Real and Artificial Differential Item Functioning in Polytomous Items

Educational and Psychological
Measurement

2015, Vol. 75(2) 185–207

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414534258

epm.sagepub.com

David Andrich¹ and Curt Hagquist²

Abstract

Differential item functioning (DIF) for an item between two groups is present if, for the same person location on a variable, persons from different groups have different expected values for their responses. Applying only to dichotomously scored items in the popular Mantel–Haenszel (MH) method for detecting DIF in which persons are classified by their total scores on an instrument, Andrich and Hagquist articulated the concept of artificial DIF and showed that as an artifact of the MH method, real DIF in one item favoring one group inevitably induces artificial DIF favoring the other group in all other items. Using the dichotomous Rasch model in which the total score for a person is a sufficient statistic, and therefore justifies classifying persons by their total scores, Andrich and Hagquist showed that to distinguish between real and artificial DIF in an item identified by the MH method, a sequential procedure for resolving items is implied. Using the polytomous Rasch model, this article generalizes the concept of artificial DIF to polytomous items, in which multiple item parameters play a role. The article shows that the same principle of resolving items sequentially as with dichotomous items applies also to distinguishing between real and artificial DIF with polytomous items. A real example and a small simulated example that parallels the real example are used illustratively.

Keywords

differential item functioning, polytomous Rasch model, rating scales, partial credit, artificial DIF

¹The University of Western Australia, Crawley, Western Australia, Australia

²Karlstad University, Karlstad, Sweden

Corresponding Author:

David Andrich, Graduate School of Education, The University of Western Australia, M428, 35 Stirling Highway, Crawley, Western Australia 6009, Australia.

Email: david.andrich@uwa.edu.au

Introduction

There are two main reasons why many instruments of assessment in the educational, psychological, health, and social sciences, in general, have more than one item. Specifically, the greater the number of items, the greater the potential validity and the greater the potential precision of the assessment. These reasons for having multiple items are justified psychometrically if the items have invariant scale values among individuals and groups to be compared. We return to these reasons later in the article.

The requirement of the invariant scale values of items was articulated by Thurstone (1928), and the requirement of a response structure that provided invariant comparisons was articulated by Guttman (1950). Working independently, Rasch (1961) incorporated both ideas and went further than either Thurstone or Guttman by rendering the requirement of invariance in the form of a probabilistic response model. Rasch's invariance requirements were the following:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; [and] Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison. (Rasch, 1961, p. 322)

The Rasch model is expressed in terms of a person parameter and relevant item parameters, and because of its properties of invariance, to the degree to which responses fit the model, to that degree the comparisons of the item parameters are invariant with respect to different values of the person parameters, and vice versa. Item-model fit is relative and no item can fit a model on its own; instead, fit of an item implies that the item is operating consistently with the other items of the instrument as summarized by the model.

General evidence that an item's responses *might not* fit the model, irrespective of group membership, concerns the operation of an item across the different values of the assessed variable. Specific evidence that an item's responses *might not* fit the model by operating differently across different identifiable groups is referred to as differential item functioning (DIF). An item is defined to have *no* DIF between groups if, for the *same value* on the variable defined by the instrument, persons from the different groups have the same expected value for their responses to the item.

The importance of detecting and dealing with DIF is testified by the vast literature in the field for both dichotomously and polytomously scored items (e.g., Brodersen et al., 2007; Budgel, Raju, & Quartetti, 1995; Holland & Wainer, 1993; Penfield & Lam, 2000; Potenza & Dorens, 1995; Roussos, Schnipke, & Pashley, 1999; Tennant & Pallant, 2007; Wang, 2000; Zwick, Donoghue, & Grima, 1993). This literature includes observations that "sometimes, for reasons unknown, calculations of a DIF detection strategy may suggest DIF, where none truly exists" (Osterlind & Everson, 2009, p. 21) and spurious DIF (Kreiner & Christensen, 2011). Andrich and Hagquist (AH; 2012) formalized this kind of DIF in the well-known method of Mantel and

Haenszel (1959) for detecting DIF in dichotomous items, a method popularized by Holland and Thayer (1988). AH showed that as an artifact of the MH procedure, the presence of DIF in one item that favors one group inevitably induces the appearance of DIF in the other items that favor the other group. To emphasize the difference between DIF that is present and that which only *appears* to be present, AH referred to the former as *real* and the latter as *artificial*—artificial because it is an artifact of the method for detecting DIF.

