotomous Rasch model. These are the curves, referred to in the Introduction, that are parallel in the PRM for all thresholds of all items. The EVC is central to the studies and the definition of DIF for the paper, and it is given by

$$E[X_{ni}] = \sum_{x=0}^{m_i} x \Pr\{X_{ni} = x\}\,, \tag{3}$$

which from Eq. (1) is a function of β. Graphs of the EVC under different circumstances are shown later in the paper. In the case of a dichotomous response, $m_i = 1$, $E[X_{ni}] = \Pr\{X_{ni} = 1\}$ and is the curve known as the ICC.

The algebra of artificial DIF

In the Rasch model the item parameters can be estimated by conditioning on the person total scores, i.e. the sum of the responses of persons across items. This total score is a sufficient statistic that fully characterise a person's profile of responses, given that the data fit the PRM. In that case, the item parameter estimates are obtained independently of the person parameters. In the estimation of the item parameters a constraint is imposed implying that within each score group the expected values of the responses sum to the actual total score. It therefore follows that real DIF in one item favouring one of two sample groups will inevitably induce artificial DIF in the other items favouring the opposite group.

In anticipation of the results of the present simulation study, item set 1 in Table A1 (appendix A) consisting of four items is used below to demonstrate the mathematics of artificial DIF. In that item set, Item 2 is simulated to have a 0.5 logit uniform DIF favouring girls with a location of -1.0 for girls compared to -0.5 for boys.

Because all four items in item set 1 have the same number of categories, the same parameter values for the thresholds and the data are complete, in subsequent expressions the subscript i is dropped from the maximum score m_i.

Taking the item parameters obtained from a conditional method of estimation as known, and assuming that all persons responded to all items, the maximum likelihood (ML) solution equation for estimating the person parameter is given by

$$r = \sum_{i=1}^{I} \sum_{x=0}^{m} x \Pr\{X_i = x \mid \hat{\beta}_r\}, \tag{4}$$

where $\hat{\beta}_r$ is the estimate of all persons with a total score of $r = r_n = \sum_{i=1}^{I} x_{ni}$ for all persons n, $n = 1,2,...N_r$, N_r is the number of persons with a score of r, and where I is the total number of items.

Let $p_{xri} = \Pr\{X_i = x \mid \hat{\beta}_r\}$. Then for persons with score r, $E[X_{xri}] = \sum_{x=0}^{m} x p_{xri}$.

Because the location for Item 2 in set 1 of Table A1 is greater for boys than for girls, for all β_r,

$$E[X_{xr2} \mid Girl] = \sum_{x=0}^{m} x p_{xr2} \mid Girl > E[X_{xr2} \mid Boy] = \sum_{x=0}^{m} x p_{xr2} \mid Boy \ . \tag{5}$$

We focus on Item 2 from the set 1 of four items in Table A1, and write Eq. (4) in the form

$$r = \sum_{i=1}^{4} \sum_{x=0}^{m} x p_{xri} = \sum_{x=0}^{m} x p_{xr2} + \sum_{x=0}^{m} x p_{xr1} + \sum_{i=3}^{4} \sum_{x=0}^{m} x p_{xri} \ . \tag{6}$$

The constraint in Eq. (4) and the inequality for Item 2 in Eq. (5), imply that the terms which involve Items 1, 3 and 4 on the right side of Eq. (6) must satisfy the inequality

$$\sum_{x=0}^{m} x p_{xr1} + \sum_{i=3}^{4} \sum_{x=0}^{m} x p_{xri} \mid Girl < \sum_{x=0}^{m} x p_{xr1} + \sum_{i=3}^{4} \sum_{x=0}^{m} x p_{xri} \mid Boy \ . \tag{7}$$

Moreover,

$$\begin{aligned} & E[X_{xr2} \mid Girl] - E[X_{xr2} \mid Boy] \\ &= E[X_{xr1} \mid Boy] - E[X_{xr1} \mid Girl] + \sum_{i=3}^{4} E[X_{xri}] \mid Boy - \sum_{i=3}^{4} E[X_{xri}] \mid Girl. \end{aligned} \tag{8}$$

Thus Eq. (8) shows that the real DIF in Item 2 with a higher score for girls induces artificial DIF with a higher score for boys in the other items, and further, that this artificial DIF is distributed across all three remaining items.

Simulation of data

All simulated DIF is understood to be real, even though in the analysis of the simulated data, both real and artificial DIF will be evident using the initial method of detection of DIF. Fifteen different sets of data consisting of items with five response categories were simulated to reveal uniform DIF, varying with respect to the total numbers of items, the proportion of items with uniform DIF, the magnitude of uniform DIF and the mean of the person distribution relative to the mean of the item parameters.

Seven different sets of data consisting of items with five response categories were simulated to reveal non-uniform DIF, varying with respect to the total numbers of items, the proportion of items with non-uniform DIF, and the mean of the person distribution.

All data were simulated according to the polytomous Rasch model implying equal discrimination at the item thresholds.

Uniform DIF

The data sets simulated to reveal uniform DIF were:

- Three sets of data with DIF of different magnitude (0.5; 1.0; 1.5) on one of four items (person mean of 0.00 and a person standard deviation of 2.0),
- Four sets of data with DIF of different magnitude (0.5; 1.0; 1.5) on one of eight items (person mean of 0.00 and a person standard deviation of 2.0),
- One set of data with DIF of the magnitude of 1.5 on one of eight items with a person mean of -3.00 and a person standard deviation of 2.0,
- One set of data with DIF of the magnitude of 1.5 on one of eight items with a person mean of +3.00 and a person standard deviation of 2.0,
- Three sets of data with DIF of different magnitude (0.5; 1.0; 1.5) on two of eight items favouring the same groups (person mean of 0.00 and a person standard deviation of 2.0),
- Three sets of items with DIF of different magnitude (0.5; 1.0; 1.5) on two of eight items favouring opposite groups (person mean of 0.00 and a person standard deviation of 2.0).

Each data set comprised 1000 boys and 1000 girls. In Table A1-3 (appendix A) the input values for the item location parameters are shown for all items in all 15 subsets. All items consisted of five response categories and the values for the threshold parameters were the same for all items across all subsets of items [-1.500; -0.500; 0.500; 1.500].

Non uniform DIF referenced to the EVC

The data sets simulated to reveal non-uniform DIF were:

- Two sets of data, differing in item locations, with a person mean of 0.00 and a person standard deviation of 2.0 with non-uniform DIF on one of four items,

- Two sets of data, differing in item locations, with a person mean of 0.00 and a person standard deviation of 2.0 with non-uniform DIF on one of eight items,
- One set of data with a person mean of 0.00 and a person standard deviation of 2.0 with non-uniform DIF on two of eight items,
- One set of data with a person mean of -3.00 (sd 2.0), with non-uniform DIF on one of eight items,
- One set with a person mean of +3.00 (sd 2.0), with non-uniform DIF on one of eight items.

Each data set comprised 1000 boys and 1000 girls. The basic input values for the item location parameters were the same for boys and girls, and identical to those values shown for girls in Tables A1-2. For the items intended to show non-uniform DIF the range of the threshold values was shrunk for one group (boys), making the slope (see appendix B) relatively steeper, and stretched for the other group (girls) making the slope relatively flatter than the slopes of the other items. The following input values were used for the threshold parameters: Boys: -0.5 (t 1), -0.167 (t 2), 0.167 (t 3), 0.5 (t 4); Girls: -3.0 (t 1), -1.0 (t 2), 1.0 (t 3), 3.0 (t 4).

