PSYCHOMETRIKA—VOL. 80, NO. 2, 317–340
JUNE 2015
DOI: 10.1007/s11336-014-9408-y

A STATISTICAL TEST FOR DIFFERENTIAL ITEM PAIR FUNCTIONING

TIMO M. BECHGER AND GUNTER MARIS

CITO

GUNTER MARIS

UNIVERSITY OF AMSTERDAM

This paper presents an IRT-based statistical test for differential item functioning (DIF). The test is developed for items conforming to the Rasch (Probabilistic models for some intelligence and attainment tests, The Danish Institute of Educational Research, Copenhagen, 1960) model but we will outline its extension to more complex IRT models. Its difference from the existing procedures is that DIF is defined in terms of the relative difficulties of pairs of items and not in terms of the difficulties of individual items. The argument is that the difficulty of an item is not identified from the observations, whereas the relative difficulties are. This leads to a test that is closely related to Lord's (Applications of item response theory to practical testing problems, Erlbaum, Hillsdale, 1980) test for item DIF albeit with a different and more correct interpretation. Illustrations with real and simulated data are provided.

Key words: item response theory, differential item functioning, identifiability, Rasch model.

1. Introduction

This paper is about the statistical testing of *Differential Item Functioning (DIF)* based on *Item Response Theory (IRT)*, more specifically, about methods to detect DIF items. The simple fact that the difficulty of an item is not identified from the observations enticed us to develop a test for DIF in relative difficulty. Thus, while the existing tests are focussed on detecting differentially functioning *items*, we discuss a test to detect differentially functioning *item pairs*. The focus will be on the Rasch model which contains all the important ingredients of the general case but we will outline extensions to the other IRT models.

The outline of the paper is as follows: In the ensuing sections, we discuss the Rasch model, reiterate the common definition of DIF, survey the problems with the existing procedures, and motivate our focus on relative rather than absolute difficulty. Having thus set the stage, we derive a statistical test for *differential item pair functioning* and discuss its operating characteristics using real and simulated data for illustration. Then, we sketch how the idea underlying our approach can be used to develop statistical methods to investigate DIF in more complex IRT models. The paper ends with a discussion.

2. Preliminaries

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 item response variables

Correspondence should be sent to Timo M. Bechger, Cito, Amsterdamseweg 13, Arnhem, The Netherlands. Email: timo.bechger@cito.nl

X_{pi}, $p = 1, \ldots, n$, $i = 1, \ldots, k$. Associated with each item response is the probability $\pi_{pi} = P(X_{pi} = 1)$. The Rasch (1960) model assumes that the X_{pi} are independent and π_{pi} is a logistic function of a person parameter θ_p and an item parameter δ_i. Specifically,

$$\pi_{pi} = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)} \tag{1}$$

where all the parameters are real valued. The person parameter is interpreted as a person's ability, and the item parameter is interpreted as item difficulty; and implicitly defined as the ability for which $\pi_{pi} = 0.5$. We obtain a *random-effects* Rasch model by making the additional assumption that abilities, c.q. persons, are a random sample from some distribution. In DIF research we are mostly dealing with naturally occurring groups whose ability distributions are unknown[1] and not necessarily equal. Note that the two Rasch models are conceptually quite different (e.g., San Martin & Roulin, 2013, 3.6) but the material in this paper applies equally to both.

2.2. *Identifiability*

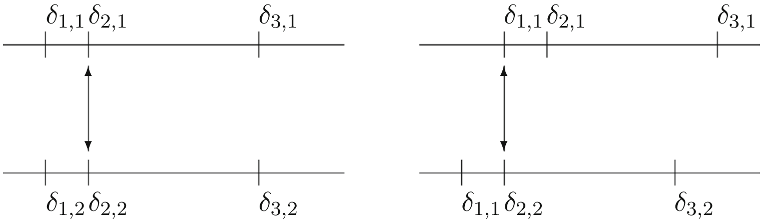
It is well known that the Rasch model is not identified by the item responses. Specifically, the origin of the latent scale is arbitrary which implies that we cannot determine the value of δ_i but we can determine the *relative difficulties* $\delta_i - \delta_j$. The Rasch model is identified if we impose a restriction of the form $\sum_i a_i \delta_i = 0$, where the a_i are arbitrary real constants such that $\sum_i a_i = 1$ (SanMartin and Roulin, 2013). This is an identification restriction called a *normalization*. A normalization puts the zero point somewhere on the ability scale, and nothing more. It is not a genuine, i.e., refutable, restriction like $\delta_2 - \delta_1 = 0$ and, from an empirical point of view, we may, therefore, choose any normalization we like. Here, we adopt the normalization $\delta_r = 0$, where r is the index of an arbitrary *reference item*. It follows that δ_i can be interpreted as the difficulty of item i relative to that of the reference item (San Martin, González, & Tuerlinckx, 2009). If we change the reference item, we change the parametrization of the model. Estimates of the item parameters, as well as their variances and covariance, will change (Verhelst, 1993): the fit of the model will not. The relative difficulties, in contrast, are independent of the normalization (e.g., Fischer, 2007, p. 538).

In this paper, we consider the situation where n persons in two independent, naturally occurring groups answer to the same k items. We do not *a priori* know how able one group is relative to the other, and we need *two* normalizations to identify the model; i.e., one for each group. To see this, note that the likelihood can be written 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, and we write $\delta_{i,g}$ to denote the difficulty parameter of the ith item in the gth group to allow for the possibility that the same item has a different difficulty in the groups (see e.g., Davier & Davier, 2007). It is easily seen that

[1] Technically, such groups are called non-equivalent to indicate that assignment to group was not random. We assume non-equivalent groups because (random) equivalent groups would not show any systematic difference, let alone DIF. The distribution of ability in non-equivalent groups may be the same or different.



(a) $\delta_{2,1} = \delta_{2,2} = 0$ **(b)** $\delta_{1,1} = \delta_{2,2} = 0$

FIGURE 1.
Two different normalizations. 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.

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

where $\delta_{i,1}^* = \delta_{i,1} - c, \theta_p^* = \theta_p - c, \delta_{i,2}^{**} = \delta_{i,2} - d$, and $\theta_q^{**} = \theta_q - d$, with c and d arbitrary constants. The constants need not be equal and, in principle, we could adopt different normalizations in each group. This is inconvenient, however, because the item parameters cannot be directly compared when the parametrization is different in each group.

To illustrate all this, Figure 1 shows two *item maps*: each representing a Rasch model with three items administered in two groups. The maps represent the same model and differ only in the normalization. In the left panel, the second item is the reference item in both groups, whereas, in the right panel, we have chosen a different reference item in each group. It is seen that the normalization fixes the origin on each scale *and* establishes a relation between them; i.e., it “anchors” the scales. The relative position of the two ability scales can be changed by adopting a different normalization in each group. This illustrates the well-known fact that between-group differences are not 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. The relative difficulties/positions of the items within each group are the identified parameters and independent of the normalization.

2.3. *Differential Item Functioning: Lord’s Test*

Assuming that a Rasch model holds within each group, an item is said to exhibit DIF if the item response functions (given by Eq. 1) for the two groups differ. There is no DIF when none of the items show DIF and the same Rasch model holds for both groups. The item response functions are determined by the item difficulty parameters and Lord (1977, 1980, 14.4) noted that the question of DIF detection could be approached by computing estimates of the item parameters within each group. Specifically, he proposed a Wald (1943)-type test using the standardized difference

$$d_i = \frac{\hat{\delta}_{i,1} - \hat{\delta}_{i,2}}{\sqrt{\text{Var}\left(\hat{\delta}_{i,1}\right) + \text{Var}\left(\hat{\delta}_{i,2}\right)}} \tag{2}$$

as a statistic to test the hypothesis $\delta_{i,1} = \delta_{i,2}$ against $\delta_{i,1} \neq \delta_{i,2}$, where the hats denote estimates (see also Wright, Mead, & Draba, 1976 or Fischer, 1974, p. 297). Lord notes that the estimates $\hat{\delta}_{i,g}$ (and their variances) must be obtained using the same normalization/parametrization in each group. Figure 1 illustrates why. The left panel shows that all the item parameters can be aligned and there is no DIF. Nevertheless, we may erroneously find DIF items if we use a different normalization in both groups as seen in the right panel.