In detecting DIF in usual data sets, parameter values of persons are unknown. In the MH method, the unknown values are substituted by their total scores on the items and persons are classified by their total scores. AH formalized the artificial DIF in the MH procedure by relating it to the dichotomous Rasch model, a relationship also explicated by Holland and Thayer (1988). This formalization is facilitated by the Rasch model because the total score of a person on the items of an instrument is the sufficient statistic for the person parameter (Andersen, 1977; Rasch, 1961). As a result, classifying persons by total scores is justified by the model and is not merely an ad hoc, convenient method of classification.

AH also showed that the formalization of artificial DIF provided a rationale for differentiating between real and artificial DIF. The rationale implies a sequential procedure of resolving items in which, from a simultaneous assessment of DIF of all items, the item with the greatest magnitude of DIF is resolved to create two distinct items, one for each group. Not only does this resolution of an item remove the artificial DIF it induced in other items, but it quantifies the DIF in the metric of the variable defined by the items that show no DIF. When no further DIF is evident, items having real DIF are identified. This procedure for resolving items has been used in Rasch model analyses (Brodersen et al., 2007) without the procedure being connected to eliminating artificial DIF.

This article generalizes the AH procedure for distinguishing between real and artificial DIF in dichotomous items using the dichotomous Rasch model to polytomous items using the polytomous Rasch model (PRM). Integral to the generalization is the role of resolving items sequentially.

The literature on DIF distinguishes between uniform and nonuniform DIF. The former reflects consistent differences across the continuum of the variable, the latter an interaction implying different magnitudes of DIF across the continuum. Because of the greater number of item parameters in the PRM than in the dichotomous model, there are differences in the assessment of nonuniform DIF between the two models. Although the procedure of resolving items sequentially for distinguishing between real and artificial DIF generalizes to nonuniform DIF, the article is concerned with uniform DIF; nonuniform DIF is broached only where relevant.

AH demonstrated that classifying persons by their total scores, which is tantamount to classifying them by their *estimates* in the dichotomous Rasch model rather than using known person location values, is the source of artificial DIF in the MH method. The same principle applies with the PRM. It is recognized that other models are used in assessing DIF. However, as with the dichotomous model, if persons are

classified by their estimates from the data within which DIF is assessed, then because identifying constraints are imposed on the equations in any estimation in any model, artificial DIF will be induced. We leave the study of this effect with other models for other occasions.

The rest of the article is structured as follows. Using just three items with five categories, the second section shows the mathematical form of artificial DIF in the PRM. To distinguish between real and artificial DIF using the sequential resolution of items, beginning with the item showing the largest DIF, it is necessary to have evidence regarding DIF and its relative magnitude among items. Therefore, although not the main point of the article, the third section summarizes briefly an efficient method for identifying the relative magnitude of any DIF among items, which may include artificial DIF using an analysis of person-item residuals. The fourth section shows an analysis of a real example with eight polytomous items each with five categories assessing mental health in early adolescence. It shows that although initially there are four items that show DIF, only two have real DIF, both having higher scores for girls. To confirm the efficacy of the procedure in detecting and quantifying real DIF, Appendix A provides a simulated example. The fifth section considers the implication for interpreting the means of the groups when items are resolved to remove artificial DIF, and the last section is a summary.

The PRM, Model Fit, and Real and Artificial DIF

There are different expressions for the PRM (Andersen, 1977; Andrich, 1978; Wright & Masters, 1982), which follow from Rasch's (1961) form of the polytomous model. The one convenient for this article is given by

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

where $X_{ni}=x \in \{0, 1, \dots, m_i\}$ are the scores associated with the $m_i + 1$ successive categories of item i , τ_{ki} are m_i thresholds defining the $m_i + 1$ successive categories on the continuum, $\tau_0 \equiv 0$, $\sum_{k=0}^{m_i} \tau_{ki} = 0$, β_n and δ_i are the respective location parameters of person n and item i on the same variable expressed in logits, and γ_{ni} is a normalizing factor. Clearly, with a single location parameter for each person the model is uni-dimensional. For purposes of estimation, only one identifying constraint is required, usually $\sum_{i=1}^I \hat{\delta}_i = 0$.

