Measurement and Research Department Reports 2010-4

A Different View on DIF

Timo M. Bechger
Gunter Maris
Huub H.F.M. Verstralen

Measurement and Research Department Reports 2010-4

A Different View on DIF

Timo M. Bechger, Cito

Gunter Maris, Cito & University of Amsterdam

Huub H.F.M. Verstralen, Cito

Cito
Arnhem, 2010

Abstract

When subjects from non-equivalent groups take the same test, the items may change their characteristics relative to each other giving rise to what is known as *Differential Item Functioning (DIF)*. In this paper, we sketch our view on DIF: i.e., what it is, how it is detected, and how we deal with it when we find it. We focus on DIF in difficulty. Our start point is that DIF in difficulty can only be defined meaningfully in terms of differences in difficulty between items. Thus, to investigate DIF, we compare the relative difficulties estimated in separate calibrations in each group. If there is DIF, it does not imply that *all* relative difficulties are different. Items whose relative difficulties are invariant across groups form *clusters*. Each cluster may be used as an anchor in a concurrent analysis and the set of anchor items or, equivalently, the set of "DIF-items", is not unique. This issue relates to what Camilli (1993) refers to as the ipsative nature of DIF and appears to be well-known among psychometricians What enticed us to bring this issue around is the observation that, when there is DIF, the relation between the latent traits in two groups depends on which set of DIF (or anchor) items one considers. Consequently, when the purpose of the study is to equate scores on two test forms, there may be two or more equating functions that can not be distinguished with the data at hand. An explanation is offered by showing that clusters measure different latent abilities and DIF occurs because the relation between the abilities is different across groups.

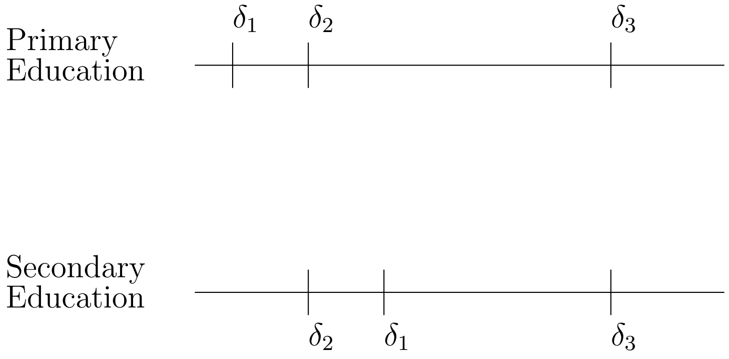


FIGURE 1.

Results of separate analysis. The positions of the items are their difficulties δ_i.

1. Introduction

When subjects from non-equivalent groups take the same test, the items may change their characteristics relative to each other giving rise to what is known as *Differential Item Functioning (DIF)*. When this happens, we have to decide, in consultation with content experts, how to deal with DIF in order to make a meaningful comparison between the groups. This paper is about DIF: i.e., what it is, how it is detected, and how we deal with it when we find it. It is assumed that there are only two groups and there is no DIF with respect to "hidden" groups. When there are more groups, we would compare them pair-wise. Throughout this paper, we assume that Rasch model (Rasch, 1960) holds and focus on DIF in difficulty.

We continue the Introduction with an invented example to provide a concrete context. In Paragraph 1.2 we explain why DIF in difficulty can only be defined meaningfully in terms of differences in difficulty between items. Paragraph 1.3 provides the outline for the rest of the paper.

1.1. Motivating Example

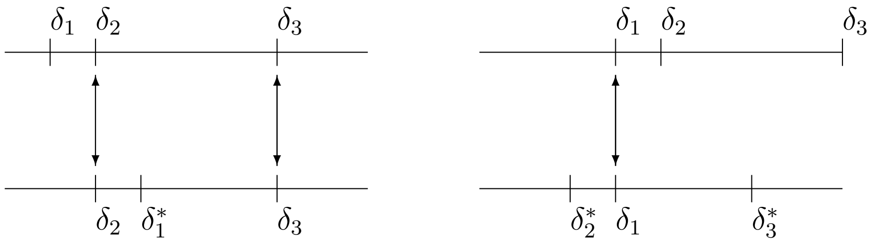
The same test of three items has been taken by two groups of students: students from the final year of primary school, and students from the first year of secondary school. The items are:

1. $365 - 277$
2. $(4 + 2 \times 3) - 7$
3. $\frac{27}{38} - \frac{17}{45}$

We assume that the items follow the Rasch model (Rasch, 1960) in each group. However, separate analyses of the data collected in each group show that the relative difficulties of the items are *not* the same for both groups. Figure 1 shows that items 1 and 2 have changed positions relative to one another, while the relative position of item 2 and item 3 is the same in both groups. Thus, the characteristics of the items depend on the group to which they are administered and we recognize this as a simple case of DIF.

Remark 1. *Note that in practice, one might be inclined to think that DIF is the consequence of a qualitative difference between the abilities of primary and secondary school children. The intuition is that DIF signals multidimensionality. For now, we ignore this interpretation but we get back to it in Section 3.*

Next, we wish to do a concurrent analysis and analyze data from both groups simultaneously. The groups are samples drawn from populations whose ability distributions are *a priori* unknown to us. This means that we need a set of *anchor items* to establish a relation between the two ability scales. Items that show DIF are kept in the analysis but removed from the anchor. They are treated as different items in the two populations.



(a) Choosing the largest DIF-free set (b) Looking at the content of the items

Figure 2.

Two different choices to link the two scales. An asterisk indicates that the item is considered a different item in the secondary education group.

It will be clear that items 1 and 2 (or items 1 and 3) can not be both in the anchor but there is still a number of alternatives to choose from. In consultation with content experts, we consider the following two possibilities, illustrated in Figure 2.

A1: Item 2 and item 3

A2: Item 1.

The first option A1 is based on the argument that DIF items are always a minority, and items 2 and 3 constitute the largest set without DIF. The second option A2 is based on inspection of the curricula of primary and secondary school. The substantive argument is that the first item is the only item that both primary and secondary school children should find easy to solve. According to content experts, item 2 shows DIF because it requires knowledge of precedence rules of operations, which confuses some of the primary school children. Item 3 is difficult for both groups, but secondary school children have less difficulty understanding that the fractions must be converted into like quantities in order to add them.

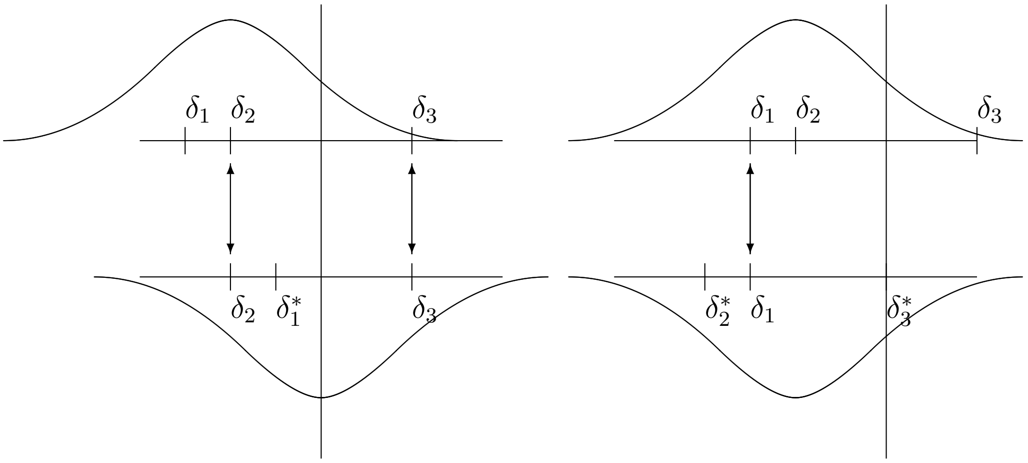
When we look at the output of our computer program, however, both analyses give the same value of the likelihood function and we conclude that the models, although with different sets of DIF items, give the same goodness-of-fit. On second thought, the reason is that differences between difficulties of items within each group are the same in the two models (see Figure 2), and we have in fact specified the same model (see Appendix 6.1).