Analysis of DIF

This paper integrates two complementary approaches to studying DIF in data analysed according to the PRM. As stressed in the Introduction, both approaches are referenced to the EVC. The first approach estimates a single EVC for each item and residuals identified by groups are analysed according to the two way analysis of variance (ANOVA). In the analysis one factor has the levels of the gender groups and the other has class intervals on the continuum. In addition to a main gender effect and a main class interval effect, the ANOVA provides evidence of an interaction effect between the class interval and gender. Hence, the ANOVA identifies the general fit of the data to the ICC across the continuum and both uniform and non-uniform DIF (Hagquist & Andrich, 2004). Also as indicated in the Introduction, because estimates are used to construct the residuals, artificial DIF is induced.

The second approach resolves the items identified by the first procedure to show DIF across groups, and estimates a separate EVC for the item for each group. The EVC is identified by its location and slope (see appendix B) and therefore accounts for both uniform and non-uniform DIF. The resolution also helps distinguish between real and artificial DIF and quantifies any real DIF. In addition, as indicated earlier and although not a part of this paper, it is possible that on resolution of an item, it still does not fit the model. In that case DIF would also be non-uniform.

Steps in the analysis of DIF

First, to demonstrate real and artificial DIF all 22 sets of data were analysed based on analysis of responses of persons to items using ANOVA. *Second*, to quantify real DIF in all data sets real DIF items were resolved to create two distinct items, one for each gender.

Similarly, to confirm artificial DIF, in some sets artificial DIF items were resolved into gender specific items. These procedures enable comparisons of the differences in location and slope values for boys and girls. *Third*, these revised item sets were re-analysed using ANOVA. Resolving real DIF provides an estimate of the magnitude of real DIF in the metric of the scale values of the items with no DIF. The difference of the item location values between two groups (e.g. boys and girls) is a measure of the magnitude of uniform DIF for a specific item. The difference of the slope values between two groups (e.g. boys and girls) is a measure of the magnitude of non-uniform DIF for a specific item.

The critical significance level used was 0.5 with Bonferroni adjustment (Bland & Altman, 1995).

The DIF-analyses are reported in Tables as well as using graphs showing EVCs before and after resolution of DIF, along with the mean values of the observed scores for the DIF items as well as the estimates of the item location and slope parameters. In the analysis of non-uniform DIF a comprehensive set of graphs is used to illustrate the impact of the item locations and the person distributions on the magnitude of non-uniform DIF.

Also, the results from the ANOVA are summarized and the estimates of the means of the person parameters reported.

The data were analysed with the software RUMM2030 which uses a pairwise conditional method of estimation for the item parameters in which person parameters are eliminated and the person estimates were obtained using a weighted likelihood method which reduces the bias in the person estimates taking the item parameters as known (Andrich, Sheridan, & Luo, 2014).

Results

Uniform DIF

In Table A4 (appendix A) three sets of items comprising four items with one DIF item and two sets of items comprising eight items with one DIF item are shown. In Figures 1-4 and Tables 2-5 additional four sets of eight items with one DIF item are shown.

Table A4 shows how artificial DIF is induced and affected by the total number of items and the magnitude of the DIF items. With one DIF-item of the magnitude of 0.5 logit in item set 1 with four items artificial DIF is induced in one item, while in item sets 2 and 3 with higher magnitude of DIF additional two items show artificial DIF. In item set 4 with a DIF item of the magnitude of 0.5 logits in a larger set with eight items there are no artificial DIF items induced. With higher magnitude of DIF in the other three sets with eight items and person mean values of 0, two or three items with artificial DIF are induced.

In Table 1 six sets of eight items including two DIF items are shown. Three sets are favouring the same group and three sets are favouring opposite groups.

In item set 10 with two DIF-items of the magnitude of 0.5 logits artificial DIF is not induced in any item, while in item sets 11 and 12 with higher magnitude of DIF three and five items respectively show artificial DIF.

With two DIF-items favouring opposite groups artificial DIF-items are induced working in the opposite direction to the DIF-item with the largest magnitude. With two DIF items of the same magnitude working in the opposite directions the impact of DIF is balanced out.

The results shown in Table 1 demonstrate that the manifestation of artificial DIF is affected by the magnitude of DIF on individual items, the number of DIF items, the direction of DIF, and the total number of items.

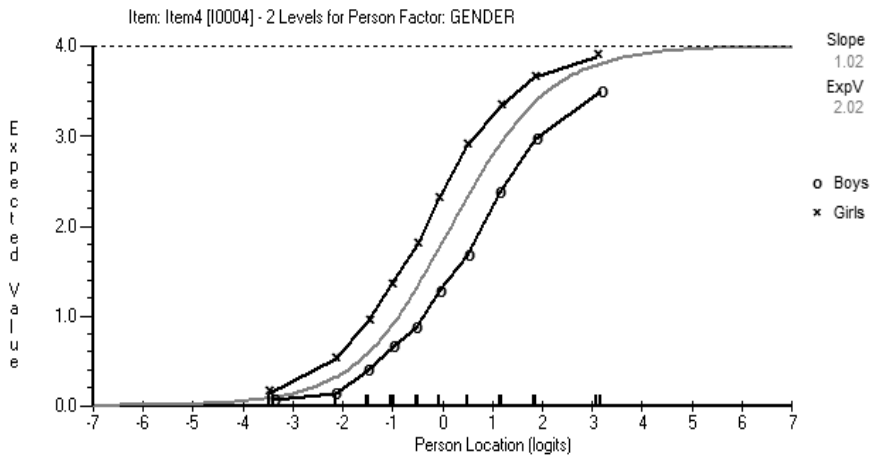
Table 1:

Three sets of eight items with two uniform DIF items favouring the same groups and three sets of eight items with two uniform DIF items favouring opposite groups. Items showing real uniform DIF marked in bold. Mean values of person estimates for boys and girls.

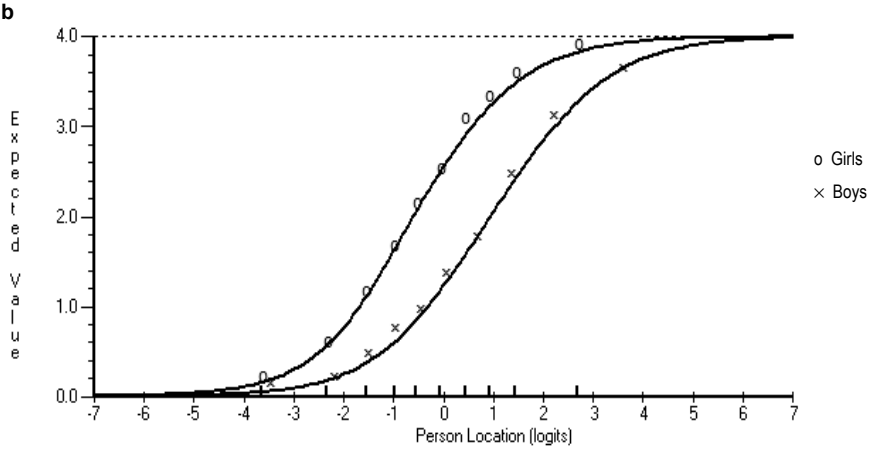
Item set	Number of DIF items	Direction	Magnitude	Total number of items	Items with sign uniform DIF F-values	Person Mean values		
						Boys	Girls	Diff
10	2	G>B	It 4 B +0.5 It 6 B +0.5	8	**It4: 41.77903 G>B** **It6: 20.38926 G>B**	-0.119	-0.059	-0.06
11	2	G>B	It 4 B +1.0 It 6 B +1.0	8	It3: 16.64039 B>G **It4: 132.86080 G>B** It5: 32.41593 B>G **It6: 96.15041 G>B** It7: 19.61691 B>G	-0.244	-0.050	-0.194
12	2	G>B	It 4 B +1.5 It 6 B +1.5	8	It2: 10.82893 B>G It3: 33.74949 B>G **It4: 302.87300 G>B** It5: 61.13791 B>G **It6: 188.65920 G>B** It7: 30.31273 B>G It8: 21.94899 B>G	-0.395	-0.025	-0.37
13	2	G>B B>G	It 4 B +1.5 It 6 G +0.5	8	It3: 17.13523 B>G **It4: 527.72600 G>B** It5: 20.17254 B>G **It6: 127.64710 B>G**	-0.236	-0.098	-0.138
14	2	G>B B>G	It 4 B +1.5 It 6 G +1.0	8	It1: 11.35218 B>G **It4: 561.20210 G>B** It5: 9.58796 B>G **It6: 296.04260 B>G**	-0.220	-0.196	-0.024
15	2	G>B B>G	It 4 B +1.5 It 6 G +1.5	8	**It4: 609.59030 G>B** **It6: 458.94840 B>G**	-0.208	-0.210	0.002