As in the dichotomous model, the total score $r_n = \sum_{i=1}^I x_{ni}$ across I items is the sufficient statistic for the person parameter β_n (Andersen, 1977; Andrich, 1978), which implies that the person parameters can be eliminated when the item parameters are estimated. The analyses of the examples in this article were carried out with the software RUMM2030 (Andrich, Sheridan, & Luo, 2013), which operationalizes the conditional pairwise algorithm (Andrich & Luo 2003) in which the person parameters

are eliminated while the item parameters are estimated. The person parameters are then estimated taking the item parameters as known.

Because the definition of the DIF involves the expected value of a response to an item for persons with the same location on the variable but different group membership, the expected value curve, referred to as the item characteristic curve (ICC), provides the frame of reference for assessing DIF in this article. The expected value for an item as a function β is given by

$$E[X_i; \beta] = \sum_{x=0}^{m_i} x \Pr\{X_i = x | \beta\}, \quad (2)$$

which specializes simply to $E[X_i; \beta] = \Pr\{X_i = 1 | \beta\}$ for dichotomous items.

The ICC and Model Fit

As indicated above, because of the sufficiency of the total score for the person parameter in the PRM, it is natural in testing fit to classify persons by their total scores, or in the presence of a large number of total scores, classify them into class intervals based on adjacent total scores. A general test of fit, irrespective of group membership, checks the degree to which the observed means of the responses in the class intervals across the variable are close to the ICC. A specific test of DIF checks the degree to which the observed means of the responses in the class intervals across the variable *for each group* are close both to each other and to the ICC. Although the principles of assessing DIF generalize beyond two groups, in this article we are concerned with just two groups. In addition, although one can be seen as a dominant group and a second as a focal group, both groups may have equal a priori status, such as groups defined by gender. This article considers an example of the latter kind.

Artificial DIF

To illustrate the mathematics of artificial DIF, three items with the parameters shown in Table 1 are used. Suppose two groups of persons, designated boys and girls for convenience, respond to these three items each of which has five ordered response categories. Item 1 has a 1 logit DIF favoring girls with a location of -0.5 compared to 0.5 for boys. That is, Item 1 has uniform DIF favoring girls. For example, for $\beta = 0$, $E[X_1 | Girl] = 2.546$ and $E[X_1 | Boy] = 1.454$.

Without loss of generality, the condition that all items have the same number of categories can be relaxed, but for simplicity of illustration, we retain the same number of categories, and the same parameter values for the thresholds for all items. Therefore, in subsequent expressions, the subscript i is dropped from the maximum score m_i . We assume that all persons have responded to the same items, though again, without loss of generality, this condition can be relaxed.

Table 1. Location and threshold values of three items with maximum score 4.

Item	Locations					Thresholds		
	1 Girls -0.5	1 Boys 0.5	2 -1.0	3 1.0	1 -1.5	2 -0.5	3 0.5	4 1.5

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

$$r = \sum_{i=1}^I \sum_{x=0}^m x \Pr\{X_i = x | \hat{\beta}_r\}, \quad (3)$$

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

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

As illustrated above for $\beta = 0$, because the location for Item 1 of Table 1 is greater for boys than for girls, for all β_r ,

$$E[X_{xr1} | Girl] = \sum_{x=0}^m x p_{xr1} | Girl > E[X_{xr1} | Boy] = \sum_{x=0}^m x p_{xr1} | Boy. \quad (4)$$

We focus on Item 1 from the set of three items in Table 1, and write Equation (3) in the form

$$r = \sum_{i=1}^3 \sum_{x=0}^m x p_{xri} = \sum_{x=0}^m x p_{xr1} + \sum_{i=2}^3 \sum_{x=0}^m x p_{xri}. \quad (5)$$

The constraint in Equation (5) and the inequality for Item 1 in Equation (4) imply that the terms that involve Items 2 and 3 on the right-hand side of Equation (5) must satisfy the inequality

$$\sum_{i=2}^3 \sum_{x=0}^m x p_{xri} | Girl < \sum_{i=2}^3 \sum_{x=0}^m x p_{xri} | Boy. \quad (6)$$

Moreover,

$$E[X_{xr1} | Girl] - E[X_{xr1} | Boy] = \sum_{i=2}^3 E[X_{xri} | Boy] - \sum_{i=2}^3 E[X_{xri} | Girl]. \quad (7)$$

Thus, Equation (7) shows that the real DIF in Item 1 with a higher score for girls induces artificial DIF with a higher score for boys in the other items and,

Table 2. $r : E[X_{xri}]$ for Three Values of r in the Three-Item Example.