This outcome worries us for two reasons: First, following e.g., Mellenbergh (1989), Penfield and Camelli (2007, par. 14) or Zumbo (2007) we would like to identify the DIF items, and help content experts to develop theories about the nature of the differences between the populations. Thus, when asked to identify which items show DIF, and which do not, and we are left with two equally good options. The options are "equally good" in the sense that there is no empirical ground to prefer one over the other. Second, the relation between the ability scales depends on which items were selected for the anchor. Figure 3 shows that the average ability in the groups can either be the same or different depending on the anchor used. If the vertical line represents a passing score on the test, half of the pupils in the second group pass in the first analysis, while it is about 10% in the second analysis. It clearly matters which anchor is used. Eventhough, from an empirical point of view we are free to choose any of the anchors, the substantial conclusions are very different. It is this observation that enticed us to prepare this report.

1.2. Separate Calibrations as a Start Point for Inquiry

1.2.1. The Rasch Model

When each of n persons answers each of k items and answers to these items are scored as right or wrong, the result is a two-way factorial design with Bernoulli variates X_{pi}, $p = 1, \ldots, n$, $i = 1, \ldots, k$. Associated with each variate is the probability $\pi_{pi} = P(X_{pi} = 1)$. The Rasch model expresses π_{pi} as a function of a row or *person* effect θ_p and a column or *item* effect δ_i, and assumes that the X_{pi} are independent.


δ_1 δ_2 δ_3
δ_2 δ_1^* δ_3
δ_1 δ_2 δ_3
δ_2^* δ_1 δ_3^*

(a) Choosing the largest DIF-free set

(b) Looking at the content of the items

FIGURE 3.

Two different choices to link the two scales, similar to Figure 2. The vertical line indicates a ability threshold defined on the scale of the first group.

Specifically,

$$P(X_{pi} = x_{pi} | \theta_p, \delta_i) = \frac{\exp[x_{pi}(\theta_p - \delta_i)]}{1 + \exp(\theta_p - \delta_i)} \tag{1}$$

where x_{pi} denotes the observed response. The probability to find the correct responses is an increasing function of θ_p which is taken to represent a person's ability. When δ_i increases, it becomes harder to find the correct response. Hence, δ_i can be interpreted as the difficulty of an item.

The responses of person p will be put into a vector $\mathbf{x}_p = (x_{p1}, \ldots, x_{pk})$ called a *response pattern*. From independence, the data matrix $\mathbf{x}$ is observed with probability

$$P(\mathbf{X} = \mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_p P(\mathbf{X}_p = \mathbf{x}_p | \theta_p, \boldsymbol{\delta}) \tag{2}$$

$$= \prod_p \prod_i \frac{\exp\left[x_{qi}(\theta_q - \delta_i)\right]}{1 + \exp(\theta_q - \delta_i)} \tag{3}$$

$$= \frac{\exp\left(\sum_p x_{p+}\theta_p - \sum_i x_{+i}\delta_i\right)}{\prod_p \prod_i [1 + \exp(\theta_p - \delta_i)]} \tag{4}$$

where $x_{p+} = \sum_i x_{pi}$, and $x_{+i} = \sum_p x_{pi}$. This is the Rasch model, named after

Georg Rasch who presented it in 1960 (Rasch, 1960). The Rasch model will be used throughout this paper. This keeps the mathematics simple while the main arguments extent readily to more complex models. Note that the row and column sums x_{p+} and x_{+i} determine the probability (4). This property of the Rasch model fits well to the ubiquitous use of the number correct score to make decisions about persons.

1.2.2. Identifiability

It is well-known that the Rasch model is not *identifiable*. The value of its parameters is not unique because an arbitrary constant can be added to δ_i and θ_p without changing the probability of a correct answer. That is, $P(X_{pi} = 1|\theta_p, \delta_i) = P(X_{pi} = 1|\theta_p + c, \delta_i + c)$, where c is an arbitrary constant. Consequently, it makes no sense to say: "the difficulty of this item is 2". Only relative difficulties, i.e., differences in difficulty, are uniquely defined.

The item parameter values produced by a program like OPLM (Verhelst, Glas, & Verstralen, 1994) are not to be interpreted as estimates of item difficulty. In fact, each is an estimate of difficulty relative to an aribtrary chosen *point of reference*. Estimated abilities are relative to the same point of reference. Both the estimates and their standard errors may change when a different point of reference is chosen (cf. Verhelst, 1993). Simply because it is a different thing that is estimated.

The point of reference is chosen by imposing a restriction, called a *normalization*. For example, if $\delta_1 = 0$ we obtain difficulties relative to the first item. A normalization is a restriction of the form: $\sum_{i=1}^{k} a_i \delta_i = c$, where c is a constant that may arbitrarily be set to zero. The coefficients a_i are arbitrary except that $\sum_i a_i \neq 0$. A linear combination of the item parameters with $\sum_i a_i = 0$ is called *estimable* because its value is uniquely determined under the Rasch model. It is easily proven that any estimable function can be expressed as a linear combination of relative difficulties $\delta_i - \delta_j$ (cf. Bechger, Verstralen, & Verhelst, 2002, appendix). When we restrict an estimable combination of the item parameters (e.g., $\delta_1 = \delta_2$) we are truely imposing a restriction on the model, i.e., one that can be true or false. In contrast, a normalization merely serves to identify the models: i.e., it cannot be false.

Now, suppose we have persons in two independent groups taking the same test. To allow for the possibility that the same item may have a different difficulty in the groups, we write $\delta_{i,g}$ to denote the difficulty parameter of the ith item in the gth group (see e.g., Davier & Davier, 2007). We may then write the likelihood as:

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\delta}) = \overbrace{\prod_{p=1}^{n_1}\prod_{i=1}^{k}\frac{\exp\left[x_{pi}(\theta_p - \delta_{i,1})\right]}{1+\exp(\theta_p - \delta_{i,1})}}^{\text{group 1}} \overbrace{\prod_{q=n_1+1}^{n}\prod_{i=1}^{k}\frac{\exp\left[x_{qi}(\theta_q - \delta_{i,2})\right]}{1+\exp(\theta_q - \delta_{i,2})}}^{\text{group 2}},$$

where n_1 denotes the number of persons in group 1. It is easily seen that

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{p=1}^{n_1}\prod_{i=1}^{k}\frac{\exp\left[x_{pi}(\theta^*_p - \delta^*_{i,1})\right]}{1+\exp(\theta^*_p - \delta^*_{i,1})} \prod_{q=n_1+1}^{n}\prod_{i=1}^{k}\frac{\exp\left[x_{qi}(\theta^{**}_q - \delta^{**}_{i,2})\right]}{1+\exp(\theta^{**}_q - \delta^{**}_{i,2})},$$

where $\delta^*_{i,1} = \delta_{i,1} - c$, $\theta^*_p = \theta_p - c$, $\delta^{**}_{i,2} = \delta_{i,2} - d$, and $\theta^{**}_q = \theta_q - d$, where c and d are arbitrary constants that need not be equal. That is, we are free to choose a *different* point of reference in each group. It follows that only differences in difficulties within groups are unique. As long as we do not change the relative difficulties of items in each group, we do not change or violate the model. Another way of saying this is that differences between groups are not uniquely determined because we can add a constant to the abilities in a group and add the same constant to the item parameters without changing the model.

1.2.3. Conclusion

With non-equivalent groups, only the relative difficulties of items within each group are defined. What this means is that only differences in relative difficulties across the groups can be detected. Furthermore, we cannot distinguish between a uniform shift in difficulty and a difference in average ability between the groups. Note that this has been known for a long time (e.g., Thissen, Steinberg, & Gerrard, 1986). For example, Scheuneman (1981) writes that

> . . . the mean bias effect across items appears as part of the apparent differences between groups on item difficulty. What remains to be detected is only the variation across items. (pp. 24-25)

Differences in difficulty can be estimated separately with the data from each group. Hence, in contrast to what is sometimes believed, e.g., Millsap (2010), it not necessary to analyse the data from two groups simultaneously in order to test for DIF.

1.3. Outline

In the motivating example, we found that a different Rasch model holds in each group. However, the two concurrent analyses also revealed that the items can be partitioned into two subsets, X and Y, such that the items in each set conform to the same Rasch model in both groups. As illustrated in Figure 4, we found that the Rasch model holds for:

1. $\{X^{(g)}, Y^{(g)}\}$ in each group g.
2. $\{X^{(1)}, X^{(2)}\}$
3. $\{Y^{(1)}, Y^{(2)}\}$

Interestingly, each concurrent analysis proved two of our three findings: i.e., 1 and 2, or 1 and 3. The first analysis established finding number 2 with $X = \{\text{item } 2, \text{item } 3\}$. The second analysis finding number 3 with $Y = \{\text{item } 1\}$. Thus, it proved to be useful to do two concurrent analyses, although we would rather use a model that is consistent with all three findings.