In Figures 1 a-b and Table 2, item 4 with uniform DIF of the magnitude of 1.5 logit is shown, in a set with eight items with a person mean of 0.

1a



1b



Figures 1 a-b:
Item 4 showing uniform DIF before (a) and after resolving DIF (b), in a set with eight items of which one is simulated to reveal uniform DIF (+1.5 logit favouring girls).

Table 2:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of 0 simulated to show uniform DIF for Item 4. Items showing real DIF marked in bold. Mean values after resolving for real DIF are marked in italics. Mean values after resolving artificial DIF items are underlined.

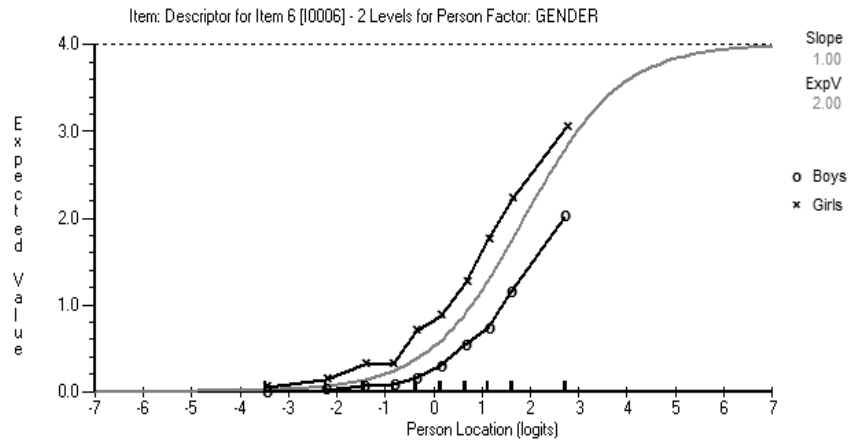
	A. Mean of Observed Scores for item 4 before resolved	**B. Location value for item 4 after resolved**	**C. Slope value for item 4 after resolved**	**D. Person Mean Values**			**E. F-values for items with sign uniform DIF**		
				Original set	Item 4 resolved	Item 5 resolved	Original set	Item 4 res.	Item 5 resolved
Boys	1.37	0.96	0.85	-0.248	*-0.185*	<u>-0.313</u>	It3: 21.24652 B>G	No	It3: 30.89816 B>G
Girls	2.23	-0.54	0.95	-0.049	*-0.196*	<u>-0.060</u>	**It4: 460.60870 G>B**		It4: 413.07740 G>B
Diff	-0.86	1.50	-0.10	-0.199	0.011	-0.253	It5: 32.23714 B>G		It6: 42.86140 B>G
							It6: 30.55280 B>G		It8: 11.71924 B>G

In Figures 2 a-b and Table 3, item 6 with uniform DIF of the magnitude of 1.5 logit is shown, in a set with eight items with a person mean of 0.

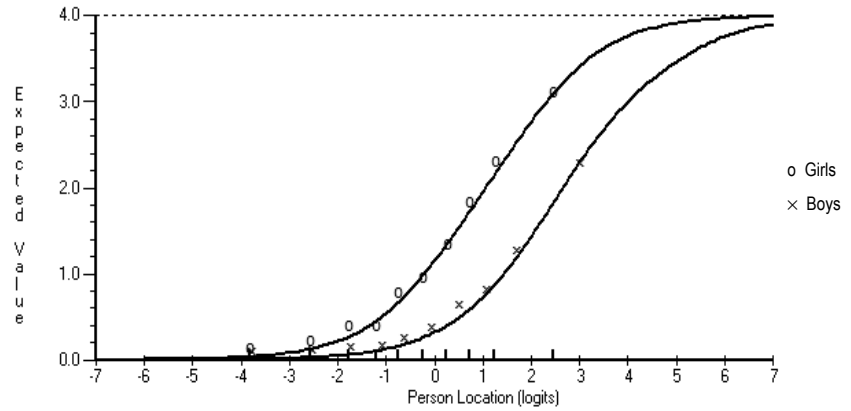
Table 3:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of 0 simulated to show uniform DIF for Item 6. Items showing real DIF marked in bold.

	A. Mean of Observed Scores for item 6 before resolved	**B. Location value for item 6 after resolved**	**C. Slope value for item 6 after resolved**	**D. Person Mean Values**	**E. F-values for items with sign uniform DIF**
				Original set	Original set
Boys	0.53	2.74	0.87	-0.222	It5:11.55437 B>G
Girls	1.20	1.05	0.83	-0.046	**It6: 354.37890 G>B**
Diff	-0.67	1.69	0.04	-0.176	It7:10.60062 B>G

2a



2b



Figures 2 a-b:
Item 6 showing uniform DIF before (a) and after resolving DIF (b), in a set with eight items of which one is simulated to reveal uniform DIF (+1.5 logit favouring girls).

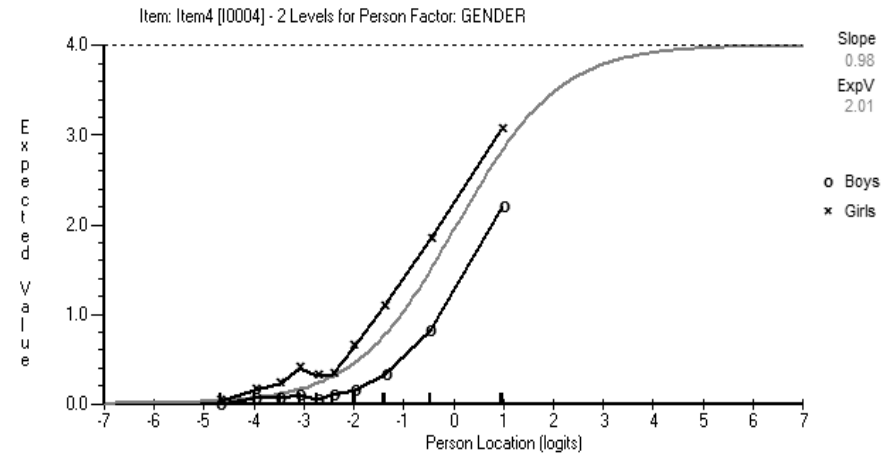
Figures 1 a-b and 2 a-b show that within both items 4 and 6 the expected value curves for boys and girls are parallel with different locations of the item i.e. that for the same values of the variable the difference between the expected value curves is the same along the whole variable. This gender difference is reflected by the mean of the observed scores confirming that item 4 and item 6 favour girls, which is shown in Tables 2 and 3. These Tables also show that the difference in person mean value is bigger in the set with DIF-item 4 than in the set with DIF-item 6, although both items were simulated to have the

same magnitude of DIF. This indicates that the effect of DIF depends on the location of the items relative to the distribution of the persons.

