Educational and Psychological
Measurement
71(6) 1023–1046
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164411400734
http://epm.sagepub.com

Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel–Haenszel Procedure in Short Test and Small Sample Conditions

Insu Paek[1] and Mark Wilson[2]

Abstract

This study elaborates the Rasch differential item functioning (DIF) model formulation under the marginal maximum likelihood estimation context. Also, the Rasch DIF model performance was examined and compared with the Mantel–Haenszel (MH) procedure in small sample and short test length conditions through simulations. The theoretically known relationship of the DIF estimators between the Rasch DIF model and the MH procedure was confirmed. In general, the MH method showed a conservative tendency for DIF detection rates compared with the Rasch DIF model approach. When there is DIF, the *z* test (when the standard error of the DIF estimator is estimated properly) and the likelihood ratio test in the Rasch DIF model approach showed higher DIF detection rates than the MH chi-square test for sample sizes of 100 to 300 per group and test lengths ranging from 4 to 39. In addition, this study discusses proposed Rasch DIF classification rules that accommodate statistical inference on the direction of DIF.

[1] Florida State University, Tallahassee, FL, USA
[2] University of California, Berkeley, Berkeley, CA, USA

Corresponding Author:
Insu Paek, Educational Psychology and Learning Systems, Florida State University, 3204D Stone Building,
1114 W. Call Street, Tallahassee, FL 32306-4453, USA
Email: ipaek@fsu.edu

Keywords

DIF, Mantel–Haenszel test, Rasch DIF classifications, Rasch model

Differential item functioning (DIF) is one of the most active research areas in psychological and educational measurement (Camilli & Shepard, 1994; Holland & Wainer, 1993; Mapuranga, Dorans, & Middleton, 2008; Millsap & Everson, 1993; Osterlind & Everson, 2009). A popular definition of the No DIF condition for dichotomously scored item (Lord, 1980; Meredith & Millsap, 1992) is

$$P(X=1, G|\theta) = P(X=1|\theta)P(G|\theta)$$

or equivalently,

$$P(X=1|\theta, G) = P(X=1|\theta), \forall\theta, \tag{1}$$

where $P(\cdot)$ is a probability, X is an item indicator variable (1 for *correct* and 0 for *incorrect*), G is a group indicator variable, and θ is a latent ability. The No DIF definition above states that the probability of a correct response at the same ability should be the same regardless of G. Item response theory (IRT) models have attracted attention from researchers because they can test the null hypothesis of No DIF above directly using item response function (IRF) differences. Popular parametric IRF forms in IRT include the Rasch (Rasch, 1980), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) models (Birnbaum, 1968). Proposed statistical tests to directly evaluate the No DIF hypothesis in IRT were the use of the Wald test (Lord, 1977, 1980), the likelihood ratio (LR) test (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), the score test (or Lagrange multiplier test; Glas, 1998), and simultaneous item bias test (SIBTEST; Shealy & Stout, 1993). The first three are well-known large sample tests in the maximum likelihood estimation approach, and they are appropriate for the parametric IRT models. The last, SIBTEST, is a nonparametric IRT approach in which the IRFs do not assume a particular parametric form.

This article focuses on the use of Rasch modeling in DIF investigation. We show formulations of the Rasch DIF modeling in detail in the marginal maximum likelihood (MML) estimation context and discuss its relationship with one of the most widely used DIF methods, the Mantel–Haenszel (MH) procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959) and with Educational Testing Service (ETS) DIF categories. We also discuss DIF classification rules for the Rasch DIF model. In addition, we investigate the performance of two types of DIF tests (the Wald test, which we call z test here, and the LR test) under the Rasch DIF model, comparing them with the MH procedure through simulations. Our simulation and elaboration of the Rasch DIF model formulation and discussion of its relationship with the MH method facilitate understanding, both theoretically and empirically, of the behavior of both procedures with regard to DIF detection. For the simulation part of the study, our interest lies especially in the small sample and short test length situations. Because of the relationship known in the literature between the Rasch model and the MH method, which is elaborated in a later section, and because their statistical tests are asymptotic-based, a large sample size with a long test length is theoretically expected to display the same behavior in the two approaches in terms of DIF detection and recovery of the relationship between DIF estimators. However, the Rasch model is said to have the potential to work efficiently with a small sample size because of its concise parameterization for item response behavior. Hence, the study of the impact of small sample sizes with short test lengths (although applications for DIF investigation with such occasions may not be common) is not only of theoretical interest, in that the results can show how robust the two procedures are and how well the theoretical relationship holds between the Rasch DIF model and the MH procedure under these conditions, but also has some practical value for practitioners who are faced with the challenging situation of a small sample size with a short test.

Formulation of the Rasch DIF Model

The general form of the Rasch DIF model can be expressed as follows for a response of 1 using a logit link function:

$$\text{logit}\{P(x_{ni} = 1|\theta_n, g)\} = \theta_n - \delta_i + \gamma_i G, \tag{2}$$

where x_{ni} is the score of person n for item i (1 = *correct*, 0 = *incorrect*; $i = 1, 2, \ldots, I$; $n = 1, 2, \ldots, N$), θ_n is a scalar proficiency (ability) parameter for person n, δ_i is the item difficulty parameter for item i, γ_i is the DIF index parameter for item i, and g indicates either the reference group or the focal group, $G = 1$ if $g = R$ (reference group), $G = 0$ if $g = F$ (focal group). This is equivalent to $P(x_{ni} = 1|\theta_n, g) = \exp(\eta)/\{1 + \exp(\eta)\}$, where $\eta = \theta_n - \delta_i + \gamma_i G$. Under this model presentation, γ is the item difficulty difference between the focal group and the reference group (i.e., $\gamma = \delta_F - \delta_R$). Because of the dependency induced by the nested structure (i.e., items nested within person), or equivalently to model the repeated measurement design for a given person, θ_n is considered a random parameter, and it is assumed that $\theta_n|g \overset{iid}{\sim} N(\mu_g, \sigma)$.[1] Note that this population modeling takes care of the overall group ability differences, called "impact," which should not be confounded with DIF. This impact modeling can be specified by latent regression of θ onto the group indicator variable.

$$\theta = \beta_0 + \beta_1 Y + \varepsilon, \tag{3}$$

where Y is the group indicator variable that is the same as G in Equation (2), the βs are scalar regression coefficient parameters, and $\varepsilon \overset{iid}{\sim} N(0, \sigma)$. The person notation n is suppressed in the above and the following equations since they are all formulated for person n. For the identification of the DIF model above and to make sure a common comparable scale is established between compared groups, $\beta_0 = 0$ and the number of elements in $\{\gamma_i\} < I$ (i.e., the number of free γ parameters less than the total number of items in the test) were used (see Wang, 2004, for the effects of different types of constraints on the DIF modeling). From the perspective of IRT as being a multilevel model (e.g., Adams, Wilson, & Wu, 1997; Kamata, 2001; Rijimen, Tuerlinckx, De Beock, & Kuppens, 2003; Skrondal, & Rabe-Hesketh, 2004; Van den Noortgate, De Boeck, & Meulders, 2003; Van den Noortgate & Paek, 2004), Equations (2) and (3) represent a random intercept model.