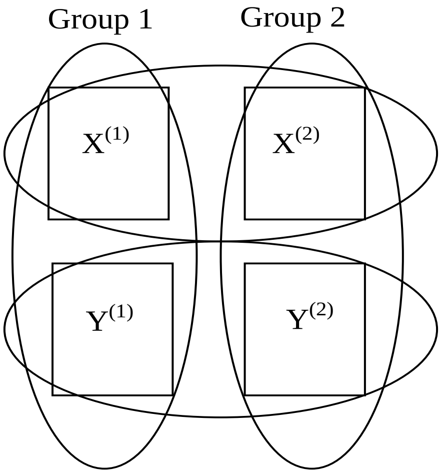


FIGURE 4.

Subsets of items. The ovals indicate which data conform to a Rasch model.

In the sequel, we will take our findings as a starting point. That is, we assume that there is a partition of the items such that our findings are true. We focus on two questions. The first is: How do we find a partition? This is addressed in Section 2. Second, if we have a partition, is there a model that is consistent with all three findings? This is discussed in Section 3. The paper ends with a discussion in Section 4. To keep the discussion focused, some of the technical details are in the Appendix and a separate report (Bechger, Maris, & Verstralen, 2010).

2. Detecting DIF Item-pairs

2.1. A Statistical Test for DIF

While parameter values depend on the arbitrary normalization imposed in each analysis, the relative difficulties do not and are uniquely determined. If we know the item parameters in each group, we can construct two matrices: $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ that contain the differences in difficulties in each of the two groups. That is, $R_{ij}^{(g)} = \delta_{i,g} - \delta_{j,g}$. These matrices contain all information that is available for inferences about

DIF with respect to difficulty: When there is no DIF, $\Delta\mathbf{R} = \mathbf{R}^{(1)} - \mathbf{R}^{(2)} = \mathbf{0}$. For later reference we state this as:

$$\begin{aligned} &H_0 : \Delta\mathbf{R} = \mathbf{0} \quad \text{(No-DIF)} \\ &H_1 : \Delta\mathbf{R} \neq \mathbf{0} \quad \text{(DIF)} \end{aligned}$$

Example 2 (Motivating example continued)**.** *In the motivating example,* $\mathbf{R}^{(1)}$ *and* $\mathbf{R}^{(2)}$ *may have looked like this:*

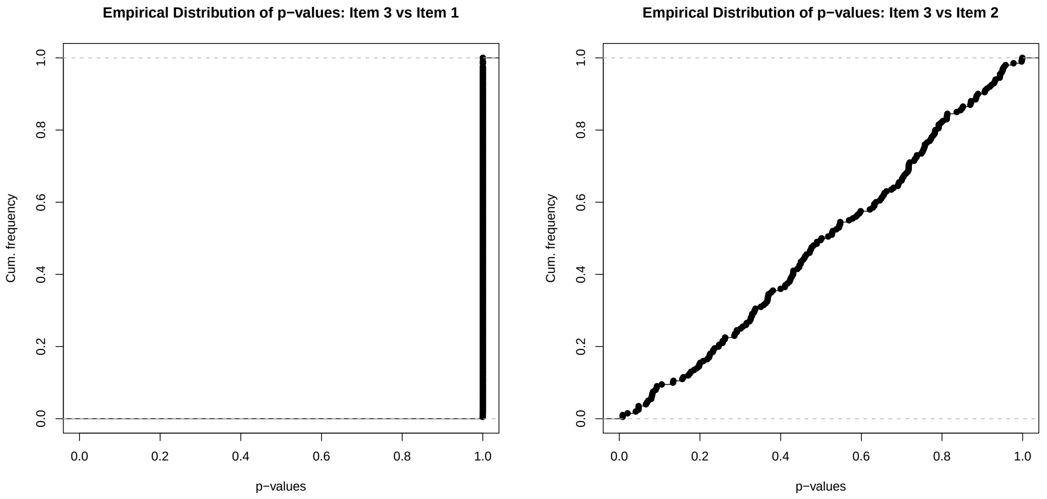
$$\mathbf{R}^{(1)} = \begin{pmatrix} 0 & -0.5 & -1 \\ 0.5 & 0 & -0.5 \\ 1 & 0.5 & 0 \end{pmatrix} \quad \textit{and} \quad \mathbf{R}^{(2)} = \begin{pmatrix} 0 & 1 & 0.5 \\ -1 & 0 & -0.5 \\ -0.5 & 0.5 & 0 \end{pmatrix}. \tag{5}$$

It follows that

$$\Delta\mathbf{R} = \begin{pmatrix} 0 & -1.5 & -1.5 \\ 1.5 & 0 & 0 \\ 1.5 & 0 & 0 \end{pmatrix} \tag{6}$$

and we immediately see that there is DIF. Note that $\Delta\mathbf{R}$ *is, by construction,* skew-symmetric*: i.e.,* $\Delta R_{ij} = -\Delta R_{ji}$. *Note further there are zero as well as non-zero entries in* $\Delta\mathbf{R}$. *The fact that entry (3,2) equals zero means that the relative difficulty of item 2 compared to item 3 is the same in both groups. The non-zero entries in the first row (column) are due to the fact that* δ_1 *has changed its position relative to* δ_2 *and* δ_3.

By construction, the matrices have a very specific structure, and we can repro-



(a) $D_{3,1} > 0$

(b) $D_{3,2} = 0$

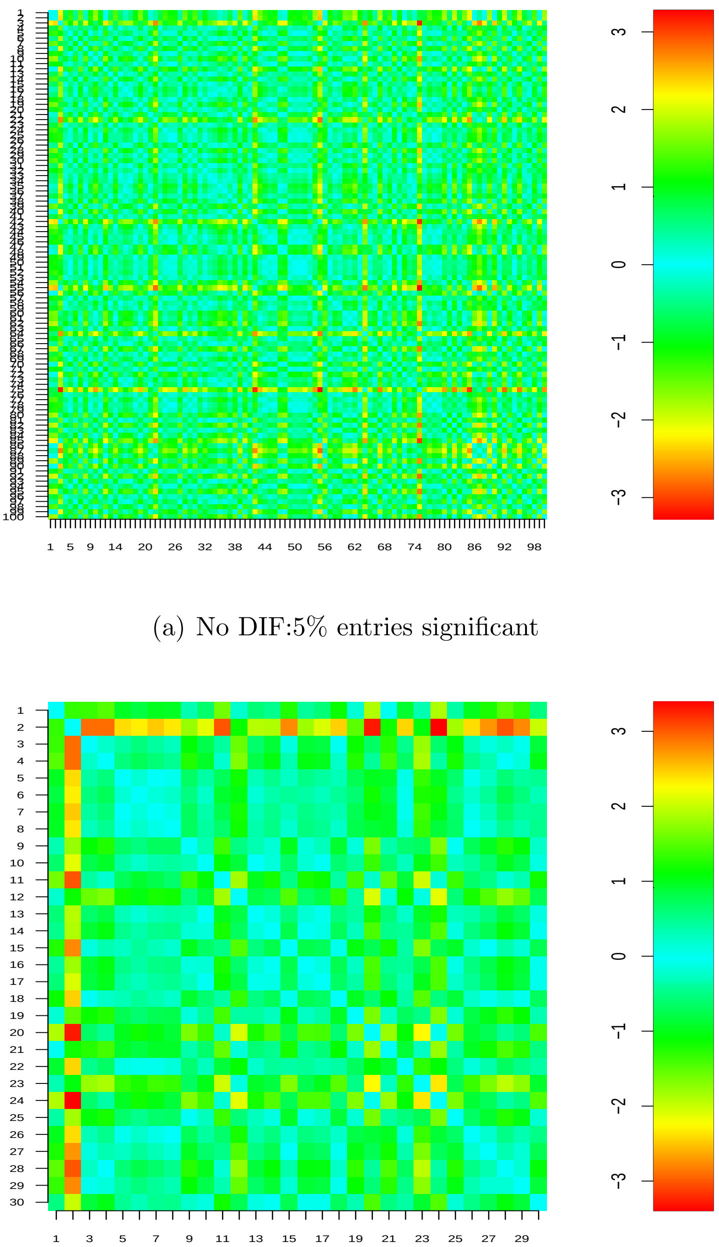
Figure 5.
Empirical distribution of the probability of $\hat{D}_{3,1}$ and $\hat{D}_{3,2}$ under the standard normal distribution based on a simulation. $n = 300$.