3. The Problem with the Detection of DIF Items

In spite of their widespread use, it is well known that common statistical methods for the detection of DIF items do not perform as one would hope. First of all, conclusions about DIF may be sensitive to the between-group differences while DIF concerns the relation between ability and the observed responses and has nothing to do with the between-group differences or *impact* (Dorans & Holland, 1993, pp. 36–37). This was found for the *simultaneous item bias test* (SIBTEST: Shealy & Stout, 1993) by Penfield and Camelli (2007, p. 125), Gierl, Gotzmann, and Boughton (2004), and DeMars (2010), for the *Mantel–Haenszel (MH) test* by Magis and Boeck (2012), and for logistic regression (Swaminathan & Rogers, 1990) by Jodoin and Gierl (2001). Second, many procedures show a tendency to detect too many DIF items. This has been observed in numerous simulation studies (e.g., Finch & French, 2008; Stark, Chernyshenko, & Drasgow, 2006; DeMars, 2010) and is known as "type-I error inflation." It is best documented for procedures which use the raw score as a proxy for ability (Magis & Boeck, 2012) but it is easy checked (using simulation) that there are procedures based more directly on IRT that share the same problem: For instance, the S_i test (Verhelst & Eggen, 1989; Glas & Verhelst, 1995a,b) incorporated in the OPLM software (Verhelst, Glas, & Verstralen, 1994). Finally, adding or removing items may turn non-DIF items into DIF items or vice versa. This issue is not as well documented as the previous ones which is why we illustrate it with a real data example.

Example 1. (*Preservice Teachers*)The dataset consists of the responses of 2, 768 Dutch preservice teachers to a subset of six items taken from a test known to fit the Rasch model. The students differ in their prior education: 1, 257 students have a background in vocational education, and 1, 511 entered directly after finishing secondary education. It was hypothesized that prior education affects item difficulty, and the difference in prior education might show up as DIF. We used three methods to detect DIF items: Lord's test, Raju's signed area test (Raju, 1988, 1990), and the likelihood ratio test (LRT: Thissen, Steinberg, & Wainer, 1988; Thissen, 2001). The results are in Table 1. After removing the DIF items 14 and 16, we obtain the results in Table 1 indicating that item 3 has now become a DIF item while it was not a DIF item in the first analysis. The reverse happened to items 22 and 24.

Over the years, a large body of research was developed aimed at explaining and remedying this behavior. This revealed that all the procedures are, to some extent, conceptually or statistically flawed.[2] For starters, some statistical tests are formulated in such a way that the null hypothesis and the alternative hypothesis may both be false. This is most clearly the case with Lagrange multiplier (Glas, 1998) and the related tests (e.g., Tay, Newman, & Vermunt, 2011). Assuming a Rasch model for each of two groups, the null hypothesis is

$$H_0 : \forall i \quad \delta_{i,1} = \delta_{i,2}$$

and tested against the alternative:

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

[2] Bad practice is a frequent source of problems: in particular, the misspecification of the measurement model, giving the appearance of DIF, or capitalization on chance (e.g., Teresi, 2006). We assume here a fitting Rasch model in each group.

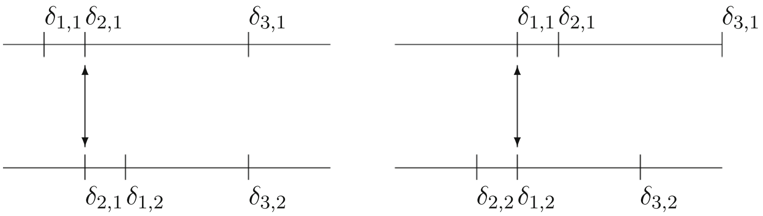
TABLE 1.
DIF analyses of six items from a test for preservice teachers with the original item labels.

Item	Lord	Raju	LRT
(a) First analysis			
3	NoDIF	NoDIF	NoDIF
14	DIF	DIF	DIF
16	DIF	DIF	DIF
20	NoDIF	NoDIF	NoDIF
22	DIF	DIF	DIF
24	DIF	DIF	DIF
(b) Results after removing items 14 and 16			
3	DIF	DIF	DIF
20	NoDIF	NoDIF	NoDIF
22	NoDIF	NoDIF	NoDIF
24	NoDIF	NoDIF	NoDIF

When H_0 is rejected in favor of the alternative hypothesis H_{1i}, 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 items show no DIF. If this condition does not hold, neither H_0 nor H_{1i} is true and it is difficult to say what we expect for the distribution of the test statistic. We do know that there are less restrictions under the alternative hypothesis, which explains the tendency to conclude that all the items show DIF. This observation is not new (e.g., Hessen, 2005, p. 513) and has enticed researchers to suggest *anchoring strategies* (e.g., De Boeck, 2008, §7.4; Wang & Yeh, 2003, 2004; Stark et al., 2006), and *heuristics* for iterative *purification* of the anchor (Lord, 1980, p. 220; Van der Flier, Mellenbergh, Adèr, & Wijn, 1984; Candell & Drasgow, 1988; Wang, Shih, & Sun, 2012). Unfortunately, none of this provides a valid statistical test when there is more than one DIF item. Furthermore, it is known that purification may be associated with type-I error inflation and tends to increase the sensitivity to impact (Magis & Facon, 2013; Magis & Boeck, 2012).

A more principled solution would be to follow the rationale of Andersen's goodness-of-fit tests (Andersen, 1973) and compare estimates $\hat{\delta}_{i,g}$ obtained from separate calibrations as suggested by Lord. This solves the aforementioned problem but now we are faced with a new one. To wit, the results *depend on the normalization*, even if we use the same normalization in each group. Choosing a reference item means, first of all, that we arbitrarily decide that one of the items is DIF free. This by itself is inconvenient: how do we know this item is DIF free? However, if there is DIF, the normalization also determines which items are DIF items. We illustrate this with Figure 2 showing a situation where the relative difficulties, except $\delta_3 - \delta_2$, are different in the two groups and there is DIF. The logic underlying Lord's test is that an item exhibits DIF when its parameter (i.e., its location) is different for the two groups. Thus, from Figure 2a, we would conclude that item 1 has DIF. Figure 2b, however, would lead us to conclude that items 2 and 3 have DIF. From an empirical point of view both the conclusions are equally valid which means that we cannot tell which items are the DIF items, unless we take refuge to a substantive argument[3]. Unfortunately, Lord's test is not the only test where the results depend on the normalization. Related tests, like Raju's signed area test, suffer from the same defect. While these tests are widely used, close

[3] Some researchers would report item 1 on the argument that DIF items are necessarily a *minority* (e.g., De Boeck, 2008): DIF items are viewed as outliers with respect to other (non-DIF) items (Magis & Boeck, 2012). This, however, is an opinion; it cannot be proven correct or incorrect using empirical evidence. Furthermore, as suggested to us in a personal communication by Norman Verhelst, this working principle implies that items that show DIF in a test, where they are a minority, may not show DIF in another test where they constitute a majority.



(a) $\delta_{2,1} = \delta_{2,2} = 0$ **(b)** $\delta_{1,1} = \delta_{1,2} = 0$

FIGURE 2.
The effect of different normalizations when there is DIF.

reading of the literature reveals that many psychometricians are aware of the dependency of their test statistics on the arbitrary normalization (e.g., Shealy & Stout, 1993, p. 207; De Boeck, 2008, p. 551; Soares, Concalves, & Gamerman, 2009, pp. 354–355; Glas & Verhelst, 1995a, pp. 91–92; Magis & Boeck, 2012, p. 293, or Wang, 2004). However, we could not find a publication dedicated to this issue and, as far as we know, a solution has not been offered.

Finally, we mention a class of procedures which do not suffer from the aforementioned flaws but have a dependence between different items, as it where, “build-into” the test. Suppose, for instance, that we compare the *item-total regressions* to detect DIF.[4] (Holland & Thayer, 1988; Glas, 1989) Under the Rasch model we have

$$P(X_i = 1 | X_+ = s) = \frac{\epsilon_i \gamma_{s-1}(\boldsymbol{\epsilon}^{(i)})}{\gamma_s(\boldsymbol{\epsilon})}, \tag{3}$$

where $\epsilon_i \equiv \exp(-\delta_i)$, $X_+ = \sum_i X_i$ denotes the number of correct answers on the test, $\gamma_s(\boldsymbol{\epsilon})$ the *elementary symmetric function* of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_k)$ of order s (e.g., Andersen, 1972; Verhelst, Glas, & van der Sluis, 1984), and $\boldsymbol{\epsilon}^{(i)}$ denotes the vector $\boldsymbol{\epsilon}$ without the ith entry ϵ_i. An item shows no DIF when its item-total regression function is the same in both the groups which are tested against its negation. Interestingly, we now have a valid statistical test. It is easily checked that the regression functions are independent of the normalization[5] and, in contrast to Lord’s test, any item can be flagged as a DIF item—including the reference item. Thus, the previously mentioned problems are out of the way. Unfortunately, we still cannot identify the DIF items. As explained in more detail in Meredith and Millsap (1992), one DIF (or in general, misfitting) item affects *all* the item-total regression functions and a dependence between items is inherent to (hence an artifact of) the procedure. This issue is reminiscent of what in the older DIF literature was coined the *ipsative* or circular nature of DIF (e.g., Angoff, 1982; Camilli, 1993, pp. 408–411; Thissen, Steinberg, & Wainer, 1993, p. 103; Williams, 1997). We will not discuss this further but note that no solution has been reported (Penfield & Camelli, 2007, pp. 161–162).