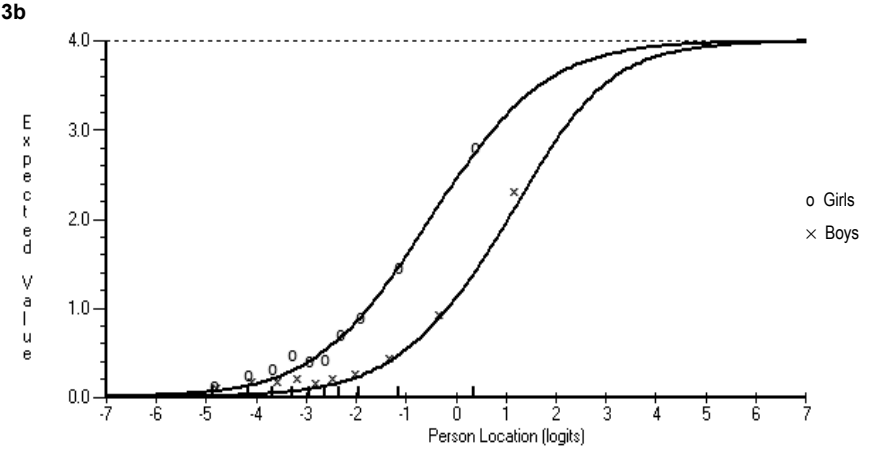
In addition, in Table 2 the effect of resolving real DIF items and artificial DIF items is shown. The resolution of real DIF is manifested at the person level where the difference between boys and girls disappears after resolving DIF. In contrast, Table 2 also clearly shows that resolving artificial girls, thereby confirming that artificial DIF is just an artefact of the procedure for identifying DIF.

In Figures 3 a-b and Table 4, item 4 with uniform DIF of the magnitude of 1.5 logit is shown, in a set with eight items with a person mean of -3.0.

3a



3b



Figures 3 a-b:
Item 4 showing uniform DIF before (a) and after DIF is resolved (b) in a set with eight items of which one is simulated to reveal uniform DIF (+1.5 logit favouring girls), **M -3.0**.

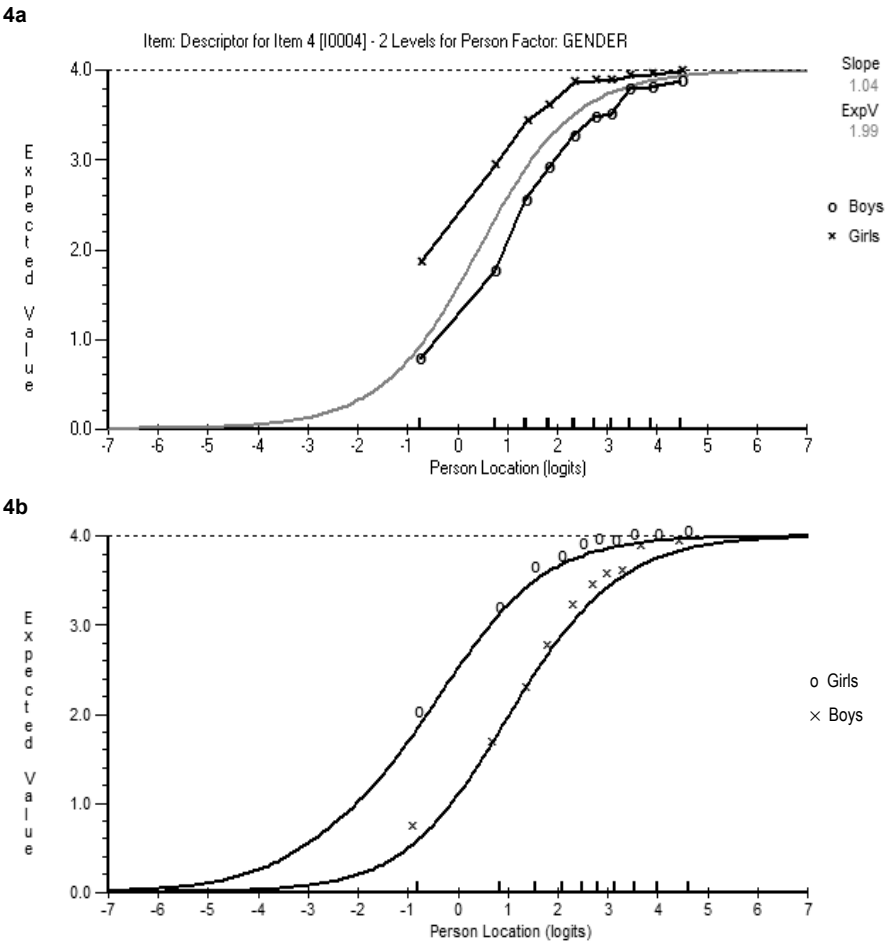
Table 4:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of -3 simulated to show uniform DIF for Item 4. Items showing real DIF marked in bold.

	A. Mean of Observed Scores for item 4 before resolved	B. Location value for item 4 after resolved	C. Slope value for item 4 after resolved	D. Person Mean Values	E. F-values for items with sign uniform DIF
				Original set	Original set
Boys	0.29	0.97	0.93	-3.003	**It4: 196.11040 G>B**
Girls	0.75	-0.52	0.87	-2.887	
Diff	-0.46	1.49	0.06	-0.116	

In Figures 4 a-b and Table 5, item 4 with DIF of the magnitude of 1.5 logit is shown, in a set with eight items with a person mean of +3.0.

Table 5:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of +3 simulated to show uniform DIF for Item 4. Items showing real DIF marked in bold.

	A. Mean of Observed Scores for item 4 before resolved	B. Location value for item 4 after resolved	C. Slope value for item 4 after resolved	D. Person Mean Values	E. F-values for items with sign uniform DIF
				Original set	Original set
Boys	2.93	1.06	0.92	2.633	**It4: 321.08870 G>B**
Girls	3.56	-0.72	0.82	2.783	It6: 10.16088 B>G
Diff	-0.63	1.78	0.10	-0.150	It7: 10.64495 B>G
					It8: 10.15287 B>G



Figures 4 a-b:
Item 4 showing uniform DIF before (a) and after DIF is resolved (b) in a set with eight items of which one is simulated to reveal uniform DIF (+1.5 logit favouring girls), **M +3.0.**

Figures 3 a-b and 4 a-b show that girls are favoured in the person distribution skewed to the left as well as in the person distribution skewed to the right. This is reflected by higher scores on average for girls in both distributions. With the person distribution skewed to the left most of the observed values are located at the lower parts of the curves. With the person distribution skewed to the right most of the observed values are located at the higher parts of the curves.

Tables 4 and 5 show another example of that the effect of DIF depends on the location of the items relative to the distribution of the persons. The difference in person mean value is bigger in the item set skewed to the right than in the item set skewed to the left, although in both items the same item was simulated to have DIF of the same magnitude. Noticeable, for item set skewed to the right as well as the set skewed to the left the difference in person values is less than in the item set with a mean of zero of the person distribution (Table 2).