duce the whole matrix $\Delta\mathbf{R}$ from the entries in any single row or column. Specifically,

$$\begin{aligned}\Delta R_{ij} &= \delta_{i,1} - \delta_{j,1} - (\delta_{i,2} - \delta_{j,2}) \\ &= \delta_{i,1} - \delta_{r,1} - [\delta_{j,1} - \delta_{r,1}] - (\delta_{i,2} - \delta_{r,2} - [\delta_{j,2} - \delta_{r,2}]) \\ &= \Delta R_{ir} - \Delta R_{jr}\end{aligned}$$

for all rows i and columns $j \neq r$. In fact, we need only know the off-diagonal entries, since the diagonal entries are always zero. Let $\boldsymbol{\beta}^{<r>}$ denote the rth column of $\Delta\mathbf{R}$ but without the zero diagonal entry. Then, $\Delta\mathbf{R} = \mathbf{0} \Leftrightarrow \boldsymbol{\beta}^{<r>} = \mathbf{0}$, for any $r \in \{1, \ldots, k\}$. It turns out that we can take any column we like, and to avoid excessive notation we refer to $\boldsymbol{\beta}^{<r>}$ as $\boldsymbol{\beta}$.

In practice, the item parameters are not known but estimated: $\hat{\boldsymbol{\beta}}$ is calculated with estimates of item parameters and used to make inferences about $\boldsymbol{\beta}$. If the estimators of the item parameters are asymptotically normal, the asymptotic distri-



(a) No DIF:5% entries significant

(b) $\forall i \neq 2$:$R_{2i}^{(2)} \neq R_{2i}^{(1)}$

FIGURE 6.

A colour map of the matrix $\hat{\mathbf{D}}$ based on simulated data ($n = 1000$). Rows and columns are item indices. See Appendix 6.4.

bution of $\hat{\boldsymbol{\beta}}$ is multivariate normal, with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ with entries:

$$\begin{aligned}\Sigma_{i,j} = Cov(\hat{\beta}_i, \hat{\beta}_j) &= Cov\left(\hat{R}^{(1)}_{ir} - \hat{R}^{(2)}_{ir}, \hat{R}^{(1)}_{jr} - \hat{R}^{(2)}_{jr}\right) &(7)\\ &= Cov\left(\hat{R}^{(1)}_{ir}, \hat{R}^{(1)}_{jr}\right) + Cov\left(\hat{R}^{(2)}_{ir}, \hat{R}^{(2)}_{jr}\right), &(8)\end{aligned}$$

where the second equality follows from the fact that the calibrations are based on independent samples. The (co)variances in (7) are obtained from the asymptotic variance-covariance matrix of the item parameter estimates (see Appendix 6.3). Testing for DIF is now easy. Under H_0, the *squared Mahalanobis distance* (SMD: Mahalanobis, 1936)

$$\chi^2_{\Delta\mathbf{R}} \equiv \hat{\boldsymbol{\beta}}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}} \tag{9}$$

follows a chi-squared distribution with $k-1$ degrees of freedom (e.g., Rao, 1973, pp. 418-420). In Appendix 6.2, we show that the index r is arbitrary: i.e., the value of $\chi^2_{\Delta\mathbf{R}}$ is independent of the column of $\Delta\mathbf{R}$ we use. Appendix 6.3 provides a small **R**-script and explains how the test is calculated in practice.

2.2. Clusters

To assess the hypothesis $H_0^{(ij)} : R^{(1)}_{ij} = R^{(2)}_{ij}$ against $H_1^{(ij)} : R^{(1)}_{ij} \neq R^{(2)}_{ij}$, we calculate:

$$\begin{aligned}\hat{D}_{ij} &= \frac{\hat{R}^{(1)}_{ij} - \hat{R}^{(2)}_{ij}}{\sqrt{Var\left(\hat{R}^{(1)}_{ij} - \hat{R}^{(2)}_{ij}\right)}} &(10)\\ &= \frac{\hat{R}^{(1)}_{ij} - \hat{R}^{(2)}_{ij}}{\sqrt{Var\left(\hat{R}^{(1)}_{ij}\right) + Var\left(\hat{R}^{(2)}_{ij}\right)}}. &(11)\end{aligned}$$

The *standardized difference* $\hat{D}_{ij}$ is asymptotically standard normal under $H_0^{(ij)}$. When $R^{(1)}_{ij} \neq R^{(2)}_{ij}$, the mean of $\hat{D}_{ij}$ is non-zero. The *power* of the test depends only on the difference $R^{(1)}_{ij} - R^{(2)}_{ij}$, and on the standard errors which depend on the difficulty of the items relative to the corresponding ability distribution, and the sample size in each group. Furthermore, the marginal distribution of $\hat{D}_{ij}$ is not affected by other entries. This means that each entry can be tested *consistently*, i.e., the result for one item-pair have no bearing on the result for others. A small simulation study confirms that this works. See Figures 5 and 6.

Example 3 (Motivating example continued)**.** *We randomly generate* 200 *responses in each group, with* $\Delta\mathbf{R}$ *as in Equation 6. Ability distributions are normal with* $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 0.2, \sigma_2 = 0.8$. *The item parameters are estimated with conditional maximum likelihood. The* $\chi^2_{\Delta\mathbf{R}} = 20.52$, *which is highly significant with* 2 *degrees of freedom. Hence, we reject the hypothesis that there is no DIF. When we now look at item-pairs, we find*

$$\widehat{\Delta\mathbf{R}} = \begin{pmatrix} 0 & -1.36 & -0.87 \\ 1.36 & 0 & 0.50 \\ 0.87 & -0.50 & 0 \end{pmatrix}, \hat{\mathbf{D}} = \begin{pmatrix} 0 & -4.46 & -2.86 \\ 4.46 & 0 & 1.62 \\ 2.86 & -1.62 & 0 \end{pmatrix} \qquad (12)$$

Compared to $\widehat{\Delta\mathbf{R}}$, $\hat{\mathbf{D}}$ *has the advantage that we interpret the value of the entries: i.e., if its absolute value is smaller than* 2, *an entry is not significantly different from zero. At a glance, it is clear that we may conclude that item 1 has changed its position relative to the others, while the distance between item 2 and item 3 is invariant.*

Example 3 illustrates that items form *clusters*: i.e., subsets of items whose difficulties relative to each other are invariant, and visual inspection of the matrix $\hat{\mathbf{D}}$ may help to spot them. In the absence of DIF, there is one cluster containing all items (e.g., Figure 6.a). When one item has changed its difficulty relative to all other items, there are two clusters: one consisting of one item, and one consisting of all others (see Figure 6.b). On the other extreme, when *all* relative distances are different there are k clusters: each consisting of a single item. Testing whether a particular group of items form a cluster is straightforward. A cluster corresponds to a submatrix of $\Delta\mathbf{R}$ and we simply test whether an arbitrary column in this submatrix is equal to zero. Note, however, that when we wish to do an exploratory analysis and consider multiple clusters we run into the problem of multiple comparisons or *multiple testing.*

It is useful to define a *cluster* as a set of items such that, if one more item is added to the set, the relative difficulties of the items in the set are no longer invariant. A single item i can classify as a cluster in this way when $\delta_{i,1} - \delta_{h,1} \neq \delta_{i,2} - \delta_{h,2}$ for all $h \neq i$. Both anchors, $X = \{1\}$, and $Y = \{2, 3\}$, in the motivating example are such clusters. Defining an anchor in this way guarantees that different anchors are mutually exclusive. Note that we define the clusters with the true parameter values: clusters may not appear so clear-cut with limited data.

2.3. A Real Data Example

To see our test in practice we investigate DIF using real data collected with an IQ-test for children in the age between 11 and 15. For our present purpose, we have taken two samples of (424 and 572) 12-year old children that have taken a sub-scale of the test consisting of 15 items: This subscale was found to fit the Rasch model in each group. The first group has taken the test on the computer, while the second group has taken the paper-and-pencil version of the test.