This hierarchical structure for DIF modeling can be easily embedded as well in at least two of the general Rasch family IRT models. One is the multidimensional random coefficient multinomial logit model (MRCLM; Adams, Wilson, & Wang, 1997), and the other is linear logistic test model (LLTM; Fischer, 1973, 1983). The unidimensional version of the MRCLM in the log of odds form is

$$\log\left[\frac{P(X_{ik}=1|\theta)}{P(X_{i(k-1)}=1|\theta)}\right] = (b_{ik} - b_{i(k-1)})\theta + (\mathbf{a}'_{ik} - \mathbf{a}'_{i(k-1)})\boldsymbol{\xi}, \tag{4}$$

where X_{ik} = 1 if the response to item i is in category k and 0 otherwise ($1 \leq i \leq I$, and $1 \leq k \leq K_i$), θ is a scalar person ability random parameter that follows a density function $g(\theta; \boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is a vector of person distribution parameters, b_{ik} is a predetermined item score for category k of item i, $\boldsymbol{\xi}$ is a $p \times 1$ fixed unknown item parameter vector, and $\mathbf{a}'_{ik}$ is a $1 \times p$ vector to specify a linear combination of p elements of $\boldsymbol{\xi}$ for each response category. For K_i = 2, let $b_{ik=2}$ = 1 and $b_{ik=1}$ = 0, which is a typical dichotomous item scoring. In addition, let $\mathbf{a}'_{ik=1} = \mathbf{0}_{1\times(I+1)}$, $\mathbf{a}'_{ik=2} = [\lambda_{1m}]_{1\times(I+1)}$, $\boldsymbol{\xi}' = [\delta_1, \delta_2, \ldots, \delta_i, \ldots, \delta_I, \gamma_i]_{(I+1)\times 1}$, and $g(\theta; \boldsymbol{\alpha}) = N(\mathbf{Y}'\, \boldsymbol{\beta}, \sigma^2)$, where $\mathbf{0}$ is a null vector serving as a reference category; $\lambda_{1m} = -1$ when $m = i$, $\lambda_{1m} = G$ when $m = I + 1$, and λ_{1m} = 0 otherwise; $\mathbf{Y}' = [1, Y]$ and $\boldsymbol{\beta}$ is a 2×1 regression coefficient vector. Under these specifications, Equation (4) with $g(\theta; \boldsymbol{\alpha}) = N(\mathbf{Y}'\, \boldsymbol{\beta}, \sigma^2)$ becomes the DIF model formulated as in Equations (2) and (3).

LLTM incorporates a linear regression structure into item parameters. For DIF modeling in the LLTM, item difficulty is partitioned into three parts: one is the common item difficulty (for both focal and reference groups), another is an impact parameter that models group mean, and the third is a parameter for DIF. With the same assumptions specified for the Rasch DIF model and $\mu_g = \mu + \Delta G'$, where $G' = -1$ if g = focal group and G' = 1 if g = reference group, it can be shown that the DIF modeling under the LLTM context has the following form (see Appendix A):

$$\text{logit}\{P(x_i = 1|\theta^*, g)\} = \theta^* - \delta_i + \Delta G' + \gamma_i G \tag{5}$$

with $\theta^* \overset{iid}{\sim} N(\mu, \sigma)$. For the model identification and to make sure a common comparable scale, μ = 0 and the number of elements in $\{\gamma_i\} < I$ (i.e., the number of free γ parameters less than the total number of items in the test) can be used. The person ability and its distribution in Equation (5) are not identical as in Equation (2), but Equation (5) models the ''impact'' factor by Δ and has the same item difficulty (δ_i), DIF index parameter (γ_i), and σ in the person ability distribution. This confirms that when the person ability distribution follows a normal distribution (with homogeneous variance), the LLTM DIF modeling using the conditional maximum likelihood (CML) estimation works the same way as the Rasch DIF model by the MML estimation with the normal distribution assumption.[2]

The Rasch DIF model has a DIF effect size measure of γ, offering two types of DIF test statistics. One is the *z* test that is the ratio of $\hat{\gamma}$ to the standard error (*SE*) of it, and the other is the LR test. By incorporating the DIF size estimator γ in the IRF, the Rasch DIF model estimates γ and $SE(\gamma)$ directly from the data, so that the *z* test can be carried out with ease.[3] The LR test statistic, sometimes called the ''deviance,'' is defined as $D = -2 \log$ (likelihood of nested model/likelihood of full model). *D* follows a chi-square distribution with degrees of freedom equal to the difference between the number of parameters in the full and the nested models. The full model (DIF model) consists of Equations (2) and (3). The nested model (No DIF model) consists of Equation (2) without the term γ and Equation (3). Note that a separate equating step to put item estimates on the same scale to calculate the DIF effect size is not needed in this Rasch DIF model formulation.

To estimate the Rasch DIF model, statistical packages equipped with a mixed model estimation routine or IRT models with the functionality for flexible user-specified IRF parameterization can be used. Such programs include, for example, HLM (Bryk, Raudenbush, & Congdon, 1996), SAS NL MIXED procedure (SAS Institute, 1999), or ConQuest (Wu, Adams, & Wilson, 1998). See also, for example, Kamata (2001) and De Boeck and Wilson (2004) on how to use a mixed model framework to estimate IRT models. We used the program ConQuest in this study, and we have provided an appendix for practitioners that shows in detail how to specify and estimate the Rasch DIF model in using ConQuest (see Appendix B).

The Relationship Between the Rasch DIF Model and the MH Procedure

Mellenbergh (1982) called the constant statistical relationship between item response (*X*) and group (*G*) for all levels of a matching variable ''uniform DIF.'' Following Mellenbergh's definition of uniform DIF, Hanson (1998) differentiated uniform DIF from ''parallel DIF'' caused by IRT IRF difficulty parameter differences only. The uniform DIF was defined as

$$\alpha(\theta) = \frac{P_R(\theta)[1 - P_F(\theta)]}{P_F(\theta)[1 - P_R(\theta)]} = \alpha \neq 1,$$

where $P(\theta) \equiv P(X = 1|\theta)$. He showed that in the case of the 3PL model with the parallel DIF,

$$\alpha(\theta) = \frac{g + e^{a(\theta - \delta_F)}}{g + e^{a(\theta - \delta_R)}},$$

where *g* is a guessing parameter and *a* is a slope parameter, indicating the parallel DIF in the 3PL model is not the same as the uniform DIF definition. In contrast, parallel

DIF with the 2PL and the Rasch models result in $\alpha(\theta) = e^{a(\delta_F - \delta_R)}$ and $\alpha(\theta) = e^{(\delta_F - \delta_R)}$, respectively, satisfying the uniform DIF definition. The MH procedure estimates a common odds ratio α_{MH} as a DIF effect size measure typically conditioning on an observed number-right score $Z = \sum X$. For the Rasch model, Z is a sufficient statistic for θ, so $\alpha(\theta) = \alpha(Z) = \alpha$ and $\alpha_{MH} = \alpha = e^{(\delta_F - \delta_R)}$. The common odds ratio α_{MH} is sometimes transformed by $-2.35\ln(\alpha_{MH})$ to put it into ETS difficulty scale called the ‘‘delta scale’’ for practitioners who use the ETS DIF classification rules (Dorans & Holland, 1993, p. 41). Thus, we have

$$\Delta_{MH} = -2.35\ln(\alpha_{MH}) = -2.35(\delta_F - \delta_R) = -2.35\gamma \tag{6}$$

Equation (6) shows that Δ_{MH} is linearly related to γ. Holland and Thayer (1988) also derived $\alpha_{MH} = e^{(\delta_F - \delta_R)}$ in the Rasch model to show that testing H_0: $P(X = 1|\theta, G) = P(X = 1|\theta)$ in the Rasch model is equivalent to testing H_0: $\alpha_{MH} = 1$ in the MH procedure, indicating that observed score matching with the MH procedure can be the same as the unobserved latent variable IRT approach.[4] They stated that the exact conditions needed to hold for $\alpha_{MH} = e^{(\delta_F - \delta_R)}$ or Equation (6) are as follows: (a) the Rasch model fits the data, (b) all items are DIF-free except the studied item that may exhibit DIF, (c) the matching variable is number-right score that includes the studied item, and (d) the data are random samples from the reference and the focal group. Linacre and Wright (1987, 1989) also showed $\ln(\alpha_{MH})$ is equivalent to the Rasch item difficulty difference, $\delta_F - \delta_R$, estimated by the noniterative normal approximation algorithm (PROX; Wright & Stone, 1979).