The conclusion is that the magnitude of artificial DIF is determined by the location of the items showing DIF relative to the distribution of the persons along the variable. This is indicated by no items showing artificial DIF in the item set with a distribution skewed to the left but three artificial DIF items in the distribution skewed to the right.

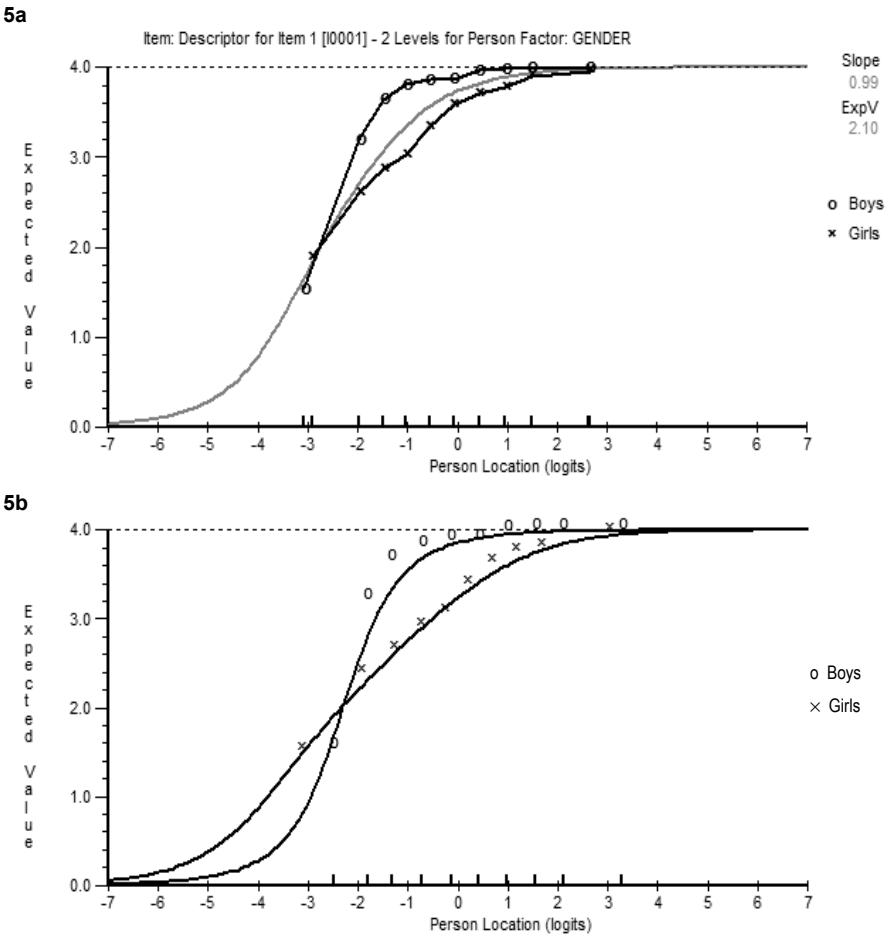
Non-uniform DIF

In Figures 5 and 6 and Tables 6 and 7, single items of different severity but with the same magnitude of non-uniform DIF are shown. The items pertain to two different item sets consisting of four items of which one item is simulated to show non-uniform DIF referenced to the expected value curve.

Figures 5 a-b show a very easy item (1). Most of the observed values are located above the intersection point of the two expected value curves, at the upper parts of the curves where boys are favoured. This is reflected by a higher mean of the observed scores for boys than for girls on that item.

Table 6:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of 0 simulated to show non-uniform DIF for Item 1.

	A. Mean of Observed Scores for item 1 before resolved	**B. Location value for item 1 after resolved**	**C. Slope value for item 1 after resolved**	**D. Person Mean Values**		**E. F-values for items with sign uniform DIF**	**F. F-values for items with sign non-uniform DIF**
				Original set	Item 1 resolved	Original set	Original set
Boys	3.63	-2.30	1.78	-0.025	0.476	IT1: 201.657	IT1: 11.8967
Girls	3.21	-2.14	0.60	-0.182	0.457	IT2: 15.11708	IT2: 7.11439
Diff	0.42	-0.16	1.18	0.157	0.019	IT3: 10.6078	

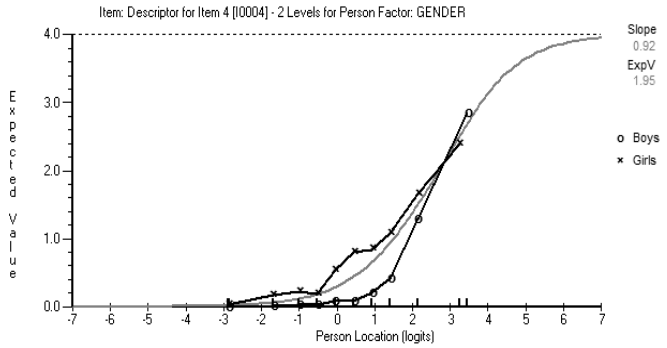


Figures 5 a-b:
Item 1 showing non-uniform DIF, before (a) and after resolving DIF (b), in a set with four original items of which one is simulated to reveal non-uniform DIF.

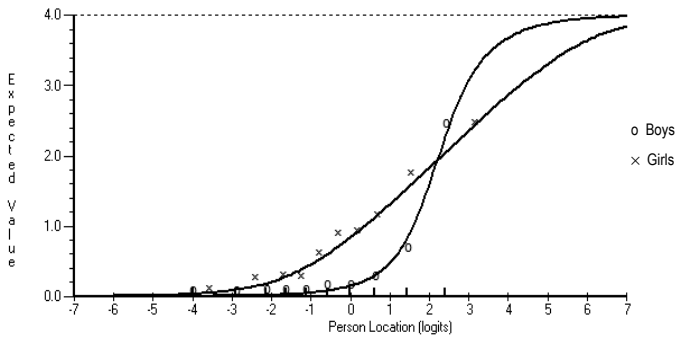
Figures 6 a-b show a very difficult item (4). Most of the observed values are located below the intersection point of the two expected value curves, at the lower parts of the curves where girls are favoured. This is reflected by a higher mean of the observed scores for girls than for boys on that item.

In Figures 7 a-d and Table 8 two items of different severity but with the same magnitude of non-uniform DIF are shown. The items pertain to one item set consisting of eight items.

6a



6b

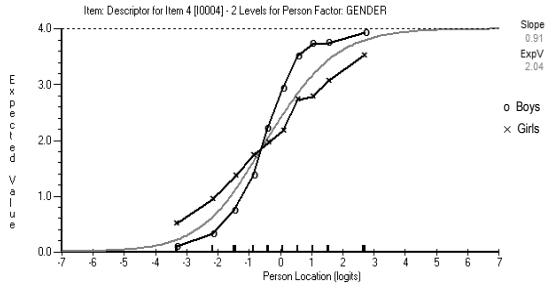


Figures 6 a-b:
Item 4 showing non-uniform DIF before (a) and after resolving DIF (b), in a set with four original items of which one is simulated to reveal non-uniform DIF.

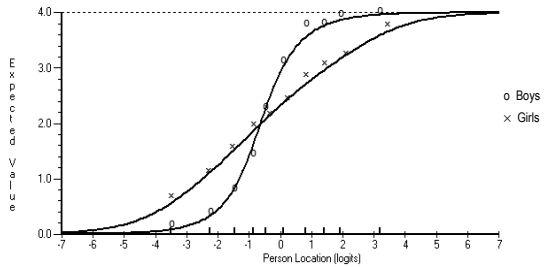
Table 7:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of 0 simulated to show non-uniform DIF for Item 4.