Item	3 : $E[X_{x3i}]$	6 : $E[X_{x6i}]$	9 : $E[X_{x9i}]$
1 (Boy, Girl)	(0.691, 1.126)	(1.712, 2.287)	(2.874, 3.309)
1 Girl-Boy	0.435	0.575	0.435
2 (Boy, Girl)	(1.855, 1.541)	(3.011, 2.723)	(3.667, 3.546)
2 Girl-Boy	-0.314	-0.288	-0.121
3 (Boy, Girl)	(0.454, 0.333)	(1.277, 0.990)	(2.459, 2.145)
3 Girl-Boy	-0.121	-0.287	-0.314
Sum Boy, Girl = r	3	6	9

furthermore, that this artificial DIF is distributed across all remaining items. To make Equation (6) concrete, Table 2 shows the values of $E[X_{xri}|Girl]$, $E[X_{xri}|Boy]$ for three values of r for all three items in Table 1. Table 2 shows that for Item 1, shown in bold, $E[X_{xri}|Girl] > E[X_{xri}|Boy]$ for all three values of r , and that $E[X_{xri}|Girl] < E[X_{xri}|Boy]$ for the other two items, again for all values of r .

The values in Table 2 are a direct result of calculations based on the item parameters and the ML estimation equation and involve no data. The artificial DIF is an inevitable part of the constraint of the ML estimation given by Equation (5) and the grouping of persons by their total score.

In the analysis of fit in real data, the mean of the observed scores for persons with a score of r is compared to $E[X_{ri}]$. For notational convenience, let P_{xri} be the observed *proportion* of persons with a total score r who respond with score x , $x = 0, 1, 2, \dots, m$. Then the observed mean, $Mean[X_{ri}]$, of the scores of persons with each total score r is given by $Mean[X_{ri}] = \sum_{x=0}^m xP_{xri}$. Because, by definition, these scores across items have to sum to r , that is, $\sum_{i=1}^3 \sum_{x=0}^m xP_{xri} = r$, the constraint of Equations (3) and (5) also holds when the observed proportion P_{xri} replaces the estimated probability p_{xri} of score x .

To illustrate the effect of real DIF in Item 1 and artificial DIF with Items 2 and 3 with observed means in data, 1,000 boys and girls, normally distributed with means 0 and standard deviation 2, were simulated with responses to the items in Table 1. The ICCs from the known item locations and thresholds from Table 1 were formed. For Item 1 and under the hypothesis of no DIF, this ICC has a location of 0.0, which is the average location of Item 1 for the boys and girls. Then based on their total scores, and irrespective of gender, the persons were placed into five class intervals across the variable.

Figure 1 shows the ICCs for Item 1 and observed means, first with boys and girls aggregated (Figure 1A) and then with boys and girls disaggregated (Figure 1B). It is evident that with boys and girls aggregated, the observed means in the class intervals are close to the ICC and that when disaggregated, and reflecting the known real DIF, the observed means are greater for girls than for boys. On the other hand, Figure 2 shows that for both Items 2 (Figure 2A) and 3 (Figure 2B), the observed class

Figure 1. Theoretical ICCs for Item 1 and observed means (●) in five class intervals in the example with three items: (A) Groups aggregated; (B) Groups disaggregated.

interval means are greater for boys than girls. However, because the total artificial DIF in Items 2 and 3 is equal to the magnitude of the real DIF in Item 1, their DIF is smaller than that in Item 1.

Resolution of the Item With Evidence of Greatest DIF

As indicated above, AH showed that a method for distinguishing real from artificial DIF involves first resolving the item with the evidence of greatest DIF, thus creating two items, one of which is responded to only by one of the two groups, named, say Item 1 (Girl) and Item 1 (Boy). Data for Item 1 as such are eliminated. This creates missing data in the response matrix, which is readily handled in estimation using the Rasch model. Because the total score for a Girl then does not include Item 1 (Boy), and vice versa, resolved Items 1 (Girl) and 1 (Boys) no longer induce artificial DIF in the other items. In addition, after resolution Item 1 no longer has the same location estimate for boys and girls. That is, Equation (5) is not the same for boys and girls, resulting in the person estimates $\hat{\beta}_{rg}$, $\hat{\beta}_{rb}$ for the total score r being different for boys and girls. This effect in polytomous items is the same as in the dichotomous items.