4. Motivation

In DIF research we first ask whether there is DIF. If so, we look at an individual item and ask whether it shows DIF. From the previous section, we can conclude that we are currently unable to answer the second question. We can detect DIF but we cannot identify the DIF items. After some

[4] The same idea underlies the S_i test proposed by Glas and Verhelst (1995b).

[5] This follows because $\gamma_s(c\boldsymbol{\epsilon}) = c^s \gamma_s(\boldsymbol{\epsilon})$.

thought, we have come to the conclusion that the problem lies in the question rather than in the procedure—more specifically, in the notion that DIF is the property of an individual item. The snag is that we try to test a hypothesis on the difficulty of an item, ignoring that it is not identified. Consider, for example, how this explains the behavior of Lord's test. If we perform Lord's test we assume that the null hypothesis (i.e., $\delta_{i,1} = \delta_{i,2}$) is independent of the normalization. At the same time, the test statistic clearly depends on the normalization which, as shown earlier, leads to the predicament that, based on the same evidence, an item can and cannot have DIF. On closer look, the null hypothesis (as well as the alternative) is stated in terms of the parameters that are not identified from the observations. Unidentified parameters cannot be estimated (Gabrielsen, 1978; San Martin & Quintana, 2002) and, under the normalization $\delta_{r,g} = 0$, $\hat{\delta}_{i,g}$ is actually an estimate of $\delta_{i,g} - \delta_{r,g}$, for $g = 1, 2$. The null hypothesis, therefore, is not $\delta_{i,1} = \delta_{i,2}$ but $\delta_{i,1} - \delta_{r,1} = \delta_{i,2} - \delta_{r,2}$; that is, it changes with the normalization, just like the test statistic. It follows that a significant value of d_i does not mean that item i shows DIF. Instead, it means that the *item pair* (i, r) has DIF. Thus, it is a mistake to use Lord's test as a test for item DIF. If we abandon this idea, Lord's test works fine and this is essentially what we will do.

We will proceed as follows. We follow the rationale that underlies Lord's test and base inferences on separate calibrations of the data within each group. We *define* DIF in terms of the relative difficulties $\delta_i - \delta_j$ because these are identified by the observations. Thus, we obtain a Wald test for the invariance of the relative difficulties of item pairs, which is similar to Lord's test but independent of the normalization. Details are in the next section.

5. Statistical Tests for DIF or Differential Item Pair Functioning

In this section, we discuss how to test for DIF and how to detect differentially functioning item pairs when there is DIF. The only requirement is that we have consistent, asymptotically normal estimates of relative difficulties, no mater how they are obtained. We focus on the test for DIF pairs: DIF testing is a routine matter and can also be done using a standard likelihood ratio test (Glas & Verhelst, 1995a). Both the tests are based on relative difficulties collected in a matrix $\Delta\mathbf{R}$ which will be introduced first. R-code (R Development Core Team, 2010) is in the Appendix.

5.1. The Matrix $\Delta\mathbf{R}$

Let $\mathbf{R}^{(g)}$ with entries $R_{ij}^{(g)} = \delta_{i,g} - \delta_{j,g}$ denote a $k \times k$ matrix of the pairwise differences between the parameters of items taken by group g. When there is no DIF, $\mathbf{R}^{(1)} = \mathbf{R}^{(2)}$. For later reference we state this as

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

where $\Delta\mathbf{R} \equiv \mathbf{R}^{(1)} - \mathbf{R}^{(2)}$ is a matrix of "differences between groups in the pairwise differences in difficulty between items." The properties of this matrix are illustrated in the next example.

Example 2. (*Three-item example*) With item parameters corresponding to Figure 2, $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ may look like this:

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

It follows that

$$\Delta\mathbf{R} = \begin{pmatrix} 0 & -1.5 & -1.5 \\ 1.5 & 0 & 0 \\ 1.5 & 0 & 0 \end{pmatrix}. \tag{4}$$

Note that $\Delta\mathbf{R}$ is *skew-symmetric*: that is, it has zero diagonal elements and $\Delta R_{ij} = -\Delta R_{ji}$, for all i, j. Note further that there are zero as well as non-zero off-diagonal 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 the groups. The non-zero entries in the first row (column) are due to the fact that item 1 has changed its position relative to items 2 and 3.

As illustrated by Example 2, the matrix $\Delta\mathbf{R}$ is *skew-symmetric by construction*. Furthermore, we can reproduce the whole matrix from the entries in any single row or column. Specifically,

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

In fact, we need only know the off-diagonal entries, since the diagonal entries are always zero. It follows that $\Delta\mathbf{R} = \mathbf{0} \Leftrightarrow \boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{\beta}$ is any column of $\Delta\mathbf{R}$ but *without the zero diagonal entry*. Thus, we can restate our hypotheses in terms of $\boldsymbol{\beta}$ which suggests a straightforward way to test whether DIF is absent.

5.2. A Test for No DIF

Let $\boldsymbol{\beta}$ be taken from the rth column of $\Delta\mathbf{R}$, with r the index of any of the items. An estimate $\hat{\boldsymbol{\beta}}$ is calculated with the estimates of item parameters and used to make inferences about $\boldsymbol{\beta}$. We assume that the estimators of the item parameters are unbiased asymptotically normal. It then follows that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal, with mean $\boldsymbol{\beta}$ and variance–covariance matrix $\boldsymbol{\Sigma}$ with entries:

$$\begin{aligned}\Sigma_{i,j} = \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) &= \mathrm{Cov}\left(\hat{R}^{(1)}_{ir} - \hat{R}^{(2)}_{ir}, \hat{R}^{(1)}_{jr} - \hat{R}^{(2)}_{jr}\right) \\ &= \mathrm{Cov}\left(\hat{R}^{(1)}_{ir}, \hat{R}^{(1)}_{jr}\right) + \mathrm{Cov}\left(\hat{R}^{(2)}_{ir}, \hat{R}^{(2)}_{jr}\right),\end{aligned} \tag{5}$$

where $i, j \neq r$. The second equality follows from the fact that the calibrations are based on independent samples. The (co)variances in (5) can be obtained from the asymptotic variance–covariance matrix of the item parameter estimates.[6]. Specifically, if we estimate under the normalization $\delta_{r,1} = \delta_{r,2} = 0$, it holds that $\Sigma_{i,j} = \mathrm{Cov}\left(\hat{\delta}_{i,1}, \hat{\delta}_{j,1}\right) + \mathrm{Cov}\left(\hat{\delta}_{i,2}, \hat{\delta}_{j,2}\right)$

Testing for DIF is now reduced to a common statistical problem. When $\boldsymbol{\beta} = \mathbf{0}$, the *squared Mahalanobis distance*

$$\chi^2_{\Delta\mathbf{R}} \equiv \hat{\boldsymbol{\beta}}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}} \tag{6}$$

follows a chi-squared distribution with $k - 1$ degrees of freedom, where k is the number of items (Mahalanobis, 1936; see also Rao, 1973, pp. 418–420; Gnanadesikan & Kettenring, 1972). It is easy to prove that the value of $\chi^2_{\Delta\mathbf{R}}$ is independent of the column of $\Delta\mathbf{R}$ (see Bechger, Maris, & Verstralen, 2010, 6.2).

[6] In general, $\mathrm{Cov}\left(\hat{R}^{(g)}_{ir}, \hat{R}^{(g)}_{jr}\right) = \mathrm{Cov}\left(\hat{\delta}_{i,g}, \hat{\delta}_{j,g}\right) + \mathrm{Var}\left(\hat{\delta}_{r,g}\right) - \mathrm{Cov}\left(\hat{\delta}_{i,g}, \hat{\delta}_{r,g}\right) - \mathrm{Cov}\left(\hat{\delta}_{j,g}, \hat{\delta}_{r,g}\right).$

5.3. A Test for Differential Item Pair Functioning

If there is DIF we may wish to test whether a particular pair of items has DIF. To assess the hypothesis $H_0^{(ij)} : R_{ij}^{(1)} = R_{ij}^{(2)}$ against $H_1^{(ij)} : R_{ij}^{(1)} \neq R_{ij}^{(2)}$, we calculate

$$\hat{D}_{ij} = \frac{\hat{R}_{ij}^{(1)} - \hat{R}_{ij}^{(2)}}{\sqrt{\mathrm{Var}\left(\hat{R}_{ij}^{(1)} - \hat{R}_{ij}^{(2)}\right)}} = \frac{\hat{R}_{ij}^{(1)} - \hat{R}_{ij}^{(2)}}{\sqrt{\mathrm{Var}\left(\hat{R}_{ij}^{(1)}\right) + \mathrm{Var}\left(\hat{R}_{ij}^{(2)}\right)}}. \tag{7}$$

The sampling variance of the estimated relative difficulties can be obtained from the sampling variances of the item parameters. Specifically,

$$\mathrm{Var}\left(\hat{R}_{ij}^{(g)}\right) = \mathrm{Var}\left(\hat{\delta}_{i,g}\right) + \mathrm{Var}\left(\hat{\delta}_{j,g}\right) - 2\mathrm{Cov}\left(\hat{\delta}_{i,g}, \hat{\delta}_{j,g}\right). \tag{8}$$

Under $H_0^{(ij)}$, the *standardized difference* $\hat{D}_{ij}$ is asymptotically standard normal which provides a useful metric to evaluate the entries of $\Delta\mathbf{R}$. We reject $H_0^{(ij)}$ when $|\hat{D}_{ij}| \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is such that $2[1 - \Phi(z_{\alpha/2})] = \alpha$; i.e., $\alpha = 0.05$ implies that $z_{\alpha/2} \approx 1.959$. Note that the significance level can be adapted to correct for *multiple testing* (e.g., Holm, 1979).