	A. Mean of Observed Scores for item 4 before resolved	**B. Location value for item 4 after resolved**	**C. Slope value for item 4 after resolved**	**D. Person Mean Values**		**E. F-values for items with sign uniform DIF**	**F. F-values for items with sign non-uniform DIF**
				Original set	Item 4 resolved	Original set	Original set
Boys	0.35	2.28	1.70	-0.118	-0.669	IT2: 10.65627	IT3: 5.40131
Girls	0.77	2.32	0.53	0.037	-0.639	IT3: 14.93694	IT4: 9.58124
Diff	-0.42	-0.04	1.17	-0.155	-0.03	IT4: 214.8291	

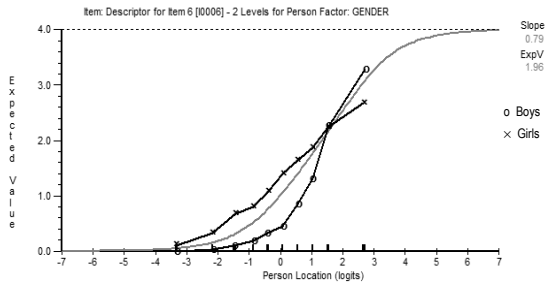
7a



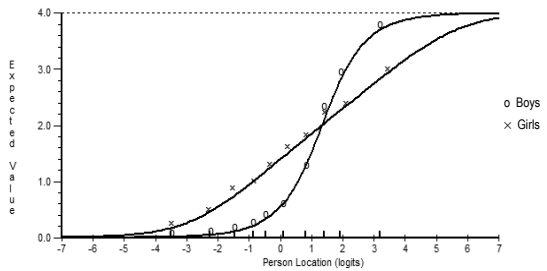
7b



7c



7d



Figures 7 a-d:
Item 4 and 6 showing non-uniform DIF before (a,c) and after DIF is resolved (b,d) in a set with eight original items of which two are simulated to reveal non uniform DIF.

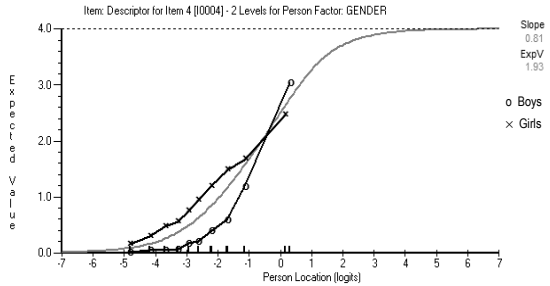
Table 8:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of 0 simulated to show non-uniform DIF for Items 4 and 6.

	A. Mean of Observed Scores for items before resolved		**B. Location value for items after resolved**		**C. Slope value for items after resolved**		**D. Person Mean Values**		**E. F-values for items with sign uniform DIF**	**F. F-values for items with sign non-uniform DIF**
	Items 4	Items 6	Items 4	Items 6	Items 4	Items 6	Original set	Item 4 & 6 resolved	Original set	Original set
Boys	2.30	1.09	-0.62	1.33	1.63	1.45	-0.069	-0.170	IT4: 41.27592	IT4: 45.63831
Girls	2.18	1.38	-0.49	1.36	0.54	0.45	-0.040	-0.167	IT6: 169.8046	IT5: 3.17446
Diff	0.12	-0.29	-0.13	-0.03	1.09	1.00	-0.029	-0.003		IT6: 33.03069

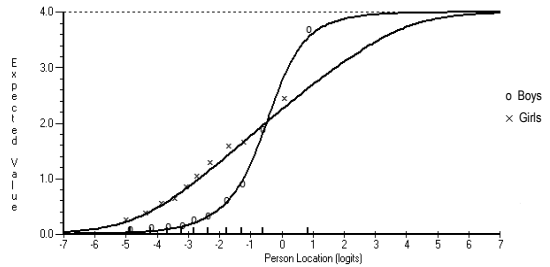
Figures 7 a-d show two non-uniform DIF items of different severity. The easier item 4 is favouring boys while the more severe item 6 is favouring girls. Because most of the individuals are located above the intersection point of the expected value curves for item 4, this item is favouring boys although the gender difference is small as a whole. In contrast the gender difference is bigger for item 6 since most observations are located below the intersection point where girls are favoured.

In Figures 8 a-d and Tables 9 a-b item four with non-uniform DIF is shown in two sets with different person distribution.

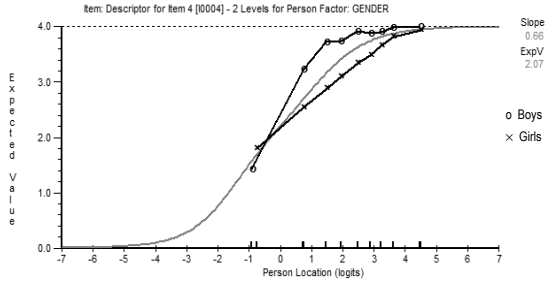
8a



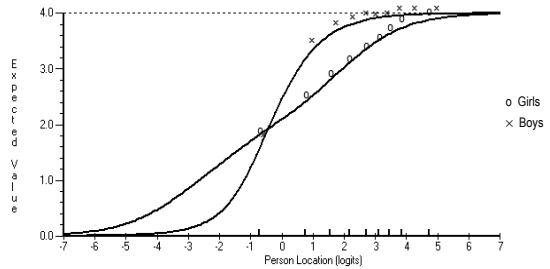
8b



8c



8d



Figures 8 a-d:
Item 4 showing non-uniform DIF before (a,c) and after DIF is resolved (b,d) in a set with eight items of which one is simulated to reveal non-uniform DIF, **Mean -3.0** and **Mean +3.0** respectively.

Table 9a:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of -3.0 simulated to show non-uniform DIF for Item 4.

	A. Mean of Observed Scores for item 4 before resolved	B. Location value for item 4 after resolved	C. Slope value for item 4 after resolved	D. Person Mean Values		E. F-values for items with sign uniform DIF	F. F-values for items with sign non-uniform DIF
				Original set	Item 4 resolved	Original set	Original set
Boys	0.60	-0.54	1.56	-3.091	-3.042	IT1: 13.90869	IT4: 17.40768
Girls	1.04	-0.46	0.48	-2.952	-3.067	IT2: 10.77397	
Diff	-0.44	-0.08	1.08	-0.139	0.025	IT4: 268.7143	

Table 9b:
Item values and person mean values before and after resolution of DIF, and significant real and artificial DIF items in item set with person mean of + 3.0 simulated to show non-uniform DIF for Item 4.

	A. Mean of Observed Scores for item 4 before resolved	B. Location value for item 4 after resolved	C. Slope value for item 4 after resolved	D. Person Mean Values		E. F-values for items with sign uniform DIF	F. F-values for items with sign non-uniform DIF
				Original set	Item 4 resolved	Original set	Original set
Boys	3.62	-0.32	1.26	2.966	3.058	IT4: 212.4304	IT4: 12.94477
Girls	3.24	-0.36	0.39	2.867	3.069		
Diff	0.38	0.04	0.87	0.099	-0.011		

Figures 8 a-d show a person distribution skewed to the left (a,b) where girls are favoured, and a person distribution skewed to the right (c,d) where boys are favoured. This is reflected by higher scores on average for girls and boys respectively.