The items look exactly the same on screen as on paper. Nevertheless, the $\chi^2_{\Delta\mathbf{R}}$ is highly significant ($p < 0.0001$). It appears that the way items are presented has had an effect on relative difficulties. The matrix $\hat{\mathbf{D}}$ is shown in Figure 7. To facilitate visual inspection we have colored its entries. Furthermore, we have permuted its rows and columns such that small entries are clustered (see Appendix 6.4). There

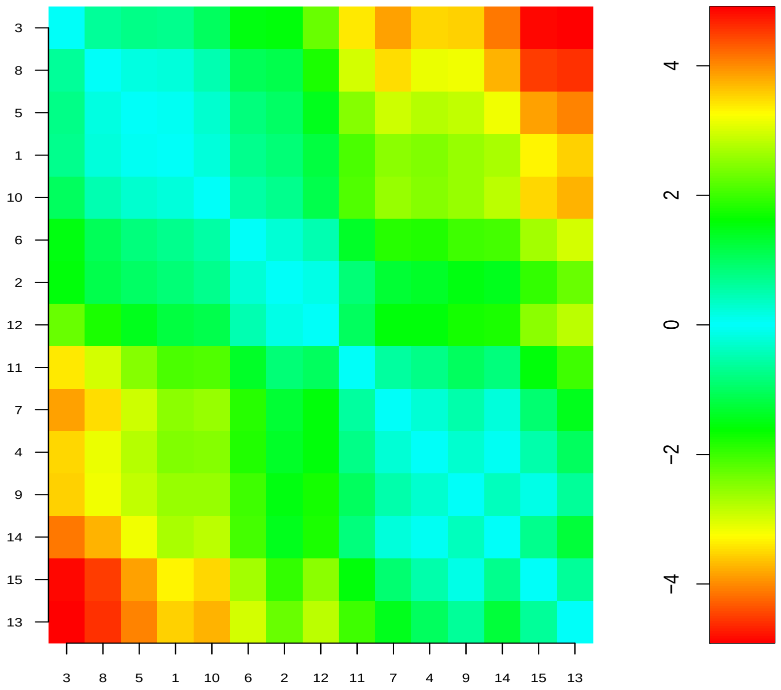


FIGURE 7.

The standardized matrix $\hat{\mathbf{D}}$ for the real data example. Labels of columns and rows are item numbers. About 22 percent of the entries is significant at the 5 percent level.

appear to be three substantial "clusters" of items: $X = \{11, 7, 4, 9, 14, 15, 13\}$, $Y = \{6, 2, 12\}$, and $Z = \{3, 8, 5, 1, 10\}$. There are also items whose difficulties relative to each other have changed. For example, items $\{3, 8, 5\}$ have changed their difficulty relative to items $\{14, 15, 13\}$, which confirms that they belong to different clusters. It is not correct to test an hypothesis on the same data that suggested the hypothesis. For illustrative purposes only, we investigate whether X and Y could be combined in one cluster. Unfortunately, this hypothesis was rejected: $\chi^2_{XY} = 26.11$ which is significant with 9 degrees of freedom ($p = 0.002$). If we remove items 13 and 15 from X, we get $\chi^2_{XY} = 15.71$ with $p = 0.027$.

3. Bringing it all together: DIF and multidimensionality

In the motivating example, one might be inclined to attribute DIF to a qualitative difference between primary and secondary school children. This suggests that a multi-dimensional model is needed, albeit one that is consistent with our findings. That is, the Rasch model must hold for:

Assumption 1 $\{X^{(g)}, Y^{(g)}\}$ in each group g.
Assumption 2 $\{X^{(1)}, X^{(2)}\}$
Assumption 3 $\{Y^{(1)}, Y^{(2)}\}$

where X and Y are clusters of Rasch items that show DIF relative to each other: e.g., $X = \{\text{Item 2, Item 3}\}$, and $Y = \{\text{Item 1}\}$, in the motivating example (see Figure 4). It will be clear how this extends to more than two clusters.

Assumptions 2 and 3 imply that X and Y are Rasch homogeneous scales across the groups. This case is discussed by Glas (1989). The (marginal) likelihood of a person randomly sampled from group g is

$$P_g(\mathbf{x}, \mathbf{y}) = \int \int P(\mathbf{x}|\theta, \boldsymbol{\delta}_x) P(\mathbf{y}|\beta, \boldsymbol{\delta}_y) f_g(\theta, \beta) d\beta d\theta, \qquad (13)$$

where $P(\mathbf{x}|\theta, \boldsymbol{\delta}_x)$ and $P(\mathbf{y}|\beta, \boldsymbol{\delta}_y)$ conform to the Rasch model, and only the distribution of abilities differs between the groups: where θ is the ability measured by the items in X, and β the ability measured by the items in Y (Glas, 1989, Eq. 7.1.2).

Assumption 1 holds when

$$\beta = \theta + b_g, \qquad (14)$$

This means that the latent variables are perfectly correlated and the marginal likelihood is:

$$P_g(\mathbf{x}, \mathbf{y}) = \int P(\mathbf{x}|\theta, \boldsymbol{\delta}_x) P(\mathbf{y}|\theta + b_g, \boldsymbol{\delta}_y) f_g(\theta) d\theta \tag{15}$$

$$= \int P(\mathbf{x}|\theta, \boldsymbol{\delta}_x) P(\mathbf{y}|\theta, \boldsymbol{\delta}_y + b_g) f_g(\theta) d\theta, \tag{16}$$

where the second equality holds because a shift in ability is equivalent to a shift in difficulty. When $b_1 \neq b_2$, the position of the items in X relative to the items in Y is different in the two groups. Hence, the latent variables are perfectly correlated, but their relation is different in the two groups and this gives rise to DIF: In this case, DIF in difficulty. Furthermore, we can identify the items measuring each latent variable.

It follows that a model where all three assumptions hold arises as a special case of a general multidimensional model, where

$$\beta = \theta + b_g + \sigma_g \epsilon, \tag{17}$$

and ϵ denotes random error with mean zero and variance one. Specifically, we obtain (14) from (17) when $\sigma_1 = \sigma_2 = 0$. Thus, when $\sigma_g = 0$, and we have Rasch homogeneous subsets X and Y, DIF can be investigated by testing whether $b_1 = b_2$.

4. Discussion

We have developed a statistical test for DIF based on relative item difficulty. The start point has been that DIF is not a property of individual items but of pairs of items. Because we look at item-pairs directly, our test is quite different from traditional procedures to identify "DIF items". To appreciate what this means, let us compare $\hat{D}_{ij}$, which was introduced earlier, to

$$\chi_i^2 = \frac{\hat{\delta}_{i,1} - \hat{\delta}_{i,2}}{\sqrt{Var\left(\hat{\delta}_{i,1} - \hat{\delta}_{i,2}\right)}} \tag{18}$$

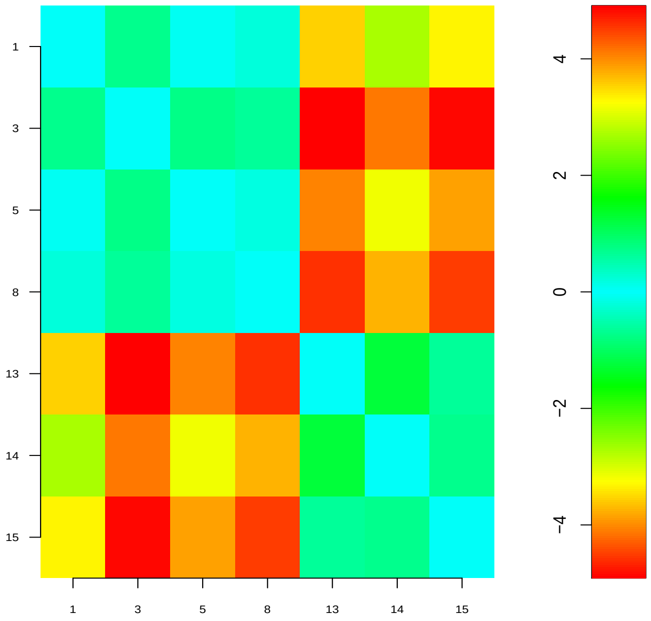


Figure 8.