Example 3. (*Three-item example continued*) We randomly generate 200 responses in each group, with $\Delta\mathbf{R}$ as in Eq. 4. Ability distributions are normal with $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 0.2, and \sigma_2 = 0.8$. 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 the item pairs, we find

$$\widehat{\Delta\mathbf{R}} = \begin{pmatrix} 0 & -1.36 & -0.87 \\ 1.36 & 0 & 0.50 \\ 0.87 & -0.50 & 0 \end{pmatrix}, \quad \text{and} \quad \hat{\mathbf{D}} = \begin{pmatrix} 0 & -4.46 & -2.86 \\ 4.46 & 0 & 1.62 \\ 2.86 & -1.62 & 0 \end{pmatrix},$$

where $\hat{\mathbf{D}}$ denotes the (skew-symmetric) matrix with entries $\hat{D}_{ij}$. Compared to $\widehat{\Delta\mathbf{R}}$, $\hat{\mathbf{D}}$ has the advantage that we can interpret the value of each entries as the realization of a standard normal random variable. At a glance, it is clear that (with $\alpha = 0.05$) 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.

It will now be clear how our test relates to that of Lord: $\hat{D}_{ir}$ is equal to Lord's d_i when $\delta_r = 0$ for identification. Thus, if we use Lord's test we look at only one column in $\hat{\mathbf{D}}$. When we test for the absence of DIF, it is irrelevant which column we use. It does matter when we test for item pair DIF.

5.4. Item Clusters

If a Rasch model holds, we expect DIF pairs to form *clusters*; sets of items whose relative difficulties are invariant across groups. To see this, image an item map like Figure 1 with three items, A, B, and C; ordered in difficulty along the latent scale. If A and B keep their distance across groups, and so do B and C, then it is implied that A and C do also. Formally, we 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. This definition guarantees that different clusters are mutually exclusive. Furthermore, a single item i can also form a cluster when $\delta_{i,1} - \delta_{h,1} \neq \delta_{i,2} - \delta_{h,2}$ for all $h \neq i$. This was the case of item 1 in the three-item example.

Testing whether a particular group of items form a cluster is straightforward: a cluster corresponds to a sub-matrix of $\Delta\mathbf{R}$ and we simply test whether an arbitrary column in this sub-matrix is equal to zero using a squared Mahalanobis distance. We expect this to be more powerful than testing each pair separately. If we have no hypotheses about clusters, visual inspection of the matrix $\hat{\mathbf{D}}$ may help to spot them, as illustrated in Figure 3. Specifically, Figure 3 shows a color map of $\hat{\mathbf{D}}$ in two situations:

1. In the absence of DIF, there is one cluster containing all the items (Figure 3a).
2. When one item has changed its difficulty relative to all the other items, there are two clusters: one consisting of one item and one consisting of all others (see Figure 3b).

Note that $\widehat{\Delta\mathbf{R}}$ is a rank-one matrix, and a technique like multidimensional scaling can be used to represent the items as points on a line such that the relative difficulty of items closer together are more alike across the groups. If we use the positions to order the rows and column of $\hat{\mathbf{D}}$ we might be able to see the clusters better.

5.5. *Small Samples*

The procedure relies on asymptotic arguments. With small data sets a *Markov Chain Monte Carlo (MCMC)* algorithm may be used to determine the reference distribution of either of the statistics in this paper. Such algorithms have been described by Ponocny (2001) and Verhelst (2008) and implemented in the raschsampler R package (Verhelst, Hartzinger, & Mair, 2007). Under the Rasch model, we can avoid repeated estimation of the item parameters and their standard errors using a simple consistent estimator:

$$d_{ji}^{(n)} = \ln \frac{\sum_p X_{pi}(1 - X_{pj})}{\sum_p X_{pj}(1 - X_{pi})} \tag{9}$$

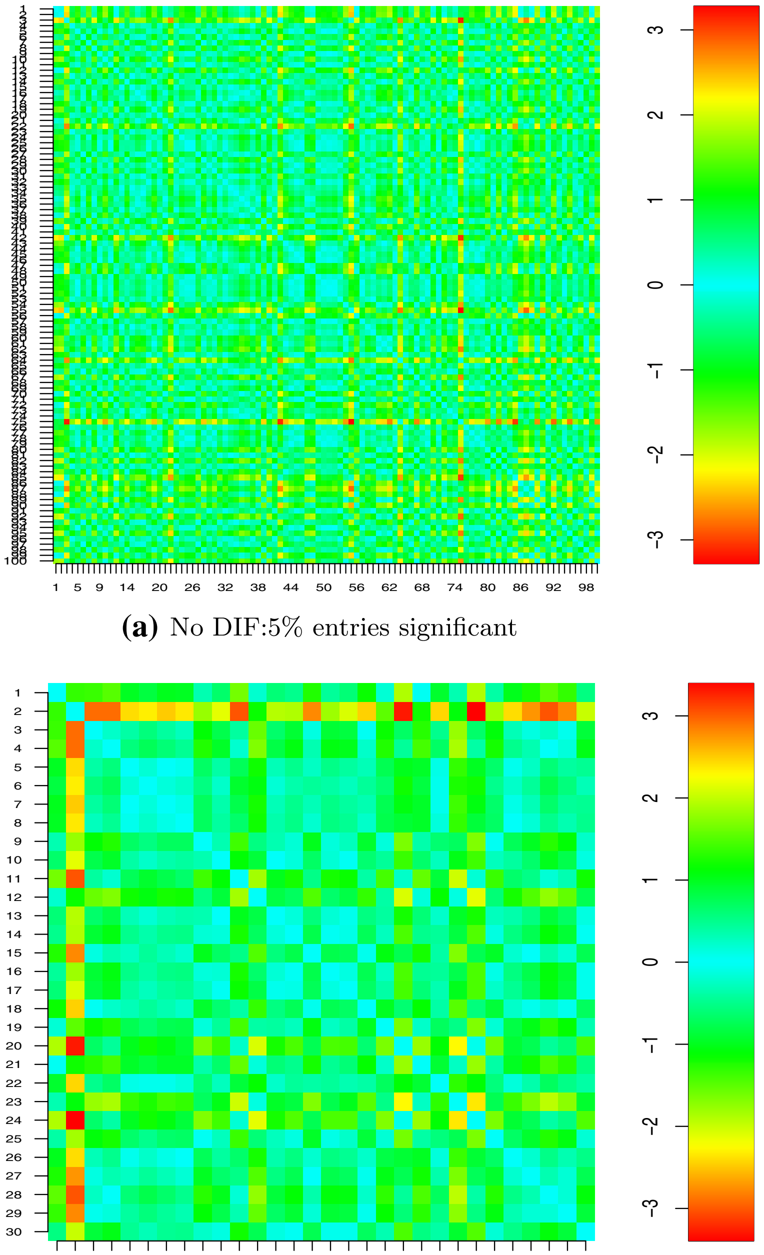
for $\delta_j - \delta_i$ (Fischer, 1974; Zwinderman, 1995).

5.6. *More than Two Groups*

The procedure, as we described it so far, is tailored for a comparison involving two groups. With more than two groups we would have a matrix $\hat{\mathbf{D}}$ for each pair of groups and the number of matrices would rapidly increase. Testing whether an item pair shows DIF for some combinations of groups can be done as follows. With G groups, we have G independent random variables $\hat{R}_{ij}^{(g)}$. When the difficulty of the item pair is invariant,

$$\hat{R}_{ij}^{(g)} \sim \mathcal{N}\left(\delta_i - \delta_j, \mathrm{Var}\left(\hat{R}_{ij}^{(g)}\right)\right). \tag{10}$$

Testing this hypothesis is again a standard statistical problem. When there is no DIF, the sum-of-squares of the standardized $\hat{R}_{ij}^{(g)}$ follow a chi-squared distribution with $G - 1$ degrees of freedom.



(a) No DIF:5% entries significant

(b) $\forall i \neq 2$: $R^{(2)}_{2i} \neq R^{(1)}_{2i}$.

FIGURE 3.
A color map of the matrix $\hat{\mathbf{D}}$ based on simulated data ($n = 1000$). Above, the situation where there is no DIF. Below, item 2 has changed its position relative to the other items.

6. Statistical Properties

In this section, we consider the statistical properties of the test for differential item pair functioning. We will first consider what can be said about these properties without simulation and then provide a simulation study.