Based on Equation (6), the Rasch difficulty difference DIF classification rules can be constructed that connect with the ETS DIF classification rules (Longford, Holland, & Thayer, 1993, p. 175), which state

- A (negligible DIF) if statistically not significant for H_0: $\Delta_{MH} = 0$ below .05 level or if $|\Delta_{MH}| < 1$;
- C (large DIF) if statistically significant for H_0: $|\Delta_{MH}| \leq 1$ below .05 level and if $|\Delta_{MH}| \geq 1.5$; and
- B (medium DIF) otherwise.

Equation (6) gives 0.426 and 0.638 for $\Delta_{MH} = -1$ and $\Delta_{MH} = -1.5$, respectively, with regard to γ. Thus, the exact γ DIF classification rule set that translates the above ETS DIF classification rules is as follows:

- A if $|\gamma| \leq 0.426$ or if H_0: $\gamma = 0$ is not rejected below .05 level;
- C if $|\gamma| \geq 0.638$ and if H_0: $|\gamma| \leq 0.426$ is rejected below .05 level; and
- B otherwise.

An alternative simpler γ DIF classification rule set (Paek, 2002) is

- A if $|\gamma| \leq 0.426$ or if H_0: $\gamma = 0$ is not rejected below .05 level;
- B if $0.426 \leq |\gamma| < 0.638$ and if H_0: $\gamma = 0$ is rejected below .05 level; and
- C if $|\gamma| \geq 0.638$ and if H_0: $\gamma = 0$ is rejected below .05 level.

Under this alternative γ DIF classification rule set, a simple null hypothesis testing of H_0: $\gamma = 0$ is performed only one time for a studied item, whereas the exact γ DIF classification rule set involves testing a simple null hypothesis H_0: $\gamma = 0$ and a composite null hypothesis H_0: $|\gamma| \leq 0.426$ for the decision of the DIF category. We will discuss some refinement of the alternative γ DIF classification rule set in the last section of this article. The exact γ DIF classification rule set and the alternative γ DIF classification rule set will not always lead to the same DIF category decisions as the ETS DIF classification rules with the MH procedure even if the conditions for $\alpha_{MH} = e^{(\delta_F - \delta_R)}$ hold, because the statistical tests used in the Rasch DIF model and the MH procedure are not the same, and may function differently for finite (especially small) samples although they are expected to perform almost the same with large sample and long test conditions; second, the alternative γ DIF classification rule set does not involve the composite null hypothesis testing that corresponds to H_0: $|\Delta_{MH}| \leq 1$ for large C DIF in the ETS categories.

Roussos, Schnipke, and Pashley (1999) showed, under the 3PL model having parallel DIF, the tendency of the MH delta (Δ_{MH}) to be squashed toward zero when the studied item has high difficulty. When the Rasch model holds, it brings stable behavior to the MH procedure. The relationship $\alpha_{MH} = e^{(\delta_F - \delta_R)}$ implies that the location of the item difficulty parameter does not have any influence on the Δ_{MH}. There is no need to be concerned about whether one is testing a hard or an easy item for DIF nor are there differences between parallel and uniform DIF.

Simulation Study for Small Sample and Short Test Length Conditions

There are several studies that examined the performance of the MH procedure under small-sample conditions (e.g., Fidalgo, Ferreres, & Muniz, 2004; Mazor, Clauser, & Hambleton, 1992; Muniz, Hambleton, & Xing, 2001; Parshall & Miller, 1995; Schulz, Perlman, Rice, & Wright, 1996), but these studies mainly varied the sample sizes with a relatively long test. Donoghue, Holland, and Thayer (1993) conducted simulation studies with a variety of conditions, one of which is the variation of the test length (e.g., 4-, 9-, 19-, and 39-item tests), but they used large sample sizes (2,000 and 500 for the reference and focal groups). Also, all these above-cited studies used the MH procedure only for DIF investigation without comparing with the IRT DIF methods. The simulations in the present study compare the Rasch DIF model approach with the MH method in the small sample and the short test length conditions.

This study used the same test lengths used by Donoghue et al. (1993) and their true IRF difficulty parameters with the Rasch IRF. (The test length of 39 is obviously not a short test length, but the same or longer test length is frequently observed in educational achievement testings, so it was included for comparison purpose.) The sample sizes in this study were (100, 100), (200, 200), and (300, 300), where the numbers in the parentheses are the reference and the focal group sample sizes, respectively. When sample sizes become large (e.g., larger than 300 per group) with longer test lengths such as 39 or higher, the theoretical relationship, $\Delta_{MH} = -2.35\gamma$ is expected to hold, but with these small finite samples and short test length conditions, DIF detection rates and the extent to which the theoretical relationship holds between the Rasch DIF model and the MH procedure are not clearly known.

Group ability differences between the reference and focal group were specified by 1 standard deviation difference with regard to the ability means, that is, $\theta \sim N(0, 1)$ and $N(-1, 1)$ for the reference and the focal groups, respectively.

In each test length, one item was the studied item with a possible parallel DIF, whereas the remaining items had No DIF. Note that this scenario is equivalent to assuming that a set of suitable items for ability matching is available. This situation is typically observed in practice in a pretest item analysis where each pilot item is investigated one by one for DIF with the existing operational test as an ability matching item set. When a matching item set is internally sought within a test, an iterative refinement procedure can be used to construct a matching item set (e.g., Holland & Thayer, 1988, p. 141; Wang, 2008).

For DIF conditions, $\gamma = 0.468$ or 0.681 [$\delta_F = \delta_R + 0.468$ (or 0.681) with $\delta_R = -.85$] and for No DIF, $\gamma = 0$ ($\delta_F = \delta_R = -.85$).[5] The γ values in the DIF conditions correspond to $\Delta_{MH} = 1.1$ and 1.6, respectively, which are slightly higher than the threshold values in medium and large DIF classifications. The choice of γ was based on the fact that the medium or large DIF items are the items of greatest concern in practice that are considered for removal from a test. The ultimate interest for the Rasch DIF model in this simulation is its ability to screen out such medium and large DIF items in a small sample with possible short test conditions.

Three DIF conditions ($\gamma = 0$, 0.468, and 0.681), four test length conditions (4, 9, 19, and 39), and three sample size per group condition (100, 200, and 300) gave a total of $3 \times 4 \times 3 = 36$ data simulation conditions. In each condition, 400 replications were made, and the two tests from the Rasch DIF modeling (*z* test and LR test) and the MH chi-square test were applied to each replication for statistical detection of DIF.

The item response data with DIF and No DIF were generated following the standard IRT data generation procedure: (a) generate predicted IRF values using Equation (2) with the true item parameters and θs; (b) compare randomly generated standard uniform distribution quantiles with the predicted IRF values; (c) assign 1 if the IRF value is less than or equal to the value from the standard uniform distribution; assign 0 otherwise.

Results

The recovery of true DIF sizes by both the Rasch DIF model and the MH procedure was good for all conditions in general. Table 1 presents summary of DIF estimates from the Rasch DIF model and its comparison with the MH DIF estimates.