A submatrix of $\hat{\mathbf{D}}$ for the real data example corresponding to items flagged as "DIF items" by a likelihood ratio test with purification. Labels of columns and rows are item numbers.

which was proposed by Lord (1980, p. 219) as a statistic to test the hypothesis that $\delta_{i,1} = \delta_{i,2}$. The similarity between the two statistics becomes apparent when we realize that $\hat{\delta}_i$ is really the estimated difficulty of item i, *relative to some arbitrary point of reference.* The notation is somewhat misleading. If, for example, the rth item is used as a anchor, we should write

$$\chi_i^2 = \frac{\widehat{\delta_{i,1} - \delta_{r,1}} - \left(\widehat{\delta_{i,2} - \delta_{r,2}}\right)}{\sqrt{Var\left(\widehat{\delta_{i,1} - \delta_{r,1}} - \left(\widehat{\delta_{i,2} - \delta_{r,2}}\right)\right)}}, \tag{19}$$

and we see that $\chi_i^2 = \hat{D}_{ir}$. It follows that, when χ_i^2 is used as a item-DIF statistic, it depends on the normalization in each calibration whether or not an item is flagged as a DIF-item: e.g., item r never has any DIF. The effect is clearly seen in the motivating example. There, if we take $r = 1$, χ_2^2 and χ_3^2 are significant even though

$D_{23} = 0$. If $r = 2$, item 1 is flagged.

While our procedure gives consistent tests for different subsets of items, many of the traditional tests are inconsistent due to a abuse of the Neyman-Pearson approach to hypothesis testing (Neyman & Pearson, 1933). Consider the following common procedure. We do two concurrent analyses. In the first analysis, all items figure in the anchor. In the second analysis, one item is removed from the anchor: i.e., the item is entered in the analysis as a different item for the second population. The null-hypothesis is $H_0 : \boldsymbol{\delta}^{(1)} = \boldsymbol{\delta}^{(2)}$. The alternative hypothesis is:

$$H_{1i} : \begin{cases} \delta_i^{(1)} \neq \delta_i^{(2)} \\ \delta_j^{(1)} = \delta_j^{(2)}, \quad (j \neq i) \end{cases}$$

A likelihood ratio is calculated to determine whether H_0 should be abandoned in favor of the alternative hypothesis H_{1i}. When this happens, it is concluded that item i shows DIF (e.g., Glas, 1998, p. 655). However, the alternative hypothesis is *not* $\delta_i^{(1)} \neq \delta_i^{(2)}$, but includes the condition that the remaining (anchor) items form a cluster. If this condition doesn't hold, neither H_0 nor H_{1i} is true and it is difficult to say what we expect for the distribution of our test statistic. Furthermore, tests for different items are no longer consistent which leads to a tendency to conclude that all items show DIF. This is easily confirmed by simulation (e.g., Finch & French, 2008; Stark, Chernyshenko, & Drasgow, 2006). Note that this observation is not new and has led researchers to suggest alternative *anchoring strategies* (e.g., De Boeck, 2008, §7.4; Wang & Yeh, 2003; Wang, 2004; Stark et al., 2006), and *heuristics* for iterative *purification* of the anchor (Lord, 1980, p. 220; Flier, Mellenbergh, Adèr, & Wijn, 1984; Candell & Drasgow, 1988; Penfield & Camelli, 2007, pp. 161-162). Lord (1980) employed one such anchoring strategy, now called the *free-baseline method.* It entails choosing a "referent" item and examining $\chi_i^2 = \hat{D}_{ir}$ for each item $i \neq r$. This provides a correct way to test the hypothesis that $\delta_{i,1} - \delta_{r,1} = \delta_{i,2} - \delta_{r,2}$.

The conceptual problem, however, remains. As an illustration, we have applied the *likelihood ration test with purification* (Thissen, 2001) to the real data analyzed earlier, and Figure 8 shows $\hat{\mathbf{D}}$ for the flagged items. It is seen that many of the "DIF items" have invariant relative difficulties. We find it hard to assess whether this provides a meaningful summary of the data.

The focus on item-pairs has naturally led to the observation that items form clusters: subsets of items whose difficulties relative to each other are invariant across groups. The real data example confirmed that there can be various clusters that are substantial in size. In the case considered here, each cluster constitutes a Rasch homogeneous scale across groups. While each cluster can be used as an anchor for concurrent analysis, the substantive outcomes will not be the same for each cluster. We have found that this can be explained in a multidimensional Rasch model: Finding DIF means that the test is multidimensional and the clusters are unidimensional subtests measuring different abilities. In future research we hope to develop practical ways to *estimate* the clusters: i.e., to determine which of the possible partitionings of the items is best supported by the data. Note that, if there is DIF, this procedure would produce multiple anchors, while current procedures to detect DIF-items produce one and only one set of DIF items. This means that there must be an implicit rule to choose between multiple empirically equivalent sets. What ever rule is applied, we suggest that it be made explicit such that users can decide whether they agree with it. The *posterior anchoring* advocated by De Boeck (2008), for instance, explicitly incorporates the rule that DIF items are a minority (p. 556). This means that items that show DIF in a test where they are a minority, may not show DIF in another tests where they constitute a majority.

Note that the definition of DIF used in this report, i.e., a difference across groups in the properties of items relative to each other, is different from the commonly accepted one: i.e., an item shows DIF when people from different groups of same underlying true ability have a different probability to give a certain response. As an illustration, we quote the definitions of DIF, or item bias as it was called at the time, given by two influential authors: First, Lord (1980) provides the following definition

> If each test item in a test had the same item response function in every group, then people of the same ability or skill would have exactly the same change of getting the item right, regardless of their group membership. Such a test would be completely unbiased. If on the other hand, an item has a different item response function for one group than for another, it is clear that the item is biased. (p. 212)

Second, Mellenbergh (1989) defines DIF, rather general, as:

> An item is unbiased with respect to the variable G and given the variable Z if and only if
>
> $$f(X|g,z) = f(X|z)$$
>
> for all values g and z of the variables G and Z, where $f(X|g,z)$ is the distribution of the item response given g and z and $f(X|z)$ the distribution of the item responses given z; otherwise the item is biased. (p. 129)

Both definitions start with a definition of no DIF, and DIF is defined as the complement. We agree on the definition of no DIF which, in this case, simply states that the Rasch model holds. However, from our point of view, the definition of DIF is incomplete because it doesn't say which anchor is chosen to define the common ability (Z in the definition of Mellenbergh). An item shows DIF with respect to a certain anchor, and researchers using different anchors will define DIF with respect to a different ability. If the anchor is explicitly chosen (e.g., the mean difficulty in each group) there is still the issue that different anchors are equally good from an empirical point of view. This was illustrated by the motivating example, where item 1 could be a DIF item or an anchor item.

Our results have implications in the following contexts:

1. *Examinations:* IRT-based equating is done to determine a score on a new test that corresponds to an established passing score on an existing test (Holland, Dorans, & Petersen, 2007). In the presence of DIF, researchers will carefully choose an anchor to equate scores from old to new. However, there will be more than one possible anchor. Furthermore, as illustrated in the motivating example, researchers working with different anchors may find different passing scores and there is no empirical basis to prefer a particular passing score over another. In view of the importance of passing scores for people taking the test, this constitutes a serious issue. However, DIF means that the test is multidimensional. Thus, we should, in retrospect, conclude that the test is measuring other latent traits besides the intended one. One should then decide how the scores on different subtests may be combined to reach a pass-fail decision. Clusters corresponding to traits that are considered irrelevant may be removed.
2. *Population Inference:* Suppose that in the motivating example, we would like to make inferences about the populations taking the test. By choosing different anchors we shift the scale in one population relative to the other which changes our inferences regarding the ability distributions in the populations (see Figure 3). Cross-national comparative studies like PISA,TIMMS, PIRLS, etc. have to deal with substantial amounts of DIF. As a rule, DIF items are identified before the main study, as part of a general effort to ensure that tests are comparable across groups (e.g., Hambleton, Merenda, & Spielberger, 2005). The DIF items are then removed and population inferences are based on an analysis without these

items (e.g., Guideline 12 in Vijver & Hambleton, 1996). However, and especially when there are two or more substantial clusters, the league tables will differ depending on the choices made. As with examinations, recognizing that DIF implies multidimensionality provides a way-out. It seems opportune not to remove items and report ability differences on the scales defined by the clusters.