Figure 2. Theoretical ICCs and observed means (●) in five class intervals for Items 2 and 3 in the example with three items showing artificial DIF: (A) Item 2: boys and girls aggregated; (B) Item 3: boys and girls disaggregated.

The reason for choosing the item with the greatest DIF to resolve first is that, as shown in Equation (7), because the real DIF of an item is distributed among all items rendering the artificial DIF in each item smaller than the real DIF, the item with the greatest DIF is most likely to have real DIF. In the example with three items, it is evident from Table 2 that Item 1 has DIF of the greatest magnitude.

Figure 3 shows the ICCs for the resolved Item 1 (Boy) and the resolved Item 1 (Girl) in both of which the observed means of the class intervals are close to the ICC. The curves themselves have different locations—they are 1 logit apart. Figure 4 shows the ICCs for Items 2 and 3 in which the observed means are close to the ICC for both boys and girls and, therefore, unlike Figure 2, no longer show artificial DIF.

Estimated Item Parameters and Artificial DIF

The demonstrations of the effect of real DIF on inducing artificial DIF reflected in Figures 1 to 4 involve only known item parameters. This permitted an analysis of the mathematics of artificial DIF free from item parameter estimates. However, in real data item parameters are not generally known, and all parameters are substituted by

Figure 3. Theoretical ICCs for resolved Item 1 and observed means (●) in five class intervals in the example with three items.

Figure 4. Theoretical ICCs and observed means (●) in five class intervals for Items 2 and 3 in the example with three items when Item 1 is resolved: (A) Item 2: boys and girls disaggregated; (B) Item 3: boys and girls disaggregated.

their estimates. The full illustration of the effect when both item and person parameters are estimated is shown in the next section with the real example. A simulated

example, which parallels the real example, and confirms the identification of artificial DIF, is shown in Appendix A.

Identifying DIF Using ANOVA of standardized Residuals From the ICC

In AH where the focus was on dichotomous items, the MH method was used to identify items with potential DIF. There are a number of other techniques in the literature for identifying DIF, not only for dichotomous but also polytomous items. Because the point of the article is to demonstrate the concept of artificial DIF in polytomous items, together with a method for distinguishing it from real DIF, this article does not set out to survey these methods (e.g., Holland & Wainer, 1993; Osterlind & Everson, 2009). Instead, this section summarizes a straightforward method for identifying possible DIF from a single analysis that can be used readily, not only to identify both uniform and nonuniform DIF for both dichotomous and polytomous items, but also to rank items by the relative magnitude of their observed DIF. This method, introduced in Hagquist and Andrich (2004), involves the analysis of variance (ANOVA) of residuals of responses from the estimated ICC.

From an initial analysis of responses of persons to items, a standardized residual, $z_{ni} = (x_{ni} - E[X_{ni}]) / \sqrt{V[X_{ni}]}$, for every person's response to every item is constructed. The persons are placed into class intervals based on their total scores, or in the presence of missing data, based on their estimates. The persons are also identified by group, giving a two-way structure of residuals, that of *class-intervals* by *groups*. If there is no DIF, then the residuals would essentially have no structure, and therefore would show no significant effect based on the grouping. However, as is evident from Figure 1B, if there is DIF, then residuals in one group will tend to be positive (girls in this example), and the residuals in the other group (boys in this example) will tend to be negative.

The standardized residual of each person n to each item i , identified by group g and class interval c , is given by

$$z_{n(cg)i} = (x_{n(cg)i} - E[x_{n(cg)i}]) / \sqrt{V[x_{n(cg)i}]}. \quad (8)$$

These residuals are analyzed according to a standard two-way ANOVA. This analysis provides evidence whether there is a main gender effect, which in the absence of any interaction is an assessment of uniform DIF. In addition, the ANOVA provides evidence of a class interval effect and an interaction between the class interval and the group. The former is a test of fit of responses across the variable, irrespective of groups, and the latter is a test of nonuniform DIF. Thus, as a by-product of assessing a main group effect, the ANOVA provides both a general test of fit for the responses, irrespective of the group classification, and a test of nonuniform DIF.

Because the numbers in each gender-by-class interval cell are unlikely to be equal or in a constant proportion, the component sum of squares in general do not sum to the total sum of squares, that is,

$$SS_T \neq SS_{gender} + SS_{classintervals} + SS_{within} + SS_{interaction}. \quad (9)$$

Therefore, the sums of squares for the interaction is calculated as the difference

$$SS_{interaction} = SS_T - SS_{betweencells} - SS_{withincells}, \quad (10)$$

where *betweencells* and *withincells* refer to groups composed of a group-by-class interval combination (Glass & Stanley, 1970). In the case that this interaction is negative, it is generally very small, and it is assumed it arises from random variation and that it is not significant.