Bias of the Rasch DIF model γ ranged from -0.03 to 0.02 across all conditions; and the mean squared error of γ ranged from 0.03 to 0.16. (When the smallest sample size of 100 per group is combined with the smallest test length of 4, the mean squared error ranged from .10 to .16. Except in this most extreme condition, the mean squared

Table 1. DIF Estimate Comparisons

Test length	Sample size per group	Mean of $\hat{\gamma}$	Mean of $-2.35\hat{\gamma}$	Mean of $\hat{\Delta}_{MH}$
No DIF (γ = 0 or Δ_{MH} = 0)				
4	100	−0.01	0.02	0.02
	200	−0.03	0.07	0.08
	300	−0.01	0.03	0.03
9	100	0.01	−0.02	−0.01
	200	0.00	−0.01	0.00
	300	0.01	−0.01	−0.01
19	100	0.02	−0.05	−0.04
	200	−0.02	0.05	0.05
	300	−0.01	0.03	0.04
39	100	0.02	−0.06	−0.02
	200	0.00	0.00	−0.01
	300	0.00	0.00	0.01
DIF (γ = 0.468 or Δ_{MH} = -1.1)				
4	100	0.48	−1.14	−1.12
	200	0.48	−1.12	−1.11
	300	0.47	−1.11	−1.11
9	100	0.48	−1.13	−1.12
	200	0.47	−1.11	−1.11
	300	0.48	−1.12	−1.10
19	100	0.47	−1.10	−1.08
	200	0.49	−1.15	−1.14
	300	0.47	−1.10	−1.09
39	100	0.48	−1.14	−1.14
	200	0.48	−1.13	−1.12
	300	0.47	−1.12	−1.11
DIF (γ = 0.681 or Δ_{MH} = -1.6)				
4	100	0.70	−1.63	−1.65
	200	0.70	−1.66	−1.66
	300	0.70	−1.64	−1.64
9	100	0.69	−1.62	−1.63
	200	0.71	−1.66	−1.65
	300	0.67	−1.58	−1.57
19	100	0.68	−1.60	−1.60
	200	0.68	−1.61	−1.59
	300	0.67	−1.58	−1.58
39	100	0.70	−1.65	−1.63
	200	0.69	−1.63	−1.63
	300	0.68	−1.60	−1.58

Note. DIF = differential item functioning; MH = Mantel–Haenszel.

error ranged from .03 to .08.) When the $-2.35\hat{\gamma}$ transformation was compared with the estimated MH delta values ($\hat{\Delta}_{MH}$), their differences ranged from -0.04 to 0.01 with a mean of -0.01. Even with such a small sample of 100 per group and the test length of 4, the delta values calculated from the Rasch DIF estimates matched the delta values from the MH procedure effectively.

Statistical rejection rates from the Rasch DIF model and the MH procedure are shown in Table 2. When there is DIF, increasing sample size brought more statistical rejections across different tests. Test length impact on DIF detection when there is DIF is not as clear as the sample size, although there seems to be a slight hint that there is an approximate tendency for increasing test length to entail higher DIF detection rates (see, e.g., the bold type rejection rates in Table 2 for this overall pattern), but the effect of increasing test length is minimal when both sample size and DIF size are large.

The third column in Table 2 is the rejection rate from $\hat{\gamma}/SE(\hat{\gamma})$, which is the *z* test. As can be seen, the *z* test showed highly inflated Type I error when there is No DIF ($\gamma = 0$). The *z* test had maximum of about 4.5 times as large as the .05 level when there is No DIF. Because the γ value is well estimated, as seen in Table 1, we suspected that $SE(\hat{\gamma})$ was underestimated. The standard deviation of $\hat{\gamma}$, $SD(\hat{\gamma})$, for each condition was calculated and compared with $SE(\hat{\gamma})$. The comparison made in Table 3 shows that the $SE(\hat{\gamma})$ is about 26% to 43% downwardly biased[6] when compared with $SD(\hat{\gamma})$.

To allow the investigation of the *z*-test performance when the $SE(\hat{\gamma})$ does not have such systematic bias, the adjusted *z* test was calculated where $SE(\hat{\gamma})$ in the *z* test was replaced by $SD(\hat{\gamma})$. The fourth column in Table 2 represents the rejection rate from the adjusted *z* test. The adjusted *z* test, LR test, and MH chi-square test rejection rates were close to or less than the .05 level when there is No DIF. To further facilitate interpretation of the rejection rate results, differences between rejection rates when the LR test was the base line were calculated and plotted in Figures 1, 2, and 3. Note that the desired results would be a line through zero at the *y*-axis when all tests perform closely.

To ensure fair comparisons when there is DIF, the Type I error rates should be at least approximately similar. The minimum and maximum of the rejection rates when there is No DIF were 0.14 and 0.23 (with an average of 0.19) for the *z* test, 0.04 and 0.06 (with an average of 0.05) for the adjusted *z* test, 0.04 and 0.07 (with an average of 0.05) for the LR test, and 0.02 and 0.05 (with an average of 0.04) for the MH chi-square test. (Except the 19-item test with sample size of 100 per group, which had the rejection rate of 0.02 when there is No DIF, all other condition rejection rates were 0.03-0.05 when there is No DIF for the MH chi-square test.) The *z* test was not comparable with other test results because of its liberal nature due to the underestimation of $SE(\hat{\gamma})$ (The *z*-test results in Figures 2 and 3 when there is DIF were displayed for completeness of the results presentation rather than for comparison purposes.) Figure 1 shows that the *z* test is too liberal. The MH chi-square test, when compared with the LR or the adjusted *z* test, showed a conservative pattern in the Type I error rates and the DIF detection rates when there is DIF, which is shown as the departure of the MH test results in a negative direction from the LR test results in Figures 1, 2, and 3. The average DIF detection rates when $\gamma = 0.468$ (or $\Delta_{MH} = -1.1$) were 0.47, 0.48, and 0.39 for the adjusted *z*, LR, and MH chi-square tests, respectively. The average DIF detection rates when $\gamma = 0.681$ (or $\Delta_{MH} = -1.6$) were 0.72, 0.74, and 0.66 for the adjusted *z*, LR, and MH chi-square tests, respectively.

Table 2. Statistical Rejection Rate

Test length	Sample Size per group	*z* test	Adjusted *z* test	Likelihood ratio test	MH chi-square test
No DIF (γ = 0 or Δ_{MH} = 0)					
4	100	0.22	0.06	0.06	0.03
	200	0.23	0.04	0.04	0.04
	300	0.21	0.04	0.04	0.04
9	100	0.20	0.06	0.07	0.03
	200	0.20	0.06	0.06	0.04
	300	0.19	0.05	0.05	0.04
19	100	0.14	0.05	0.04	0.02
	200	0.17	0.06	0.06	0.04
	300	0.16	0.06	0.05	0.04
39	100	0.21	0.05	0.06	0.03
	200	0.17	0.06	0.06	0.04
	300	0.16	0.05	0.05	0.05
DIF (γ = 0.468 or Δ_{MH} = $-$1.1)					
4	100	**0.52**	**0.25**	**0.23**	**0.16**
	200	0.69	0.39	0.44	0.37
	300	0.83	0.53	0.56	0.51
9	100	**0.52**	**0.27**	**0.27**	**0.19**
	200	0.75	0.46	0.48	0.38
	300	0.89	0.68	0.66	0.58
19	100	**0.51**	**0.31**	**0.27**	**0.19**
	200	0.75	0.54	0.57	0.47
	300	0.86	0.68	0.68	0.62
39	100	**0.54**	**0.30**	**0.30**	**0.22**
	200	0.75	0.52	0.54	0.43
	300	0.87	0.75	0.71	0.60
DIF (γ = 0.681 or Δ_{MH} = $-$1.6)					
4	100	**0.70**	**0.41**	**0.44**	**0.38**
	200	0.92	0.70	0.72	0.67
	300	0.99	0.89	0.89	0.85
9	100	**0.71**	**0.44**	**0.49**	**0.42**
	200	0.94	0.75	0.80	0.73
	300	0.99	0.93	0.94	0.87
19	100	**0.74**	**0.50**	**0.53**	**0.41**
	200	0.95	0.83	0.83	0.74
	300	0.98	0.92	0.92	0.90
39	100	**0.78**	**0.48**	**0.54**	**0.37**
	200	0.96	0.88	0.87	0.75
	300	0.99	0.94	0.95	0.86

Note. DIF = differential item functioning; MH = Mantel–Haenszel. Entries in boldface indicate that there is an approximate tendency for increasing test length to entail higher DIF detection rates.