6.1. Theoretical

6.1.1. Type-I Error Each entry $\hat{D}_{ij}$ can be estimated *consistently* and it follows that our test is consistent. Even for small datasets, we expect the result for one item pair to have a negligible effect on the result for others. The simulations described in Example 3 confirm this. Specifically, Figure 4a shows that the test for items 3 and 2 has the correct level in spite of the fact that other relative difficulties are not invariant. Figure 4b provides the same result for the relative difficulty of item 3 with respect to item 1. This item pair was not invariant across the groups and Figure 4b suggests that, in this case, our test was quite powerful.

6.1.2. Power Asymptotically, the statistic $\hat{D}_{ij}$ is distributed as the difference between two normal variates. As illustrated in Figure 5, the power to detect DIF for a pair of items depends on the amount of DIF (i.e., $R_{ij}^{(1)} - R_{ij}^{(2)}$), and on how well we estimate the relative difficulties; that is, on their sampling variance. As seen from Eq. 8, the sampling variance of the relative difficulties is independent of the amount of DIF but it does depend on the sample size in each group as well as on the difficulty of the items relative to the ability distribution: To wit, we get more information about the items when they are neither too difficult nor too easy for the persons taking them. It follows that impact may influence power and also that the power may differ across item pairs.

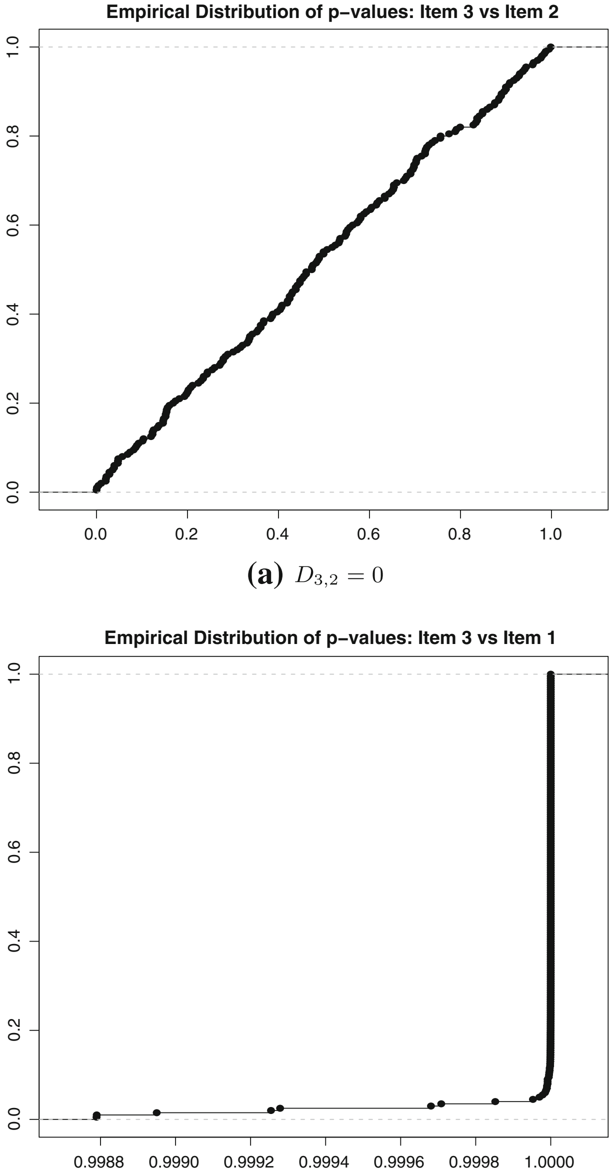
With an estimate, σ_{ij}^2, of the sampling variance of $\hat{R}_{ij}^{(1)} - \hat{R}_{ij}^{(2)}$ we can determine an asymptotic estimate of the power to detect a difference $\Delta R_{ij} = R_{ij}^{(1)} - R_{ij}^{(2)}$. Asymptotically, the sampling distributions in Figure 5 are normal and we can easily calculate the probability that $\left|\Delta R_{ij}\right| > z_{\alpha/2}\sigma_{ij}$. Note that we can use these estimates to "filter" the matrix $\hat{\mathbf{D}}$; i.e., show only those entries for which we have a minimally required power to detect a pre-determined amount of DIF. An illustration will be given below.

6.2. Simulation

We use simulation to illustrate the independence of tests for different pairs and in passing gain an impression of the power for relatively small samples. We simulated data from two groups of people taking a test of 30 items. Both the ability distributions are normal: i.e., $\mathcal{N}(0, 2)$ in the first group and $\mathcal{N}(-0.2, 3)$ in the second group. The item parameters are drawn from a uniform distribution between -1.5 and 0.5, which corresponds to a test that is appropriate for both the groups with percentages correct in the range from 0.4 to 0.7. We have the same number of observations in each group ranging between 100 and 10,000. Item parameters are estimated using *conditional maximum likelihood* (CML). DIF was simulated by changing, in the second group, the difference between items 5 and 10. Specifically,

$$\delta_{5,2} = \delta_{5,1} - c, \quad \text{and} \quad \delta_{10,2} = \delta_{10,1} - c,$$

where c ranged from zero (i.e., no DIF) to 1 and specifies the amount of the DIF; that is, $\Delta R_{10,8} = c$. Thus, the distance between item 5 and item 10 is invariant, while both the items change their position with respect to the other item. When judged against the standard deviation of ability, the amount of DIF is small to medium sized.



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

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

FIGURE 4.
Empirical distribution of the cumulative probability of $\hat{D}_{3,1}$ and $\hat{D}_{3,2}$ under the standard normal distribution based on a simulation. $n = 300$.

The results are shown in Figure 6. Figure 6b shows that the test has the correct level and the conclusion about the distance between items 5 and 10 is unaffected by the fact that the distance between items 10 and 8 was not invariant across the groups. Figure 6a shows the power of the test for items 10 and 8 as a function of effect size. It will be clear that, under the conditions of the present simulations, sample sizes less than 500 persons in each group are insufficient to detect even moderate amounts of item pair DIF. One might expect the exact test to have more power

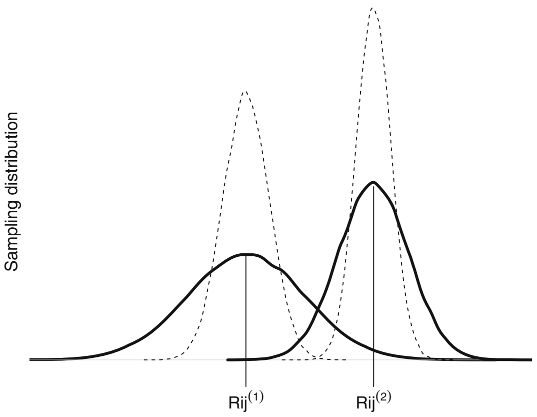


FIGURE 5.
Sampling distribution of differences between the item parameters in two groups for two different sample sizes. The *dotted lines* give the distribution for the larger sample size.

than the asymptotic test for small sample sizes, but our simulations suggest that the power is the same.

7. Real Data Examples

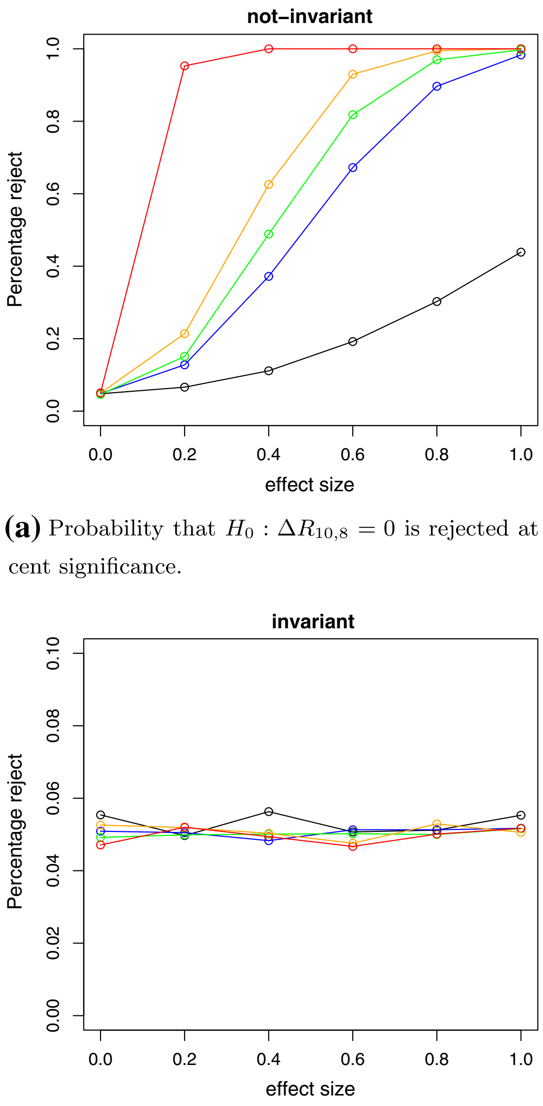
First, we re-analyze the preservice teacher data using the procedure developed in this paper. We then present two new applications using data collected for the development of an IQ test for children in the age between 11 and 15. The first application serves to show that it is possible, in practice, to encounter a number of substantial clusters. The second application illustrates cross-validation of the results. The same data are analyzed with traditional methods complementing the examples presented earlier.