Although the standardized residuals are not fully linear (Karabatsos, 2000) and not strictly normal but “approximate unit normal” (Smith, 1988), the robustness of the *F*-distribution used in ANOVA (Lindquist, 1956) should provide sound evidence of statistical effects that are compatible with the graphical displays. Reporting on a series of studies, Lindquist (1956) notes that “it is evident . . . that the *F*-distribution is amazingly insensitive to the form of the distribution of criterion measures in the parent population, granting that the same form is common to all treatment populations” (p. 81). Thus, the ANOVA of residuals permits a test of fit of an item across the variable and of uniform and nonuniform DIF among a priori specified groups to be produced simultaneously in one analysis of residuals. Importantly, the relative *F* values of the ANOVA produce a rank order of items according to the magnitude of their DIF.

The Example

The example involves survey data from a cross-sectional study collected in 1998 among students in Year 9 in the county of Värmland in Sweden. The study was part of a county study with re-current cross-sectional data collections and was carried out by the Centre for Public Health Research, Karlstad University, Sweden. The data collection involved a questionnaire handed out in the classrooms by school personnel. For the purpose of this article, eight items intended to be a composite measure of psychosomatic health are used (Hagquist, 2008). These are listed in Table 3 together with the percentages of responses in each of the categories for each of the items in the data used. The response categories for all of these items, which are in the form of questions, are “never,” “seldom,” “sometimes,” “often,” and “always.” Clearly, the categories are ordered in terms of an implied frequency, and the greater the frequency, the worse the well-being. The targeted population consisted of all students in Year 9 in all 16 municipalities in the county of Värmland, a total of 3,024 students. The response rate was 90.4%, providing a data set consisting of 2,734 cases. For the illustrative purpose of this article, all respondents from the schools in the city

Table 3. Eight Questions and Response Proportions.

<i>During this school year, have you . . .</i>	Never	Seldom	Sometimes	Often	Always
1. had difficulty concentrating?	0.03	0.25	0.45	0.23	0.04
2. had difficulty sleeping?	0.19	0.35	0.29	0.14	0.03
3. suffered from headaches?	0.18	0.30	0.30	0.18	0.04
4. suffered from stomach aches?	0.31	0.35	0.24	0.10	0.01
5. felt tense?	0.19	0.37	0.30	0.12	0.03
6. had little appetite?	0.36	0.34	0.20	0.07	0.02
7. felt sad?	0.19	0.33	0.30	0.16	0.02
8. felt giddy?	0.37	0.32	0.20	0.11	0.01

of Karlstad, except 12 persons who had missing data, were used. Although it is possible to use incomplete data, with so few missing responses, it was convenient to have complete data. The total number of persons used in the original analysis is 654, with 301 boys and 353 girls.

It was indicated above that the status of the genders was the same in the analysis of DIF. Therefore, within this group of students with complete data, 301 girls, the same number as the number of boys in the sample, were chosen at random for the DIF analysis. This equality of sample sizes ensured that if there was DIF, one group did not dominate in the estimates of parameters.

ANOVA of Standardized Residuals

Table 4 summarizes the ANOVA of residuals and the *F* ratio and its significance for each item for (a) the uniform DIF gender effect, (b) the nonuniform DIF interaction effect, and (c) the class-interval effect (general test of fit) irrespective of gender. For illustrative purposes, the criterion of significant DIF is taken as $p < .01$, this being a relatively conservative value in the case that there are 24 different significance tests carried out in Table 4, a number that increases the probability of finding DIF where there is none. A Bonferroni (Bland & Altman, 1995) correction for this error can be applied, but it was considered unnecessary in this illustrative example. The values of the significant *F* ratios are shown in bold. Whether the DIF favored boys or girls (shown in Table 4) was ascertained from graphs such as those in Figure 2A and B.