With the person distribution skewed to the left most of the observed values are located below the intersection point of the two expected value curves, at the lower parts of the curves where girls are favoured. This is reflected by a higher mean of the observed scores for girls than for boys on that item. In contrast, with the person distribution skewed to the right most of the observed values are located above the intersection point

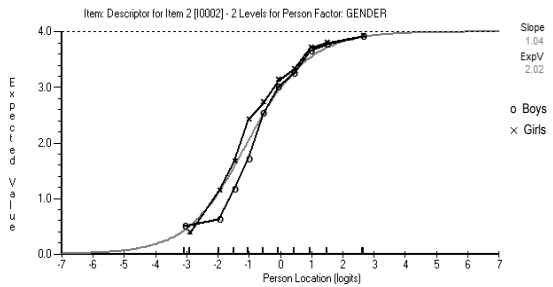
of the two expected value curves, at the higher parts of the curves where boys are favoured.

In Table A5 (appendix A) additional two sets of items comprising eight items with one non-uniform DIF item DIF are shown.

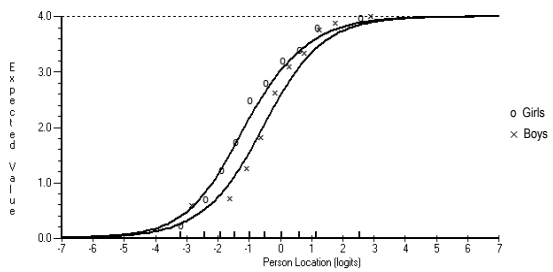
Artificial DIF induced by real non-uniform DIF

In Figures 9 a-f three items with artificial DIF induced by the real DIF Item 1 (fig 5 a-b) in the set with four original items are shown.

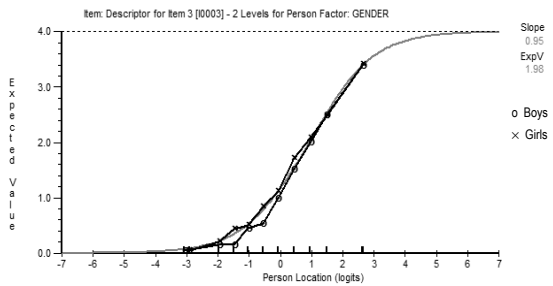
9a



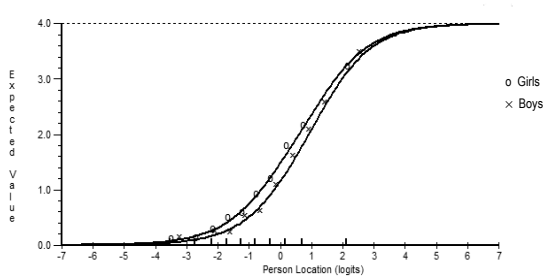
9b



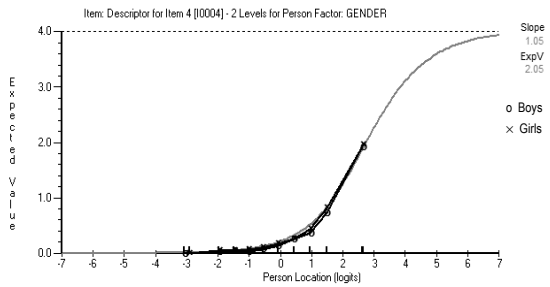
9c



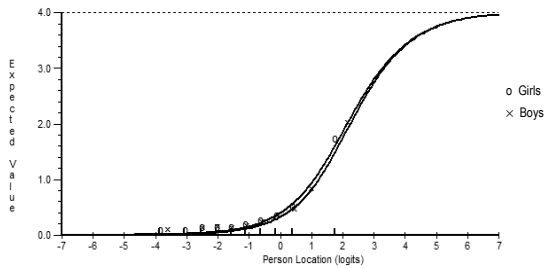
9d



9e



9f



Figures 9 a-f:
Item 2, 3 and 4 showing artificial DIF induced by real non-uniform DIF in item before (a,c,e) and after DIF is resolved (b,d,f) in a set with four original items.

Figures 9 a-f show that real non-uniform DIF in item 1 favouring boys induces artificial DIF in the remaining three items, all favouring girls, predominantly in item 3.

In contrast to Figure 5b showing the real DIF item the expected value curves are not intersecting for the artificial DIF items. The artificial DIF induced by real non-uniform DIF appears mainly as uniform DIF which is reflected by fairly similar slopes but different location of the items for boys and girls.

As a whole the results reported in Figures 5-9 show that the location of the items relative to the person distribution determines the effect on non-uniform DIF on the person level. Very easy and very difficult items located at the ends of the continuum have the biggest impact on the person measures, everything else being equal. Very easy and very difficult items are, however, favouring opposite groups. In a similar way DIF-items in very skewed person distributions have the biggest impact on the person measures, everything else being equal, and distributions skewed to the left and the right are favouring opposite groups.

Discussion

Using simulated polytomous data the present study confirms and illustrates what was proved algebraically in Andrich and Hagquist (2012, 2014) regarding the cause and consequences of artificial DIF: Re the latter, firstly, the greater the magnitude of real DIF of an individual item, the greater the artificial DIF in the other items. Secondly, although the effect of real DIF may be balanced out with respect to person estimates of groups, artificial DIF never balances out real DIF. In addition, the present paper demonstrates the impact of the alignment of the person and item locations on DIF and the effect of non-uniform DIF referenced to the expected value curve.

In both uniform and non-uniform DIF the effect of DIF depends on the location of the items relative to the distribution of the persons which implies that the magnitude of artificial DIF is determined by the location of the items showing DIF, and the distribution of the persons along the variable. Uniquely for non-uniform DIF, also the direction of DIF (e.g. favouring boys or girls) is affected by the location of the items relative to the distribution of the persons.

Given the same magnitude of non-uniform DIF, an item located at one end of the variable favouring one group may be counteracted in the person estimates by an item at the opposite end of the variable favouring another group. This is similar to uniform DIF when two items with real DIF are operating in opposite directions. In non-uniform DIF this cancelling out effect occurs not just between items but also within items. Real non-uniform DIF may also cancel out between two groups because of the item locations relative to the person distribution.

In fact, if the targeting of the person distribution to the items is good, the EVCs for two groups for a single item may intersect in a way that distributes the individuals evenly below and above each other along the variable and thereby makes the non-uniform DIF cancelling out between the two groups. Because of the symmetry inherent in non-uniform DIF the room for artificial DIF may be reduced and the effect on person measurement may be less pronounced than in uniform DIF. As a consequence non-uniform DIF requires less concern than uniform DIF.

In the ANOVA-analysis of residuals from the EVC of an item, real and artificial uniform DIF may appear along with non-uniform DIF. Noticeably, intersecting non-uniform real DIF appears as non-intersecting artificial DIF. Because the means of the thresholds are constrained to be zero in the estimations of the threshold parameters, non-uniform real

DIF will inevitably appear as intersecting non-uniform DIF. In order to appear as non-intersecting non-uniform DIF, different item location values for the two groups are required.

Uniform and non-uniform DIF may also appear simultaneously in the DIF-analysis when the person distribution is highly skewed towards one of the ends of the variable. Because of this skewness, which causes floor or ceiling effects, implying that any random response errors can only appear in one direction providing results which, if interpreted as if the data are not skewed, would be misleading.

Both uniform DIF and non-uniform DIF-items may be resolved resulting in statistical fit of responses to the model. Such fit implies that the sample group specific items discriminate equally at the thresholds. It is stressed that although each of the group specific items fits the Rasch model statistically which is manifested by different item parameters across sample groups for the same item, invariance of item parameters across groups is not retained.