3. *Student Tracking:* Student tracking systems relate scores on multiple test levels to a developmental scale that can be used to assess student growth over a range of educational levels (e.g., Kolen, 2006). Abilities may change both quantitatively and qualitatively as students progress. Hence, depending on the range of levels covered, the need for a multidimensional model is evident. When a unidimensional model is used, qualitative changes will show up as DIF. Conclusions concerning growth then depend on how the scale is anchored across levels, as demonstrated in practice by Verhelst (2010). A solution would be to use a multidimensional model, and this was suggested earlier by Andersen (1985). The model may prove difficult to interpret but once we have the multivariate distributions under control, we may use the model for *vertical prediction*. That is, foretell how a student will score on a test at a next level.

Thus, DIF affects much of what psychometricians do in practice. In this paper we have sketched our view on DIF. It will be clear that further study is needed to: a) work out the consequences in more detail, b) generalize our approach to more complex IRT models, c) develop efficient ways to apply the test in practice, and d) to establish more firmly the agreements and disagreements between our approach and existing work: Especially, the earlier work relating DIF to multidimensionality (e.g., Kok, 1988; Shealy & Stout, 1993; Nandakumar, 1993).

Finally, we think that many psychometricians share our points of view. Nevertheless, the consequences have not been put into practice. The indeterminancy of the anchor, for example, relates to what Camilli (1993) refers to as the ipsative or circular nature of DIF (e.g., Angoff, 1982; Camilli, 1993, pp 408-411; Thissen, Steinberg, & Wainer, 1993, p. 103; Williams, 1997; Penfield & Camelli, 2007, pp. 161-162) and appears to be well-known among psychometricians (e.g., Shealy & Stout, 1993, p. 207; De Boeck, 2008, p. 551; Soares, Concalves, & Gamerman, 2009, pp. 354-355; Verhelst & Glas, 1995, p. 91, or Wang, 2004). Penfield and Camelli (2007), for instance, recognize that ipsativity is a problem but continue to focus on the identification of a single set of DIF items.

5. Acknowledgement

The authors wish to thank Norman Verhelst, Cees Glas and Paul De Boeck for helpful comments and discussion on an early version of this manuscript. We thank Han van der Maas for his invitation to give a talk. Not in the least because it provided us with the opportunity to discuss the content of this paper with Don Mellenbergh whose influence on DIF research cannot be overrated.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3-16.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting item bias* (p. 96-116). Baltimore: John Hopkins University Press.

Bechger, T. M., Maris, G., & Verstralen, H. H. F. M. (2010). *Equivalence of lltms depends on the design* (R&D Report No. 2010-4). Arnhem: Cito.

Bechger, T. M., Verstralen, H. H. F. M., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, *67*, 123-136.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do test bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 397-413). Hillsdale, NJ: Lawrence Earlbaum.

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.

Davier, M., von, & Davier, A. A., von. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology*, *3*(3), 1614-1881.

De Boeck, P. (2008). Random IRT models. *Psychometrika*, *73*(4), 533-559.

Finch, H. W., & French, B. F. (2008). Anomalous tupe I error rates for identifying one type of differential item functioning in the pressence of the other. *Educational and Psychological Measurement*, *68*(5), 742-759.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, p. 515-585). Amsterdam, The Netherlands: Elsevier.

Flier, H. Van der, Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative bias detection method. *Journal of Educational Measurement*, *21*, 131-145.

Glas, C. A. W. (1989). *Contributions to estimating and testing rasch models.* Unpublished doctoral dissertation, Arnhem: Cito.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*, 647-667.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-culturar assessment.* Lawrence Erlbaum Associates.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, p. 169-204). Amsterdam: Elsevier.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (p. 263-275). Plenum.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., chap. 5). American Council on Education and Praeger Publishers.

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

Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India* (Vol. 2, p. 49-55).

Mair, P., & Hartzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*, 1-18.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Millsap, R. E. (2010). *Personal communication.* Decision letter manuscript PMET-

231. (Used with permission)

Nandakumar, R. (1993). Simultaneous dif amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *30*, 293-311.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philisophical Transactions of the Royal Society of London: Series A*, *231*, 289-337.

Penfield, R. D., & Camelli, G. (2007). Differential item functioning and bias. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26). Amsterdam: Elsevier.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New-York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago, The University of Chicago Press)

Scheuneman, J. D. (1981). A new look at bias in aptitude tests. In P. Merrifield (Ed.), *New directions for testing and measurement: Measuring human abilities.* San-Fransisco: Jossey-Bass.

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

Soares, T. M., Concalves, F. B., & Gamerman, D. (2009). An integrated bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*(3), 348-377.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Towards a unified strategy. *Journal of Applied Psychology*, *19*, 1292-1306.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the comutation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Documentation for computer program [Computer software manual]. Chapel Hill.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67-114). Hillsdale, NJ: Lawrence Earlbaum.

Verhelst, N. D. (1993). *On the standard errors of parameter estimators in the Rasch model* (Measurement and Research Department Reports No. 93-1). Arnhem: Cito.

Verhelst, N. D. (2010). *A simple model and its practical consequences: A tinge of pessimism.* (Paper presented at the 2010 RCEC workshop in Twente.)

Verhelst, N. D., & Glas, C. A. W. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 69-95). New-York: Spinger.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1994). OPLM: Computer program and manual [Computer software manual]. Arnhem.

Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *12*, 89-99.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of rasch models. *The Journal of Experimental Education*, *72*(3), 221-261.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning whith the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479-498.

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education*, *10*(3), 253-267.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223-233.

6. Appendix

6.1. Equivalence of the Models in the motivating example

In the example, we stated that analyses with two different anchors lead to the same likelihood and the models are observationally equivalent models. At first sight, however, the model *cannot* be equivalent because they differ in the number of parameters. To wit, four parameters in the first analysis, and five in the second analysis (i.e., $\delta_1, \delta_2, \delta_3, \delta_2^*, \delta_3^*$). Consequently, one expects the second analysis to furnish a better fit to the data. This requires clarification.

On closer look, the difference between the two analyses is that, in the second analysis, there is no restriction that $\delta_1 - \delta_2 = \delta_2^* - \delta_3^*$. Thus, the second model is nested in the first and it is only *after* we impose the restriction that the two models become observationally equivalent. If we do *not* impose the restriction, it is possible to find that $\hat{\delta}_1 - \hat{\delta}_2 \neq \hat{\delta}_2^* - \hat{\delta}_3^*$ in which case one would decide in favour of the second model. When, however, $\delta_1 - \delta_2 = \delta_2^* - \delta_3^*$, this will become increasingly unlikely when sample size increases. To illustrate this, we have conducted a small simulation study. For different sample sizes, n, we generated 200 data sets and analyze each using *conditional maximum likelihood.* If we gather the conditional likelihoods for each model we find that the likelihood for the second analysis tends to be higher. However, as seen in Figure 9, the ratio of the likelihoods converges to one when sample size increases.

That the models are equivalent may proven more formally by casting the models for concurrent analysis in the framework of the classical *Linear Logistic Test Model* (LLTM: see Fischer, 2007, par. 7, and references therein). A detailed proof is described elsewhere (Bechger et al., 2010).

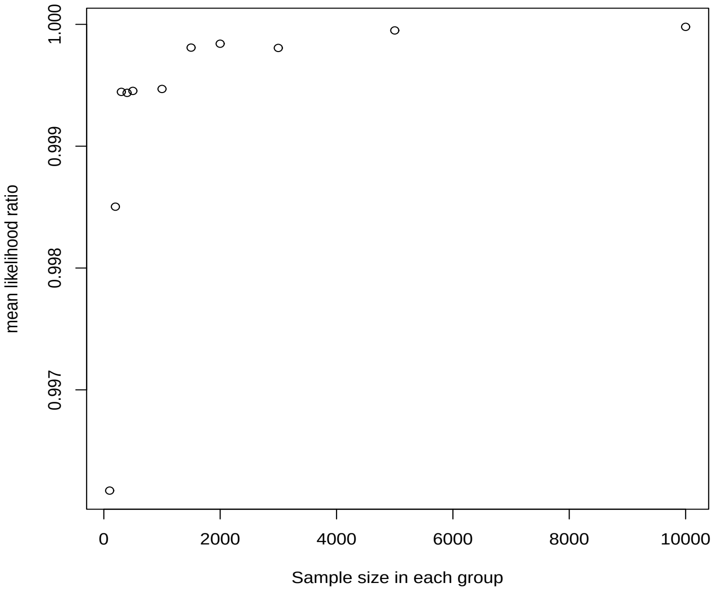


FIGURE 9.

Average Likelihood Ratio of the two concurrent analyses in the motivating example.