7.1. Preservice Teachers

The earlier analysis (see Example 1) suggested that items 14, 16, 22, and 24 are DIF items. Using methods for item DIF we observed that removing items 14 and 16 changed item 3 into a DIF item and items 22 and 24 into DIF-free items. Figure 7 shows the $\hat{\mathbf{D}}$ matrix for this example. It suggests that there are three clusters: $X = \{3\}$, $Y = \{14, 16\}$, and $Z = \{20, 22, 24\}$. Figure 8 shows how much power we have to detect the observed differences in each cell. If we remove items 14 and 16 we have the situation where all the relative difficulties are invariant except for those relative to item 3. This was picked up by the traditional methods that classified item 3 as a DIF item. However, Lord's test would likely have flagged items 22 and 24 if we had normalized by setting the parameter of item 3 to zero.

7.2. IQ: Digital Versus Paper-and-Pencil

For our present purpose, we have taken two (convenience) samples of (424 and 572) 12-year-old children who have taken a subscale of the test consisting of 15 items. This subscale consisted of items that required students to infer the order in a series of manipulated (mostly 2-dimensional) figures. The first group has taken the test on the computer, while the second group has taken the paper-and-pencil version of the test. This subscale was found to fit the Rasch model in each group.



(a) Probability that $H_0 : \Delta R_{10,8} = 0$ is rejected at 5 percent significance.

(b) Probability that $H_0 : \Delta R_{5,10} = 0$ is rejected at 5 percent significance.

FIGURE 6.
Probability to reject at 5% significance level determined using 10, 000 simulations under each condition. The *curves* belong to sample sizes of 100, 500, 700, 1000, and 10000 in each group. The effect size is $\Delta R_{10,8}$, which means that the top figure shows a power curve. The *bottom picture* shows the type-I error rate.

The items look 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. A color plot of the matrix $\hat{\mathbf{D}}$ is shown in Figure 9, where multidimensional scaling was used to permute the rows and columns such that clusters are more clearly visible. There appear to be three substantial clusters: $X = \{11, 7, 4, 9, 14, 15, 13\}$, $Y = \{6, 2, 12\}$, and $Z = \{3, 8, 5, 1, 10\}$. Items $\{3, 8, 5\}$, for instance, have changed their difficulty relative to items $\{14, 15, 13\}$, which confirms that they belong to different clusters. Unfortunately, we would capitalize on chance if we use the

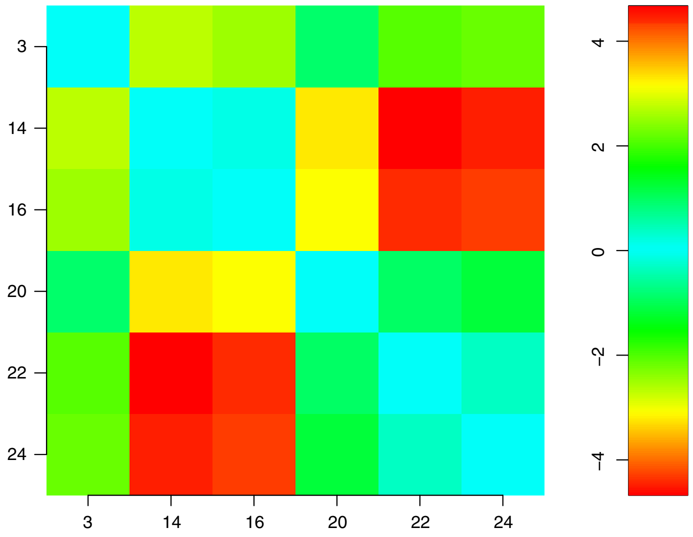


FIGURE 7.
Image of the matrix $\hat{\mathbf{D}}$ for the preservice teacher data.

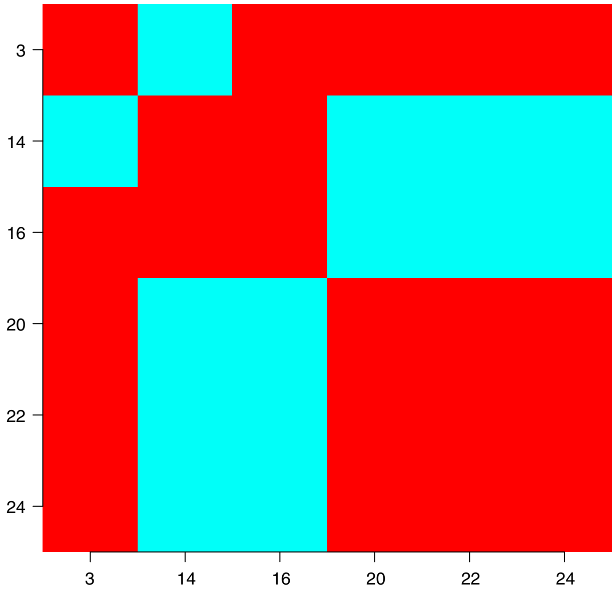


FIGURE 8.
Asymptotic power to detect the observed difference for the preservice teacher example. Entries corresponding to light areas have a power over 70 %.

same data to test, for instance, the hypothesis that X and Y could be combined in one cluster. Cross-validation is considered in the next example.

As an illustration, we have applied the *likelihood ratio test with purification* (Thissen, 2001). Note that when there is one item that has changed its relation to all the other items (as in Figure 3b), it is likely that any other test for item DIF will detect the item. This is not the case in the

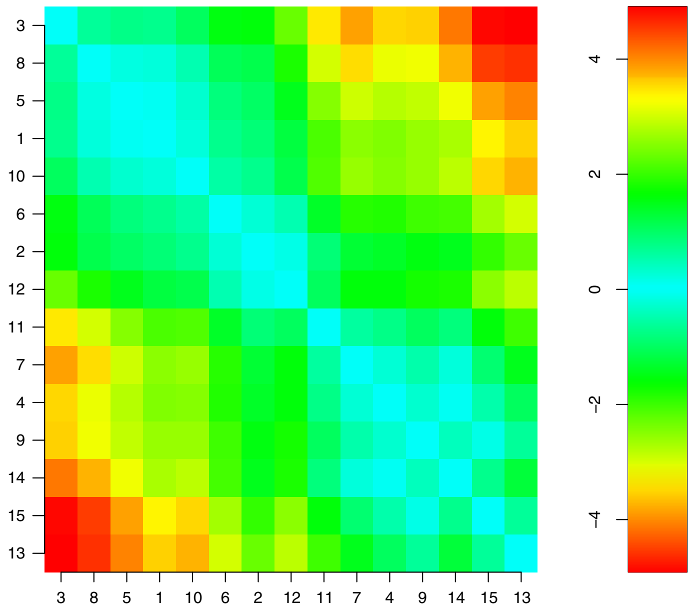


FIGURE 9.
The standardized matrix $\hat{\mathbf{D}}$ for the digital versus paper-and-pencil example. About 22 % of the entries is significant at the 5 % level.

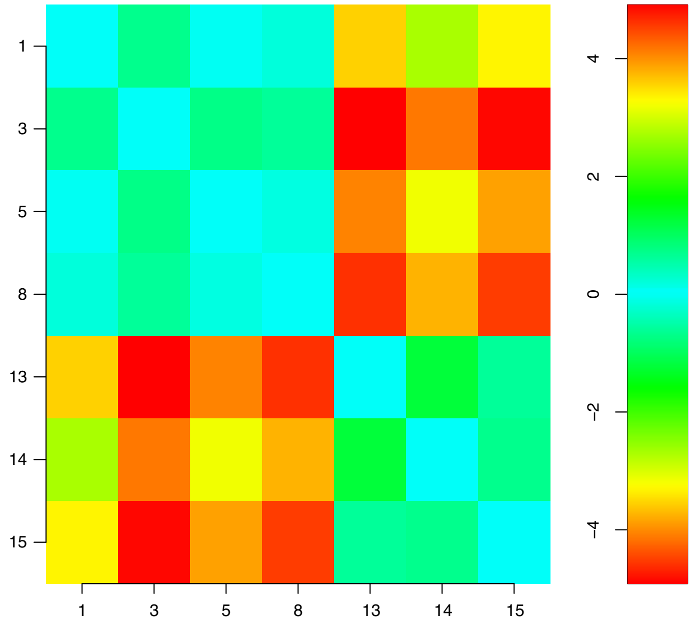


FIGURE 10.
A sub-matrix of $\hat{\mathbf{D}}$ for the digital versus paper-and-pencil example corresponding to items flagged as "DIF items" by a likelihood ratio test with purification.

present data set where there are a number of substantial clusters. Figure 10 shows $\hat{\mathbf{D}}$ for the items flagged as DIF items by the likelihood ratio test. It is seen that many of the "DIF items" have in fact invariant relative difficulties.