From a validity perspective, resolving DIF may be justified if the source of DIF can be shown to arise from some source irrelevant to the variable of assessment and therefore deemed dispensable; otherwise resolving an item and obtaining statistical fit may affect the validity of the inference from the responses.

Acknowledgement

Previous versions of this paper were presented at the 17th International Objective Measurement Workshop, Philadelphia, USA, April 1-2, 2014 and at the Sixth International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health in Cape Town, South Africa, January 12-14, 2015. The research was also supported in part by funding from the Australian Research Council and industry partners and by Pearson, and by a program grant from the Swedish Research Council for Health, Working Life and Welfare (Forte).

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-574.

Andrich, D. (1988). *Rasch Models for Measurement*. Sage University. Paper on Quantitative Applications in the social Sciences, Series 07-068. Sage Publications, Beverly Hills.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37* (3), 387-416.

Andrich, D., & Hagquist, C. (2014). Real and Artificial Differential Item Functioning in Polytomous Items. *Educational and Psychological Measurement* (Published online 16 May, DOI: 10.1177/0013164414534258)

Andrich, D., & Kline, P. (1981). Within and among population item fit with the simple logistic model. *Educational and Psychological Measurement, 41* (1), 35-48.

Andrich, D., & Sheridan, B., & Luo, G. (2014). *RUMM2030: A Windows interactive program for analysing data with Rasch Unidimensional Models for Measurement*. RUMM Laboratory, Perth, Western Australia.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, *3*, 170.

Guttman, L. (1950). The problem of attitude and opinion measurement. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star & J. A. Clausen (Eds.), *Measurement and Prediction* (pp. 46-59). New York: Wiley.

Hagquist, C., & Andrich, D. (2004). Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences 36*, 955-968.

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

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*, Educational Testing Service. Lawrence Erlbaum Associates.

Osterlind, J. S., & Everson, T. H. (2009). *Differential Item Functioning.* (2nd). Thousand Oaks, CA: SAGE Publications, Inc.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.). *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. IV*, (pp.321-334). Berkeley CA: University of California Press.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurement Transactions 20*(4).

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-54.

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 Psychological Measurement, 17*, 113-114.

Wright, B., & Masters, G. (1982). *Rating scale analysis, Rasch measurement*. Chicago: MESA Press.

Appendix A

Table A1:
Input values for the item location parameters for the sets with four items with five categories – with uniform DIF operating in one direction. Items with real DIF marked in bold.

	Set 1		Set 2		Set 3	
	Boys	Girls	Boys	Girls	Boys	Girls
Items 1	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000
Items 2	**-0.500**	-1.000	**0.000**	-1.000	**0.500**	-1.000
Items 3	1.000	1.000	1.000	1.000	1.000	1.000
Items 4	3.000	3.000	3.000	3.000	3.000	3.000

Table A2:
Input values for the item location parameters for the sets with eight items with five categories – with uniform DIF operating in one direction. Items with real DIF marked in bold.

	Set 4		Set 5		Set 6-8[1]		Set 9		Set 10		Set 11		Set 12	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Items 1	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000
Items 2	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429
Items 3	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857
Items 4	**0.0714**	-0.4286	**0.5714**	-0.4286	**1.0714**	-0.4286	-0.4286	-0.4286	**0.0714**	-0.4286	**0.5714**	-0.4286	**1.0714**	-0.4286
Items 5	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286
Items 6	1.2857	1.2857	1.2857	1.2857	1.2857	1.2857	**2.7857**	1.2857	**1.7857**	1.2857	**2.2857**	1.2857	**2.7857**	1.2857
Items 7	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429
Items 8	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000

[1] Set 6: person mean = 0; set 7: person mean = -3.0; set 8: person mean = +3.0.

Table A3:
Input values for the item location parameters for the sets with eight items with five categories – with uniform DIF favouring opposite groups. Items with real DIF marked in bold.

	Set 13		Set 14		Set 15	
	Boys	Girls	Boys	Girls	Boys	Girls
Items 1	-3.0000	-3.0000	-3.0000	-3.0000	-3.0000	-3.0000
Items 2	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429	-2.1429
Items 3	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857	-1.2857
Items 4	**1.0714**	-0.4286	**1.0714**	-0.4286	**1.0714**	-0.4286
Items 5	0.4286	0.4286	0.4286	0.4286	0.4286	0.4286
Items 6	1.2857	**1.7857**	1.2857	**2.2857**	1.2857	**2.7857**
Items 7	2.1429	2.1429	2.1429	2.1429	2.1429	2.1429
Items 8	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000

Table A4:
Five sets of items with different number of items, different magnitude of uniform DIF and different person distributions. Items with known real uniform DIF marked in bold. Mean values of person estimates for boys and girls.

Item set	Number of DIF items	Person Distribution Mean	Direction	Magnitude	Total number of items	Items with sign uniform DIF F-values	Person Mean values Boys	Girls	Diff
1	1	0	G>B	It2 B +0.5	4	**It2: 60.96416 G>B** It3: 13.90369 B>G	-0.121	-0.036	-0.085
2	1	0	G>B	It2 B +1.0	4	It1: 16.99897 B>G **It2: 171.39940 G>B** It3: 37.11440 B>G It4: 21.17592 B>G	-0.266	-0.054	-0.212
3	1	0	G>B	It2 B +1.5	4	It1: 55.79316 B>G **It2: 388.58730 G>B** It3: 84.88387 B>G It4: 13.35464 B>G	-0.462	-0.056	-0.406
4	1	0	G>B	It 4 B +0.5	8	**IT4: 48.54383 G>B**	-0.063	-0.057	-0.006
5	1	0	G>B	It 4 B +1.0	8	**It4: 217.17420 G>B** It5: 9.60383 B>G It6: 26.94744 B>G	-0.119	-0.059	-0.06

Table A5:

Two sets of items with eight items of which one item is showing significant non-uniform DIF. Location value and slope shown for DIF items resolved by gender. Person mean values reported for boys and girls.

Item set	Total number of Items	DIF Items	Mean	It Split	Location value of items after resolved		Slope of item resolved after resolved		Items with sign non-uniform DIF F-values	Items with sign uniform DIF F-values	Person Mean values		
					B	G	B	G			Boys	Girls	Diff
5	8	IT4	0						IT4: 47.99271	IT4: 23.07458	-0.051	-0.079	0.028
				4.	-0.46	-0.33	1.65	0.55			-0.020	-0.017	-0.003
6	8	IT6	0						IT6: 36.65155	IT6:135.1434	-0.085	-0.029	-0.056
				6.	1.27	1.29	1.44	0.45			-0.229	-0.230	0.001

Appendix B

The slope of the EVC at any location β is given by

$$\frac{\partial E[X_i;\beta]}{\partial \beta} = \frac{\partial \sum_{x=0}^{m_i} xP_{xi}}{\partial \beta} = V[X_i] \quad (9)$$

Incidentally, since $V[X] \geq 0$, the EVC is non-decreasing, and is generally monotonically increasing, irrespective of the values of the item parameters.

The slope varies across the continuum; therefore, the first decision that must be made is where to locate the single value of the slope. There are two obvious options. One is at the location of the item, that is, where $\beta = \delta_i$. The other is the point where the rate of change of the slope of the EVC changes from increasing to decreasing, that is, its point of inflection. If the thresholds are symmetrical about δ_i, then these points are identical; otherwise they are not. Because the point of the paper is to characterise an item by its slope and location, the first choice above is taken, that is when $\beta = \delta_i$. Let ζ_i characterise the slope at $\beta = \delta_i$: then

$$\zeta_i = V[X_i|\beta=\delta_i] \quad (10)$$