Table 3. *SD*($\hat{\gamma}$) and Mean of *SE*($\hat{\gamma}$)

Test length	Sample size per group	*SD*($\hat{\gamma}$)	Mean of *SE*($\hat{\gamma}$)	Mean of *SE*($\hat{\gamma}$) /*SD*($\hat{\gamma}$)
No DIF (γ = 0 or Δ_{MH} = 0)				
4	100	0.39	0.23	0.61
	200	0.26	0.16	0.63
	300	0.21	0.13	0.64
9	100	0.36	0.23	0.65
	200	0.25	0.16	0.66
	300	0.20	0.13	0.69
19	100	0.32	0.23	0.73
	200	0.24	0.16	0.69
	300	0.19	0.13	0.71
39	100	0.35	0.23	0.67
	200	0.23	0.16	0.70
	300	0.19	0.13	0.71
DIF (γ = 0.468 or Δ_{MH} = -1.1)				
4	100	0.37	0.23	0.64
	200	0.29	0.16	0.57
	300	0.23	0.13	0.58
9	100	0.35	0.23	0.67
	200	0.25	0.16	0.66
	300	0.20	0.13	0.68
19	100	0.32	0.23	0.73
	200	0.24	0.16	0.68
	300	0.19	0.13	0.69
39	100	0.33	0.23	0.70
	200	0.24	0.16	0.69
	300	0.18	0.13	0.74
DIF (γ= 0.681 or Δ_{MH} = -1.6)				
4	100	0.40	0.23	0.58
	200	0.28	0.16	0.58
	300	0.22	0.13	0.61
9	100	0.38	0.23	0.61
	200	0.26	0.16	0.64
	300	0.20	0.13	0.67
19	100	0.36	0.23	0.66
	200	0.24	0.16	0.69
	300	0.19	0.13	0.70
39	100	0.36	0.23	0.66
	200	0.23	0.16	0.73
	300	0.19	0.13	0.69

Note. DIF = differential item functioning; MH = Mantel–Haenszel.

Summary and Discussion

This article has made a comprehensive presentation of how the Rasch DIF model can be formulated in the MML estimation context and has investigated the performance of the Rasch DIF model with regard to the DIF parameter recovery and statistical testing

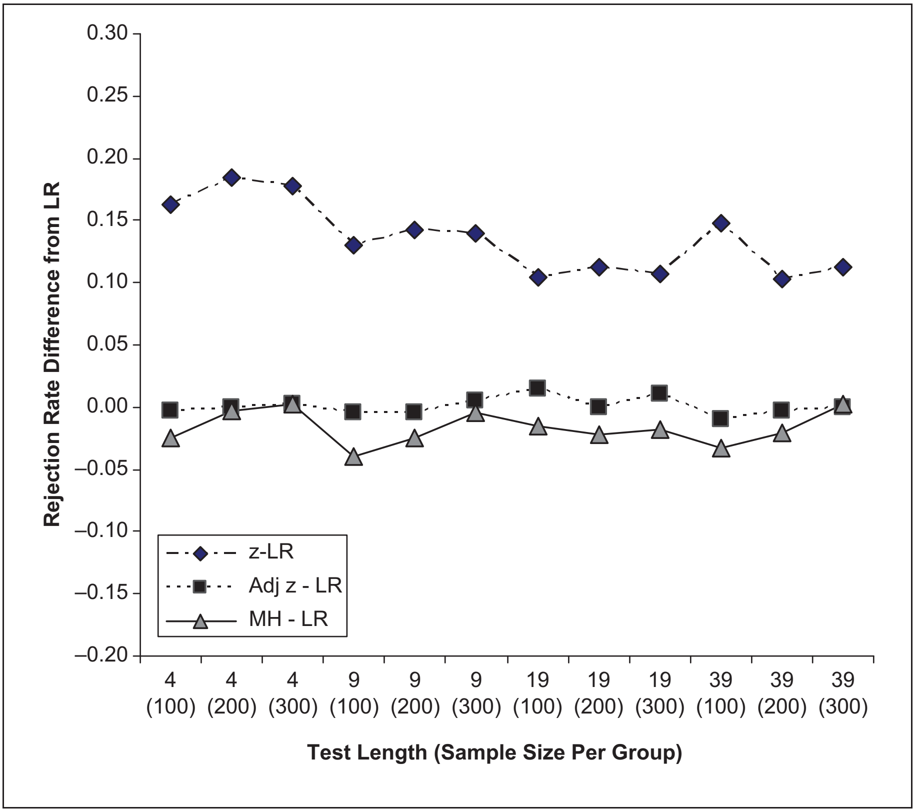

Figure 1. Rejection rate difference from likelihood ratio test result for γ = 0
Note. LR = likelihood ratio; MH = Mantel–Haenszel.

of DIF, comparing it with the MH procedure in the small sample and short test length conditions. For the statistical testing of DIF, the *z* test and the LR test were employed. Overall, the LR test and the adjusted *z* test (i.e., when its denominator is properly estimated) showed good performance with regard to false positive rate and statistical power. Their successful DIF detection rates, when there is DIF, were higher than for the MH procedure for the sample size of 100 to 300 per group with test length of 4 to 39. Note that the small sample and short test length conditions made differences mostly in terms of statistical testing, not in terms of the theoretical relationship between the Rasch DIF model and the MH DIF estimator, $\alpha_{MH} = e^{\gamma}$, which held well. Zwick (1990) mentioned that it is unwise to use the MH approach to short tests because the low reliability of a short test would lead to larger discrepancy of the relationship between the IRF in the IRT model and the DIF measured by the MH procedure. This may explain partly the observed lower rejection rates by the MH procedure than the LR (or adjusted *z* test) in the short test length conditions. Regarding sample

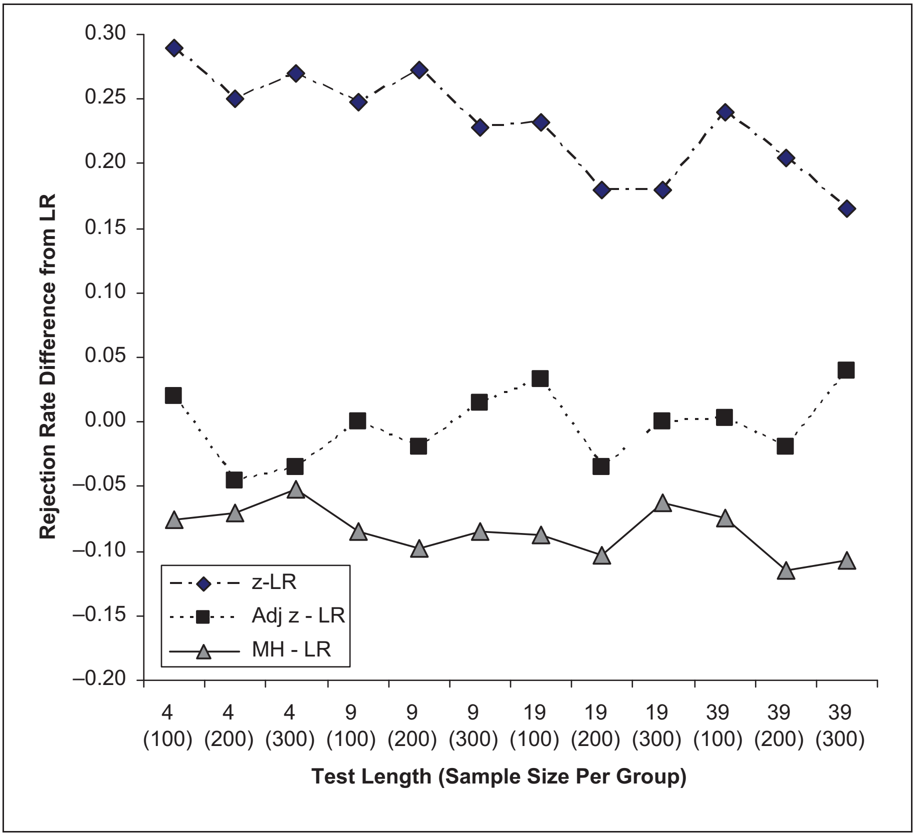

Figure 2. Rejection rate difference from likelihood ratio test result for γ = 0.468
Note. LR = likelihood ratio; MH = Mantel–Haenszel.

size, the results from this study suggest that for small sample sizes such as 100 to 300 per group, the use of the Rasch DIF model could be more powerful than the MH procedure when the Rasch model fits the data.