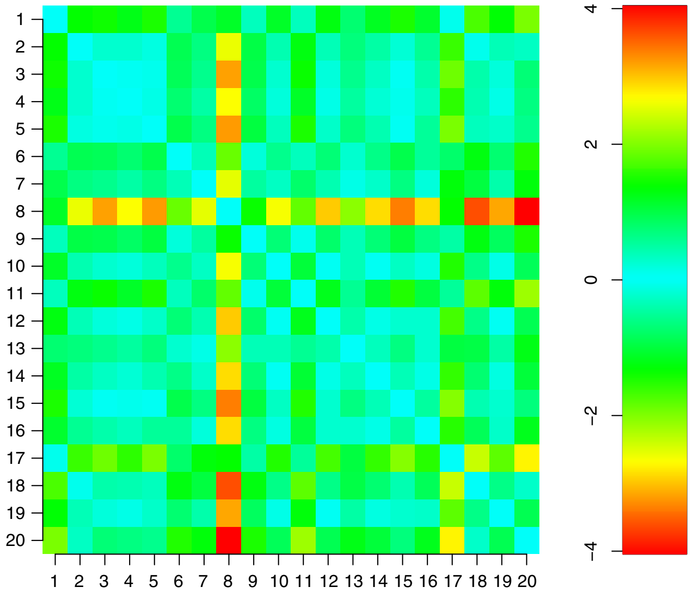


FIGURE 11.
Matrix $\hat{\mathbf{D}}$ for the IQ data example with different age groups.

7.3. *IQ: DIF Between Age Groups*

We continued with the paper-and-pencil version of the test. Inspired by the success of the Rasch model for the 12-year olds, we fitted a Rasch model to the combined data from all the age groups. Quite unexpectedly, a Rasch model did not fit the combined data while it did fit the separate age groups implying that there is DIF between age groups.[7] Our next task is to find out which item pairs show DIF.

Here, we consider samples of 12 ($n = 1142$) and 14 year old children ($n = 673$) and investigate a subset of 20 figure items. Note that these samples are different from those analyzed in the previous application. Each of the samples was randomly split into two samples of roughly equal size: an *investigation* and a *validation* sample. The investigation sample was used to calculate the matrix shown in Figure 11. Visual inspection suggests that item 8 has changed its difficulty relative to all the other items. Using the validation sample, the hypothesis that all items, except item 8, form a cluster was not rejected ($\chi^2_{\Delta\mathbf{R}} = 12.6$, $df = 19$, $p = 0.81$).

When we use the difR package (Magis, Beland, Tuerlinckx, & De Boeck, 2010) to analyze the same data with the MH and Lord's test, both procedures signal item 8 as the DIF item. The MH test also flags item 1, but only in the investigation sample. When item 8 is removed from the data, both tests conclude that there is no DIF. Note that the difR package provides no information about the normalization that is used and it is not possible to change it. It is easy, therefore, to forget that the fact that Lord's test agrees with MH is incidental.

8. Extensions

In this section, we outline how the idea underlying our procedure can be used to develop DIF tests for IRT models other than the Rasch model. This is useful because DIF is detected by virtue

[7] Even the properties of abstract IQ test items can be affected by education and/or aging. Fortunately, it is possible to develop an IQ test without concurrent calibration: A possibility that seems to be widely ignored.

of a fitting IRT model in each group and it is quite a challenge to find real data examples that fit the Rasch model.

8.1. More Complex IRT Models

8.1.1. The PCM: Item Category DIF Consider an item i with $m_i + 1$ response alternatives $j = 0, \ldots, m_i$—one of which is chosen. We obtain a *Partial Credit Model (PCM)* (Masters, 1982) when the probability that $X_{pi} = j$ given that $X_{pi} = j$ or $X_{pi} = j - 1$ is modeled by a Rasch model. That is,

$$P(X_{pi} = j | X_{pi} = j \text{or} X_{pi} = j - 1) = \frac{\exp(\theta_p - \tau_{ij})}{1 + \exp(\theta_p - \tau_{ij})}, \tag{11}$$

where τ_{ij} (where $j = 1, \ldots, m_i$) are thresholds between categories in the sense that $P(X_{pi} = j - 1|\theta) = P(X_{pi} = j|\theta)$, when $\theta = \tau_{ij}$. The PCM appears naturally as a model for "testlets" (Verhelst & Verstralen, 2008; Wilson & Adams, 1995) or rated data (Maris & Bechger, 2007). The PCM reduces to the Rasch model when $m_i = 1$ for all items i.

For identification, one item category parameter may be set to zero which is arbitrary. Similar to the Rasch model, the differences $\tau_{ij} - \tau_{hk}$ between item category parameters are identifiable and independent of the normalization. Thus, the logic of our test extends directly to the PCM. What is different from the Rasch model is that the PCM can have two different forms of DIF. Specifically, Andersen (1977) proposes a parametrization of the PCM with parameters $\eta_{ij} = \sum_{h=1}^{j} \tau_{ij}$ so that $\tau_{ij} = \eta_{ij} - \eta_{i,j-1}$, with $\eta_{i0} = 0$. Thus, we can investigate DIF in the η_{ij} *or* in the τ_{ij}.

8.1.2. The 2PL: DIF in Difficulty and/or Discrimination The Rasch model is a special case of the *two-parameter logistic model (2PL)* (Birnbaum, 1968) which allows items to differ in discrimination. Specifically,

$$P(X_{pi} = 1|\theta, \alpha_i, \delta_i) = \frac{\exp[\alpha_i(\theta_p - \delta_i)]}{1 + \exp[\alpha_i(\theta_p - \delta_i)]}, \tag{12}$$

where α_i denotes the discrimination parameter of item i, and the interpretation of δ_i is the same as in the Rasch model. That the value of the discrimination parameters is unknown implies that both the location of the origin and the size of units of the latent ability are undetermined. This complicates matters because the relative difficulties depend on the unit and the test for the Rasch model does not apply here. Specifically, only relative distances, i.e.,

$$\frac{\delta_i - \delta_j}{\delta_A - \delta_B}, \tag{13}$$

are determined where A and B are two reference items such that $\delta_A \neq \delta_B$. Hence, we now have a pair of reference items. To identify the 2PL, it is enough to fix the values of δ_A and δ_B: One parameter fixes the origin and the other determines the unit $\delta_A - \delta_B$. The values given to δ_A and δ_B are arbitrary as long as $\delta_A \neq \delta_B$. If two items are equally difficult we cannot use them as reference items and we violate the model if we do. It follows that an item pair is or is not functioning differently relative to a reference pair of items. One way to extent the test for the Rasch model to the 2PL is thus to test for differential item pair functioning for each combination of two reference items.

In the 2PL we can also have DIF in discrimination. Ratios of discrimination parameters are invariant under normalization, and DIF in discrimination is again a property of item pairs. The logarithm of the ratio of two discrimination parameters is the difference between the logarithm of the discriminations, a test for DIF in (log-)discrimination is similar to that for difficulty in the Rasch model and again related to the work by Lord (1980, p. 223).[8]

8.1.3. 3PL In the models discussed so far, the only transformation of ability that leaves the measurement model intact was a linear one and the only thing we could do was to change the origin and/or the unit. The three-parameter logistic model (3PL) has recently been shown not to be identified, even when all the parameters of a reference item are fixed (San Martín, González, & Tuerlinckx, 2014). Even if the model can be identified, the situation is likely to be more complex than that of the Rasch model of the 2PL as illustrated by Maris and Bechger (2009) for a 3PL with equal discrimination parameters. As long as we do not know what the identified parameters are for the 3PL we cannot develop a test for DIF.

9. Discussion

The detection of DIF items has become a routine matter. At the same time, it is well known that the standard procedures are flawed, and have consistently failed to give results that can be explained substantively (Angoff, 1993; Engelhard, 1989; Stout, 2002). A review of the existing procedures showed that they differ in the way that DIF is defined but none of them tests a hypothesis that is consistent with the notion that DIF is the property of an item. Lord's test, in particular, was found to be a (valid) test for the hypothesis that *relative* item difficulty is equal across groups—a hypothesis that involves two or more items and not just one. This means that these procedures are not suited to test the hypothesis that an item has DIF, which explains why we get unexpected results when we use them for this purpose. If, for example, we use Lord's test and assume that we test the hypothesis that an item has the same difficulty in two groups (i.e., $\delta_{i,1} = \delta_{i,2}$), we find ourselves in the paradoxical situation where, based on the same evidence, an item can *and* cannot have DIF. Lord's test does gives consistent results if we use it correctly as a test for the hypothesis that the relative difficulty of two items is the same (e.g., $\delta_{i,1} - \delta_{j,1} = \delta_{i,2} - \delta_{j,2}$) and this is essentially what we have done in this paper. Note that color maps are far from ideal when the number of items and groups becomes large; especially, in incomplete designs. It will be clear that further research is needed to improve the way the results are presented.