6.2. $\chi^2_{\Delta\mathbf{R}}$ is independent of the column

Let $\boldsymbol{\beta}^{<r>}$ denote the rth column in $\mathbf{R}^{(1)} - \mathbf{R}^{(2)}$, without the diagonal entry, which we know to be zero. As mentioned in the text, we can reconstruct the whole matrix $\mathbf{R}^{(1)} - \mathbf{R}^{(2)}$ from each of its columns. Hence, if we know the rth column we can determine the hth column:

$$\boldsymbol{\beta}^{<h>} = \mathbf{T}_{r,h}\boldsymbol{\beta}^{<r>} \tag{20}$$

where $\mathbf{T}_{r,h}$ is constructed from a $k \times k$ identity matrix as follows: First, each entry in the hth column is given the value -1. Then, remove the rth column and hth row. Note that $\mathbf{T}_{s,r}\mathbf{T}_{s,h} = \mathbf{T}_{s,r}$ such that $\mathbf{T}_{h,r} = \mathbf{T}_{r,h}^{-1}$.

Example 4. *Consider*

$$\mathbf{R}^{(1)} - \mathbf{R}^{(2)} = \begin{pmatrix} 0 & -0.7 & -1 \\ 0.7 & 0 & -0.3 \\ 1 & 0.3 & 0 \end{pmatrix} \tag{21}$$

It follows that

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 0.7 \\ 1 \end{pmatrix} \tag{22}$$

The second column is

$$\mathbf{T}_{1,2}\boldsymbol{\beta}_1 = \begin{pmatrix} -1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0.7 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.7 \\ 0.3 \end{pmatrix} \tag{23}$$

If we go backwards:

$$\mathbf{T}_{2,1}\boldsymbol{\beta}_2 = \begin{pmatrix} -1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -0.7 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.7 \\ 1 \end{pmatrix} \tag{24}$$

Now the $\chi^2_{\Delta\mathbf{R}}$ statistic based on row $h \neq r$ can be written as:

$$\begin{aligned} \chi^2_{\Delta\mathbf{R}} &= (\mathbf{T}_{h,r}\boldsymbol{\beta})^T \left(\mathbf{T}_{h,r}\boldsymbol{\Sigma}\mathbf{T}^T_{h,r}\right)^{-1} \mathbf{T}_{h,r}\boldsymbol{\beta} && (25) \\ &= \boldsymbol{\beta}^T\mathbf{T}^T_{h,r} \left(\mathbf{T}^T_{h,r}\right)^{-1} \boldsymbol{\Sigma}^{-1}\mathbf{T}^{-1}_{h,r}\mathbf{T}_{h,r}\boldsymbol{\beta} && (26) \\ &= \boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} && (27) \end{aligned}$$

This shows that its value is the same whatever column we take.

6.3. Detect DIF in practice

The following is an R script that calculates the squared Mahalanobis distance, and the matrix of standardized differences $\hat{\mathbf{D}}$. It uses the eRm package (Mair & Hartzinger, 2007) to estimate the parameters and calculate the asymptotic variances and covariances of the parameters. This will suffice for small problems.

```
# Estimation
 library(eRm)
 res1=RM(dataGr1,sum0=FALSE)
 res2=RM(dataGr2,sum0=FALSE)
# Calculate \Delta\mathbf{R}
  R1=kronecker(res1$betapar,t(res1$betapar),FUN="-")
  R2=kronecker(res2$betapar,t(res2$betapar),FUN="-")
  DR=R2-R1
# Calculate \chi^2_{\Delta\mathbf{R}}
  beta=DR[2:k,1]
  Sigma=vcov(res1)+vcov(res2)
  chi2=mahalanobis(beta,rep(0,(k-1)),Sigma)
  p_chi2=1-pchisq(chi2,(k-1))
# make D
  Acov1=cbind(rep(0,k),rbind(rep(0,k-1),vcov(res1)))
  Acov2=cbind(rep(0,k),rbind(rep(0,k-1),vcov(res2)))
  var1=diag(Acov1)
  var2=diag(Acov2)
  S1=kronecker(var1,t(var1),FUN="+")-2*Acov1
  S2=kronecker(var2,t(var2),FUN="+")-2*Acov2
  S=S1+S2
  diag(S)=1
  D=DR/sqrt(S)
```

Some explanation is in order. First note that in the rth column of $\mathbf{R}^{(1)}$, we have difficulties relative to item r: i.e., $R_{ir}^{(g)} = \delta_i^{(g)} - \delta_r^{(g)}$. Hence,

$$\Sigma_{i,j} = Cov\left(\hat{R}_{ir}^{(1)}, \hat{R}_{jr}^{(1)}\right) + Cov\left(\hat{R}_{ir}^{(2)}, \hat{R}_{jr}^{(2)}\right) \tag{28}$$

$$= Cov\left(\widehat{\delta_i^{(1)} - \delta_r^{(1)}}, \widehat{\delta_j^{(1)} - \delta_r^{(1)}}\right) + Cov\left(\widehat{\delta_i^{(2)} - \delta_r^{(2)}}, \widehat{\delta_j^{(2)} - \delta_r^{(2)}}\right) \tag{29}$$

Estimating with sum0=FALSE means that we have chosen $r = 1$. The estimates res1$betapar, and res2$betapar *are* estimates of difficulty relative to item r in each group. The command

```
vcov(res1)
```

produces the asymptotic variances and covariances of the $k - 1$ relative difficulties $\widehat{\delta_i^{(1)} - \delta_r^{(1)}}$, for $i = 2, \ldots, k$ in the first group. It follows from Equation 29 that

```
Sigma=vcov(res1)+vcov(res2)
```

To calculate $\hat{\mathbf{D}}$, we need to calculate:

$$Var\left(\hat{R}_{ij}^{(g)}\right) = Var(\widehat{\delta_j^{(g)} - \delta_r^{(g)}}) + Var(\widehat{\delta_j^{(g)} - \delta_r^{(g)}}) - 2Cov(\widehat{\delta_j^{(g)} - \delta_r^{(g)}}, \widehat{\delta_j^{(g)} - \delta_r^{(g)}})$$

To this aim, we construct the complete variance covariance matrix: i.e., including the first item, using the command

```
Acov1=cbind(rep(0,k),rbind(rep(0,k-1),vcov(res1)))
```

which simply adds the zero column and row for the first item that was fixed. Now, $Var\left(\hat{R}_{ij}^{(g)}\right)$ is calculated for each entry in $\mathbf{R}^{(1)}$ by

```
S1=kronecker(var1,t(var1),FUN="+")-2*Acov1
```

and similarly for $\mathbf{R}^{(2)}$.

6.4. Produce a Color-Plot

The following R-script produces a simple color plot of the matrix $\hat{\mathbf{D}}$. This one is suppose to look the same on screen as printed[1]

```
plotD <- function(x)
 {
   yLabels <- c(ncol(x):1)
   xLabels <- c(1:nrow(x))
   x <- x[yLabels,]
     #set breaks for categories
  alpha=c(.05,.01,.001)
    #set colors for categories (1 more than breaks)
  cols=c("white","lightgray","darkgray","black")
    #calc. z-scores and define breaks
  qn=-qnorm(alpha/2)                 #z-scores
  br=c(0,qn[1],qn[2],qn[3],1000)     #breaks (>1000=white!)
    #plotting commands
  par(mar=c(2,2,2,2))                      #make all margins equal
  image(abs(x),col=cols,breaks=br,axes=F)  #use image command
  axis(1,labels=xLabels,at=seq(0,1,length.out=nrow(x)))
  axis(2,labels=yLabels,at=seq(0,1,length.out=nrow(x)))
  box()                               #square box around plot
  layout(1)
}
```

[1] Thanks to Matthieu Brinkhuis!

If we wish to order the rows and columns such that clusters are more easily seen we found that multidimensional scaling is useful. By construction, $\hat{\mathbf{D}}$ is unidimensional and the scale values of the items can be used to order the rows and columns of $\hat{\mathbf{D}}$ such that the clusters may be more easily seen. This is how we produced Figure 7.

```
library(MASS)
Dpo=abs(D)
Dpo[Dpo=0]=0.0001
diag(Dpo)=0
rr=isoMDS(Dpo,k=1)
o=order(rr$points)
ff=as.matrix(colnames(DD))
ff=cbind(ff,rr$points)
ff=ff[o,]
D_order=D[o,o]
```