With a sample size of 200 or higher per group and test length of 9 or higher and when there is medium size DIF (γ = 0.468 or $\Delta_{MH} = -1.1$), both the (adjusted) z test and the LR test detected DIF more than about 50% of the time. As far as large DIF size (γ = 0.681 or Δ_{MH} = 1.6) detection is concerned, a sample size of 200 or higher per group made successful DIF detection at least about 70% across different test lengths. When 300 per group is used, DIF detection of such large DIF by both tests was more than about 90% of the time regardless of small or long test length.

In IRT model estimation, estimating the variance–covariance matrix for the model parameters is more challenging than estimating the parameters themselves. In DIF contexts (e.g., when using Lord's chi-square statistic approach or the Rasch DIF

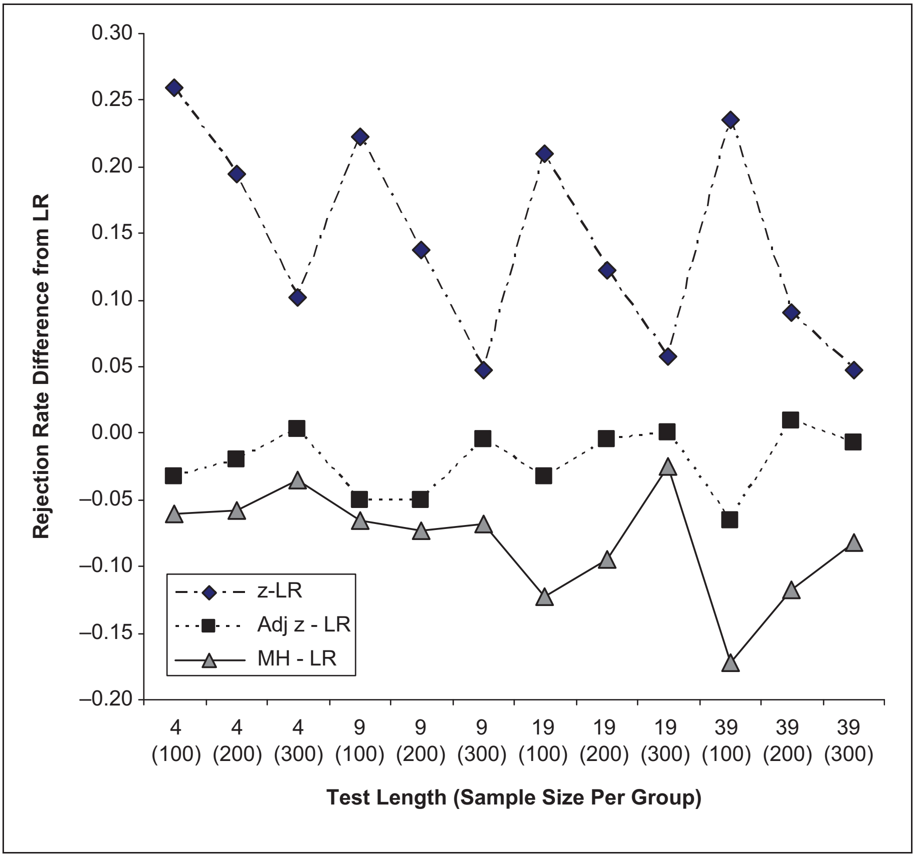

Figure 3. Rejection rate difference from likelihood ratio test for γ = 0.681
Note. LR = likelihood ratio; MH = Mantel–Haenszel.

model as in this article), it seems important for the IRT model users to pay attention to how standard errors of the model parameters are estimated. If they are estimated by an approximate method that, for instance, uses diagonal elements only in the observed information matrix (or any other approximate methods) rather than the full observed information matrix, one needs to be cautious of the extent of the potential systematic bias because of the approximation whenever the estimated standard errors are used for DIF testing (e.g., the *z* test in this study or Lord's chi-square statistic or Wald tests in general). When the Rasch DIF model is estimated, unless $SE(\hat{\gamma})$ is estimated properly, we recommend the use of the LR test.

In practice, when using the Rasch DIF model, there are at least two concerns. One is the model–data fit issue, and the other is how to find a matching item set that is DIF-free. For the first issue, when the Rasch model is adopted for scaling and equating, the model–data fit issue should be considered from the beginning of the test development process and field test analysis to maximize the model–data fit. Benefits gleaned from the use of the Rasch model once the test items are developed to fit the Rasch model are at least twofold with respect to DIF investigation. The first benefit is that testing Equation (1)—No DIF hypothesis of the same IRFs between two groups—is equivalent to the No DIF hypothesis (H_0: $\alpha_{MH} = 1$ or $\Delta_{MH} = 0$) tested by the MH test (Holland & Thayer, 1988), but when the 2PL or 3PL model is used for DIF, the same IRF hypothesis does not necessarily lead to the null hypothesis definition of No DIF in the MH procedure. Hanson (1998) showed that parallel DIF for the 3PL model violates the constant odds ratio assumption tested by the MH DIF unless the guessing parameter is zero. Zwick (1990) showed that the same IRF hypothesis of Equation (1) in the 2PL or 3PL model indicates H_0: $\alpha_{MH} = 1$ or $\Delta_{MH} = 0$ in the MH procedure only if the slope parameters are the same across all items or there are no group ability differences. Thus, when test items are developed using the 2PL or 3PL model but the MH DIF procedure is used for DIF investigation, direct inference of the same IRFs (No DIF) or parallel DIF in the IRT model cannot be guaranteed from the MH procedure results. Another benefit is based on the study by Roussos et al. (1999). They showed that α_{MH} shrinks toward 1 as the studied item difficulty becomes higher. This property of the α_{MH} would result in loss of power for those high difficulty items, a result that is also supported by the Mazor et al. (1992) study. Roussos et al. (1999) also showed that when the Rasch model fits the data, α_{MH} becomes impervious to the negative effect of high item difficulty, which in turn prevents the MH procedure from losing power for high difficulty item DIF detection. For the second issue of finding a matching item set internally when there is no available external matching item set, an iterative procedure or a two-stage matching item set refinement approach can be used (e.g., Holland & Thayer, 1988; Wang, 2008; Wang & Su, 2004). One way to implement the two-stage procedure is to first examine DIF for each item one by one to find No DIF items and to test those previously identified DIF items individually again with the No DIF items as the matching item set.

Last, we would like to discuss the use of directional two-sided test to refine statistical testing inference for the aforementioned alternative Rasch DIF classification rules. Kaiser (1960) proposed the directional two-sided test that allows directional inference. The traditional nondirectional two-sided test has two hypotheses: H_0: $\gamma = 0$ and H_A: $\gamma \neq 0$ (in our Rasch DIF model context). Rejection of H_0: $\gamma = 0$ (No DIF) only tells us that γ is not equal to zero, that is, there is DIF, but it does not provide any directional inference whether $\gamma > 0$ (DIF is against the focal group) or $\gamma < 0$ (DIF is against the reference group) within the statistical testing logic. Note that if a practitioner declares DIF against the focal group by simply observing the DIF estimates (e.g., $\hat{\gamma} = 0.47$ or $\hat{\Delta}_{MH} = -1.1$), his or her declaration of DIF against the focal group is his or her subjective conjecture, not the statistical testing result.