Our test is still inconsistent with the notion that DIF is the property of an item. Specifically, it provides DIF *pairs* and appears to be of little value to practitioners who look for DIF items. One could, however, ask whether it makes sense to consider DIF a property of an item when an item has no properties. If we follow Lord and say that item *i* has DIF when $\delta_{i,1} \neq \delta_{i,2}$, this only makes sense if the difficulty parameters are identified. Otherwise, we define DIF in terms of the parameters that do no exist; that is, whose value is arbitrary and cannot be estimated. For the Rasch model, it is well known that relative difficulties are identified which is why we defined DIF in terms of the relative difficulties. Like the Rasch model, most IRT models have the property that item difficulty is not identified and, therefore, none of these models permits DIF to be defined as the property of an item. Thus, our long lasting search for DIF items seems to have been led astray because we have been looking for the wrong thing. Time will tell whether we will be more successful if instead of asking why an item has DIF, we ask why certain relative difficulties are

[8] If $\beta_{i,g} = \ln \alpha_{i,g}$, $\ln \frac{\alpha_{i,g}}{\alpha_{j,g}} = \beta_{i,g} - \beta_{j,g}$. An application of the delta method shows that $\text{Cov}(\hat{\beta}_{i,g}, \hat{\beta}_{j,g}) = \text{Cov}(\hat{\alpha}_{i,g}, \hat{\alpha}_{j,g})(\hat{\alpha}_{i,g}\hat{\alpha}_{j,g})^{-1}$.

different. We believe that in many cases, the answer will be a difference in the education received by different groups.

The main message of this paper is that DIF can only meaningfully be defined in terms of the identified parameters. While the focus of the paper was on DIF tests for the Rasch model, we have outlined how this principle can be applied to construct DIF tests for other, more complex, IRT models. In general, if we impose a normalization we choose an identified parametrization of the model and a statement about DIF is "meaningful" if it is independent of the parametrization. Also, in general, the identified parameters are functions of the distribution of the data. We can investigate DIF in terms of the identified parameters, or one–one functions of these parameters, but in each case we must make explicit what it is that we are looking at. For the PCM, we found that DIF can be defined in terms of the relative difficulty of the item categories. Hence, this model presents no additional complications beyond the Rasch model, except that we can investigate DIF in different parametrizations. For the 2PL, the presence of a discrimination parameter meant that DIF could no longer be defined in terms of relative difficulties but in ratios of relative difficulties: Hence, we would report DIF triplets. The 2PL illustrates that it need not be trivial to determine which parameters are identified and allow meaningful statements about DIF. While the identification of the Rasch model is well studied, this knowledge is incomplete or missing for the other models. This is true in particular for the 3PL, and the multidimensional IRT models. In order to study DIF with these models we must study their identifiability.

In closing, we would like to reiterate the point that DIF is defined by virtue of a fitting measurement model in each group: here, the same kind of model. If the model does not fit this may lead to the erroneous conclusion that there is DIF, as shown by Ackerman (1992), Bold (2002), DeMars (2010, Issue 2), or Thurman (2009). This means that we are on safer ground if we use conditional maximum likelihood and avoid the specification of the distribution of ability. Misspecification of the ability distributions may cause bias in the estimates of the item parameters (e.g., Zwitser & Maris, in press) which may give the appearance of DIF. An additional complication is that the identifiability of marginal models may be (much) more complex (San Martin & Roulin, 2013).

Acknowledgments

The authors wish to thank Norman Verhelst, Cees Glas, Paul De Boeck, Don Mellenbergh, and our former student Maarten Marsman for helpful comments and discussions.

Appendix

Here is some R-code that can be used after estimates and their asymptotic variance–covariance matrix are available. It assumes two groups, nI items, and a normalization of the form $\delta_{r,1} = \delta_{r,2} = 0$.

```
        ### Overall Test for DIF
beta=par2-par1
Sigma=cov1+cov2
DIF_test=mahalanobis(beta[-r],rep(0,(nI-1)),Sigma[-r,-r])
DIF_p=1-pchisq(DIF_test,nI-1)

        ### item pair DIF
DeltaR=kronecker(par2,t(par2),FUN="-")
 -kronecker(par1,t(par1),FUN="-")
```

```
var1=diag(cov1)
var2=diag(cov2)
S=(kronecker(var1,t(var1),FUN="+")-2*cov1)
 +(kronecker(var2,t(var2),FUN="+")-2*cov2)
diag(S)=1
D=DeltaR/sqrt(S)
```

par1 and par2 are vectors with parameter estimates: the r entry equal to zero. The variance covariance matrices are called cov1 and cov2. Their rth row and column are zero.

References

Ackerman, T. (1992). A didactic explanation of item bias, impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Andersen, E. B. (1972). The numerical solution to a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, *34*, 283–301.

Andersen, E. B. (1973b). A goodness of fit test for the rasch model. *Psychometrika*, *38*, 123–140.

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

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

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bechger, T. M., Maris, G., & Verstralen, H. H. F. M. (2010). *A different view on DIF* (R&D Report No. 2010–5). Arnhem: Cito.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.

Bold, D. M. (2002). A monte carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113–141.

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* (pp. 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.

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

DeMars, C. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*, 961972.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (1989). Accuracy of bias review judges in identifying teacher certification tests. *Applied Measurement in Education*, *3*, 347–360.

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

Fischer, G. H. (1974). *Einfuhrung in die theorie psychologischer tests*. Bern: Verlag Hans Huber. (Introduction to the theory of psychological tests.).

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

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics*, *8*, 261–263.

Gierl, M., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*, 241–264.

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

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

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

Glas, C. A. W., & Verhelst, N. D. (1995b). Tests of fit for polytomous Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 325–352). New-York: Spinger.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124.

Hessen, D. J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika*, *70*, 497–516.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedur. *Scandinavian Journal of Statistics*, *6*, 65–70.

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

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Hillsdale, NJ: Lawrence Earl- baum Associates.

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

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichtomous differential item functioning. *Behavior Research Methods*, *42*, 847–862.

Magis, D., & Boeck, P. De. (2012). A robust outlier approach to prevent type i error inflation in differential item functioning. *Educational and Psychological Measurement*, *72*, 291–311.

Magis, D., & Facon, B. (2013). Item purification does not always improve DIF detection: A coun- terexample with Angoff's Delta plot. *Journal of Educational and Psychological Measurement*, *73*, 293–311.

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

Maris, G., & Bechger, T. M. (2007). Scoring open ended questions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 663–680). Amsterdam: Elsevier.

Maris, G., & Bechger, T. M. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement*, *7*, 75–88.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–175.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychomettrika*, *57*, 289–311.

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

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *66*, 437–460.

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

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

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).

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

San Martin, E., González, J., & Tuerlinckx, F. (2009). Identified parameters, parameters of interest and their relationships. *Measurement*, *7*, 97–105.

San Martin, E., & Quintana, F. (2002). Consistency and identifiability revisited. *Brazilian Journal of Probability and Statistics*, *16*, 99–106.

San Martin, E., & Roulin, J. M. (2013). Identifiability of parametric Rasch-type models. *Journal of statistical planning and inference*, *143*, 116–130.

San Martín, E., González, J., & Tuerlinckx, F. (2014). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *78*(2), 341–379.

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.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*, 485–518.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (mm-irt-c) to ascertain observed and unobserved measurement equivalence. *Organizational Research Method*, *14*, 147–155.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care*, *44*, 152–170.

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., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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* (pp. 67–114). Hillsdale, NJ: Lawrence Earlbaum.

Thurman, C. J. (2009). *A monte carlo study investigating the in uence of item discrimination, category intersection parameters, and differential item functioning in polytomous items*. Unpublished doctoral dissertation, Georgia State University.

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

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

Verhelst, N. D. (2008). An effcient MCMC-algorithm to sample binary matrices with fixed marginals. *Psychometrika*, *73*, 705–728.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peiling- sonderzoek* (PPON-rapport No. 4). Arnhem: CITO.

Verhelst, N. D., Glas, C. A. W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly*, *1*(3), 245–262.

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

Verhelst, N. D., Hartzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software*, *20*, 1–14.

Verhelst, N. D., & Verstralen, H. H. F. M. (2008). Some considerations on the partial credit model. *Psicológica*, *29*, 229–254.

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

Wald, A. (1943). tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

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., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *74*, 1–22.

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.

Wilson, M., & Adams, R. (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181–198.

Wright, B. D., Mead, R., & Draba, R. (1976). *Detecting and correcting item bias with a logistic response model*. (Tech. Rep.). Department of Education, Statistical Laboratory: University of Chicago.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, *19*, 369–375.

Zwitser, R., & Maris, G. (2013). Conditional statistical inference with multistage testing designs. *Psychometrika*. doi:10.1007/s11336-013-9369-6.

Manuscript Received: 3 SEP 2012
Accepted: 20 FEB 2014
Published Online Date: 16 SEP 2014