The directional two-sided test starts with three hypotheses[7]: H_0: $\gamma = 0$, H_1: $\gamma < 0$, and H_2: $\gamma > 0$. The construction of statistical testing with these three hypotheses showed that differences between the directional two-sided test and the traditional two-sided test lie in the hypothesis formulation and the decision rules (see Bohrer, 1974; Harris, 1997; and Kaiser, 1960, for derivations and other details of the directional two-sided test. See also Shaffer, 1972, 1974, for reformulating the two-sided directional tests as a pair of simultaneous one-sided tests). Under the same nominal significance level, the critical values for statistical significance for the two-sided directional tests are the same as the traditional two-sided test. At .05 level, they are -1.96 and 1.96 for $\delta_F - \delta_R = \gamma$ testing in the Rasch DIF model. The decision rules are: decide on H_1: $\gamma < 0$ when $z = \hat{\gamma}/SE(\hat{\gamma}) < -1.96$ (i.e., statistically significant DIF against the reference group), decide on H_2: $\gamma > 0$ when $z = \hat{\gamma}/SE(\hat{\gamma}) >$ +1.96 (i.e., statistically significant DIF against the focal group); and decide on H_0: $\gamma = 0$ when $-1.96 \leq z \leq +1.96$ (i.e., uncertain about the direction of DIF or not enough evidence to conclude in either direction). Thus, the alternative Rasch DIF classification rules are

- A if H_0: $\gamma = 0$ is adopted or if $|\gamma| \leq 0.426$;
- B+ if H_1: $\gamma < 0$ is adopted and if $-0.638 < \gamma \leq -0.426$;
- B− if H_2: $\gamma > 0$ is adopted and if $0.426 \leq \gamma < 0.638$;
- C+ if H_1: $\gamma < 0$ is adopted and if $\gamma \leq -0.638$; and
- C− if H$_2$: $\gamma > 0$ is adopted and if $\gamma \geq 0.638$.

By employing the framework of the directional two-sided test, formal statistical inference on the DIF direction is embedded inside hypothesis testing, and the DIF classifications become explicit about the direction of DIF as well.

Appendix A

Formulation of the Within-Logit-Mean DIF Model

Under the same assumptions specified in Equation (1), let $\mathbf{X} = (x_1, x_2, \cdots, x_I)$ indicate a person's item response vector. Recall $\theta|g \overset{iid}{\sim} N(\mu_g, \sigma)$ and $\text{logit}\{P(x_i = 1|\theta, g)\} = \theta - \delta_i + \gamma_i G$. Then, marginal probability of $\mathbf{X}|g$ is

$$P(\mathbf{X}|g) = \int \prod_i P_{ig}(\theta)^{x_i} Q_{ig}(\theta)^{1-x_i} \frac{1}{\sigma} \phi\left(\frac{\theta - \mu_g}{\sigma}\right) d\theta, \tag{A.1}$$

where $P_{ig}(\theta) = P(x_i = 1|\theta, g) = \exp(\eta)/\{1 + \exp(\eta)\}$, where $\eta = \theta - \delta_i + \gamma_i G$; $Q_{ig}(\theta) = 1 - P_{ig}(\theta)$; and $\phi(\,)$ is a standard normal density. Let $\mu_g = \mu + \Delta G'$, where $G' = -1$ if g = focal group and $G' = 1$ if g = reference group. Also, let $\psi = (\theta - \mu_g)/\sigma$. Then, Equation (A1) becomes

$$P(\mathbf{X}|g) = \int \prod_i P_{ig}(\sigma\psi + \mu_g)^{x_i} Q_{ig}(\sigma\psi + \mu_g)^{1-x_i} \phi(\psi) d\psi$$
$$= \int \prod_i P_{ig}(\sigma\psi + \mu + \Delta G')^{x_i} Q_{ig}(\sigma\psi + \mu + \Delta G')^{1-x_i} \phi(\psi) d\psi.$$

Let $\sigma\psi + \mu = \theta^*$. Then,

$$P(\mathbf{X}|g)=\int\prod_{i}P_{ig}(\theta^{*}+\Delta G')^{x_i}Q(\theta^{*}+\Delta G')^{1-x_i}\frac{1}{\sigma}\phi\left(\frac{\theta^{*}-\mu}{\sigma}\right)d\theta^{*} \qquad \text{(A.2)}$$

Note that compared with Equation (2), $\theta|g \overset{iid}{\sim} N(\mu_g, \sigma)$ is replaced by $\theta^{*} \overset{iid}{\sim} N(\mu, \sigma)$ in Equation (A.2).

Defining,

$$P_{ig}(\theta^{*})\equiv P_{ig}(\theta^{*}+\Delta G')=\frac{\exp(\theta^{*}+\Delta G'-\delta_i+\gamma_i G)}{1+\exp(\theta^{*}+\Delta G'-\delta_i+\gamma_i G)},$$

we can write,

$$\begin{aligned}\text{logit}\{P_{ig}(\theta^{*})\}&=\text{logit}\{P(x_i=1|\theta^{*},g)\}\\&=\theta^{*}+\Delta G'-\delta_i+\gamma_i G\\&=\theta^{*}-\delta_i+\Delta G'+\gamma_i G\end{aligned}$$

with $\theta^{*} \overset{iid}{\sim} N(\mu, \sigma)$.

Appendix B

Rasch DIF Model Specification Using the Program ConQuest

To run the Rasch DIF model, a data file, a design matrix file, a population anchor file, and a command file should be prepared in advance. ConQuest used in this study was the one built on November 21, 1997. The data structure shown below allows easy specifications of multiple γ parameters in the design matrix. The following data matrix, design matrix, and the population anchor file ensure the implementation of constraints used for the Rasch DIF model ($\beta_0 = 0$ and the number of elements in $[\gamma_i] < I$), data linkage for the matching items, and establishment of a common scale. The example here is given for a four-item test with the second and the fourth items as studied items.

Data matrix.

```
0011          1
1100          1
⋮⋮⋮⋮⋮          ⋮
0110          1
1000          1
      110100 0
      010100 0
      ⋮⋮⋮⋮⋮⋮ ⋮
      101100 0
      110000 0
```

The first block (up to the fourth column) is the data set for the focal group where columns are for items and rows are for persons. The second block (from the fifth to the eighth column) is the data set for the reference group. The last column (the ninth column) is a group indicator variable (1 for *focal group*; 0 for *reference group*).

Design matrix. The design matrix is the one in the rectangular box below. The number 6 in the first row represents the total number of parameters in the design matrix. The fifth and the sixth columns are for the DIF parameters for the second and the fourth items. If the second item was the only studied item, the design matrix should not have the last column with the total number of parameters being 5 instead of 6. When the number of elements in $[\gamma_i] > 1$ as in this example, the LR test null hypothesis is $H_0 : (\gamma_1, \gamma_2, \ldots, \gamma_i)' = (0, 0, \ldots, 0)'$.

Group	Item Number	Item Response	δ_1	δ_2	δ_3	δ_4	γ_2	γ_4
			6					
Focal	1	0	0	0	0	0	0	0
		1	-1	0	0	0	0	0
	2	0	0	0	0	0	0	0
		1	0	-1	0	0	0	0
	3	0	0	0	0	0	0	0
		1	0	0	-1	0	0	0
	4	0	0	0	0	0	0	0
		1	0	0	0	-1	0	0
Reference	1	0	0	0	0	0	0	0
		1	-1	0	0	0	0	0
	2	0	0	0	0	0	0	0
		1	0	-1	0	0	1	0
	3	0	0	0	0	0	0	0
		1	0	0	-1	0	0	0
	4	0	0	0	0	0	0	0
		1	0	0	0	-1	0	1

Population Anchor file. Inside the population anchor file the following should be the contents.

```
        1        0      0.00000
```

Command file. For the Rasch DIF model,

```
reset;
set update = yes;
set warnings = no;
datafile example.dat;
format response 1-8 grp 9;
code 0 1;
score (0 1) (0 1) ! items (1-8);
regression grp;
model item;
import designmatrix << example.dgn;
import anchor_reg_coefficients << example.rg;
```

```
estimate;
show >> example.out;
exit;
```

The import files are ‘‘example.dat,’’ ‘‘example.dgn,’’ and ‘‘example.rg’’ for the data, the design matrix, and the population anchor files, respectively. The output file is ‘‘example.out.’’ For the LR test, the No DIF model should also be estimated. The design matrix for the No DIF model does not contain the last two columns with the change in the total number of parameters (4 for the No DIF model). The modified design matrix with a proper file name should be imported. All the other aspects of running the program remain the same as the Rasch DIF model.

Output file. The item and DIF parameter estimates are shown following the order specified in the design matrix. For our example, the last two row values in the second column under the heading of ‘‘Parameter estimates’’ in the output file are the DIF estimates, and the last two row values in the third column are estimated standard errors. For the LR testing, use the deviance statistics in the outputs from the Rasch DIF model and No DIF model. The difference of the deviances (deviance of the No DIF model − deviance of the Rasch DIF model) follows approximately a chi-square distribution with degrees of freedom equal to the number of studied items.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research and/or authorship of this article.

Notes

1. In the IRT modeling context, the MML estimation is taken primarily for more effective parameter estimation. In this MML estimation, the person parameters are so-called nuisance parameters, treating them as random parameters to be marginalized in the likelihood, but in a repeated design frame or multilevel modeling perspective (Davidian & Giltinan, 1998; Rijimen et al., 2003), this is a modeling of the dependency of data such as serial correlation within person induced by the nested structure of repeated measurements for a given person. This gives another justification for the MML estimation in the IRT modeling context.
2. Equation (5) can be called the ‘‘within-logit-mean DIF model’’ in the sense that the impact factor is modeled within the logit form in contrast with the Rasch DIF model–Equations (2) and (3). Equation (5) can be also specified by Equation (4) if we let $\mathbf{a}'_{ik=1} = \mathbf{0}_{1\times(I+2)}$, $\mathbf{a}'_{ik=2} = [\lambda_{1m}]_{1\times(I+2)}$, $\boldsymbol{\xi}' = [\delta_1, \delta_2, \ldots, \delta_i, \ldots, \delta_I, \Delta, \gamma_i]_{(I+2)\times 1}$, and $g(\theta; \boldsymbol{\alpha}) = N(\mu, \sigma^2)$, where $\lambda_{1m} = -1$ for $m = i$, $\lambda_{1m} = G'$ for $m = I + 1$, $\lambda_{1m} = G$ for $m = I + 2$, and $\lambda_{1m} = 0$ otherwise. Under these specifications, Equation (4) with $b_{ik=2} = 1$ and $b_{ik=1} = 0$, becomes the within-logit-mean DIF model (Equation 5). The ConQuest program manual (Wu et al., 1998, pp. 75-84)

discusses estimating the same IRF form as Equation (5) for DIF, but with different model constraints.

3. Thissen, Steinberg, and Wainer (1993) called the LR test the general IRT-LR approach, and the squared form of the z test the IRT-D^2 approach, both of which are based on the MML estimation.
4. Theoretically, it has been argued that the latent score matching is preferred to the observed score matching. Meredith and Millsap (1992) showed that the observed score matching approach is not necessarily equivalent to the unobserved score latent score matching approach.
5. $\delta_R = -.85$ for the studied item difficulty was also from the Donoghue et al. (1993) study. In brief, the true difficulties ranged from -0.85 to 0.85 (mean = 0 and SD = 0.72) for the 4-item test; from -1.45 to 1.45 (mean = 0 and SD = 1) for the 9-item test; from -1.85 to 1.85 (mean = 0 and SD = 1.17) for the 19-item test; and from -1.85 to 1.85 (mean = 0 and SD = 1.1) for the 39-item test. See Donoghue et al. (1993, pp. 143-146) for the details of the true item difficulties used in this study.
6. ConQuest estimates parameter standard errors using the observed information matrix (Adams, Wilson, et al., 1997; Wu et al., 1998). The $SE(\hat{\gamma})$ in the Rasch DIF model was estimated from a diagonal matrix approximation to the full observed information matrix. This is a trick sometimes used in IRT variance–covariance matrix estimation to lessen computational burden in the estimation process. The full observed information matrix estimation for the Rasch DIF model was not allowed when using ConQuest. Wu et al. (1998, p. 137) mention that the impact of using the diagonal matrix approximation will be the underestimation of the sampling error in general. Our simulation results using ConQuest confirm the comment by Wu et al.
7. A similar supporting line of thought for the two-sided directional testing logic or three-valued logic was expressed by Tukey (1991). He maintained that the first thing that should be answered is not for the question ''Are the effects of A and B different?'' but for the question ''Can we tell the direction in which the effects of A differ from the effects of B?'' because the reality is that any effects under investigations are different–''in some decimal place.''

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

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. 397-479). Reading, MA: Addison-Wesley.

Bohrer, R. (1974). Multiple three-decision rules for parametric signs. *Journal of the American Statistical Association, 74*, 432-437.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago, IL: Scientific Software.

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

Davidian, M., & Giltinan, D. M. (1998). *Nonlinear models for repeated measurement data*. New York, NY: Chapman & Hall/CRC.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel–Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

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.

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

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 357-374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

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

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 3*, 244-253.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Hillsdale, NJ: Erlbaum.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67*, 160-167.

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

Linacre, J. M., & Wright, B. D. (1987). *Item bias: Mantel–Haenszel and the Rasch model* (MESA Psychometric Laboratory, Memorandum No. 39). Chicago, IL: University of Chicago Press.

Linacre, J. M., & Wright, B. D. (1989). Mantel–Haenszel DIF and PROX are equivalent! *Rasch Measurement Transactions, 3*(2), 52-53.

Longford, N. T., & Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillsdale, NJ: Lawrence Erlbaum.

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). Amsterdam, Netherlands: Swets & Zitlinger.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institution, 22*, 719-768.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (ETS Research Rep. No. 08-43). Princeton, NJ: Educational Testing Service.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel–Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-452.

Mellenbergh, G. J. (1982). Contingency table methods for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

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

Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

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

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

Paek, I. (2002). *Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context* (Unpublished doctoral dissertation). University of California, Berkeley.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press. (Original work published 1960)

Rijimen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185-205.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel–Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*, 293-322.

SAS Institute. (1999). *SAS online doc* (Version 8). Cary, NC: Author.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel–Haenszel procedures for assessing differential item functioning. In G. Engelhard, Jr., & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 65-82). Norwood, NJ: Ablex.

Shaffer, J. P. (1972). Directional statistical hypotheses and comparisons among means. *Psychological Bulletin, 77*, 195-197.

Shaffer, J. P. (1974). Bidirectional unbiased procedures. *Journal of the American Statistical Association, 69*, 437-439.

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.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. New York, NY: Chapman & Hall/CRC.

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. (1988). Use of item response theory in the study of group differences in trace lines. In H. Winter & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369-386.

Van den Noortgate, W., & Paek, I. (2004). Person regression models. In De Boeck, P., & Wilson, M. (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 167-187). New York, NY: Springer.

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

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9*, 387-408.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *Conquest: Generalized item response modeling software*. Camberwell, Victoria, Australia: ACER.

Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.