Table 4 shows that Item 4 (*Suffered from Stomach aches*) and Item 7 (*Felt Sad*) have substantial DIF with girls having higher scores than boys and that Item 1 (*Difficulty in Concentrating*) and Item 2 (*Difficulty in Sleeping*) have marginal DIF with boys having higher scores than girls. There is no evidence of nonuniform DIF. From Table 4 it was concluded that four items, Items 1, 2, 4, and 7 *potentially* show DIF, which is primarily uniform and with an equal number of items favoring each gender. This conclusion has two implications. First, that the remaining four items do not show DIF. Second, because of the way that real DIF in one item is distributed as artificial DIF among all other items, the relatively larger *F* values for Items 7 and 4,

Table 4. ANOVA of Standardized Residuals Using Five Class Intervals (CIs).

Item	Gender		Gender-by-CI		CI	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
1. Concentrating? (B > G)	8.589	.004	−0.372	1.000	1.536	.190
2. Sleeping? (B > G)	6.760	.010	1.035	.388	0.624	.646
3. Headaches?	0.059	.808	0.407	.804	0.700	.593
4. Stomachaches? (G > B)	18.749	.000	0.740	.565	1.743	.139
5. Tense?	0.061	.805	0.454	.770	2.569	.037
6. Appetite?	0.525	.469	0.412	.800	0.292	.883
7. Sad? (G > B)	43.149	.000	0.638	1.000	1.651	.160
8. Giddy?	5.501	.019	1.635	.164	0.430	.787
	<i>df</i> : 1, 602		<i>df</i> : 4, 602		<i>df</i> : 4, 602	

Note. *p* = .01 taken as significant (shown in bold).

compared to those of Items 1 and 2, suggests that the former pair favoring girls might have real DIF and the latter two favoring boys may have induced artificial DIF.

Sequential Resolution of Items Showing DIF

From the mathematics of artificial DIF, which implies that to identify real DIF the item with the largest magnitude of DIF should be resolved first, Item 7 with the largest *F* value (43.1488, *df*: 1, 602), *Felt Sad*, was resolved. Following this resolution, the only significant ANOVA statistic was the main effect for Item 4 (*Stomach aches*; *F*: 26.6093, *df*: 1, 602, *p* = .0000). That Items 1 and 2 do not show DIF suggests that Item 7 has real DIF and that their DIF was artificial. Item 4 showing DIF, even after Item 7 was resolved, suggests its DIF is also real. To confirm this suggestion, Item 4 was also resolved.

Table 5 shows the results of the ANOVA of standardized residuals with both Items 7 and 4 resolved. It is evident that the two resolved items fit their ICCs and that all other items fit their ICCs and show no gender DIF. In particular, Items 1 and 2, which showed DIF in the first analysis (Table 4), do not show DIF in the final analysis (Table 5). This lack of DIF confirms that the original DIF of these two items was artificial. In both the items showing DIF in Table 5, girls have higher scores for the same person estimates, and together they induced artificial DIF in Items 1 and 2. Table 5 also confirms that there are six items that show no DIF.

Figure 5 shows the ICCs of resolved Item 7, which confirms it has real DIF favoring girls. The resolved ICCs also show a slight difference in slopes, a difference that arises from different threshold estimates. However, because the ANOVA of residuals indicates that this difference is not statistically significant, the main focus is on the location differences between the pairs of resolved items, and therefore on uniform DIF. The study of possible nonuniform DIF and its reflection in different slopes of

Table 5. ANOVA of Standardized Residuals When Items 7 and 4 Are Resolved by Gender Using 5 Class Intervals (CIs).

Item	Gender		Gender-by-CI		CI	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
1. Concentrating?	2.361	.125	0.309	.872	1.431	.222
2. Sleeping?	0.887	.347	0.589	.671	0.533	.712
3. Headache?	3.889	.049	−0.145	1.000	1.204	.308
5. Tense?	3.634	.057	−0.467	1.000	3.377	.010
6. Appetite?	0.630	.428	0.530	.714	0.379	.823
8. Giddy?	0.667	.414	0.838	.501	0.642	.633
4a. Stomachaches? B					1.162	.328
4b. Stomachaches? G					0.688	.601
7a. Sad? B					0.468	.759
7b. Sad? G					2.899	.022
	df: 1, 602		df: 4, 602		df: 4, 602	

Note. *p* = .01 taken as significant.

the ICC in the PRM is beyond the scope of this article. Although not shown, the ICCs of resolved Item 4 are similar, though with closer locations, to those of Item 7.

Quantifying Real DIF

The resolution of the items provides an estimate of the magnitude of real DIF in the metric of the scale values of the items with no DIF. In addition, because standard errors of the estimates are available from the ML estimates, a significance test for this difference can be carried out. Table 6 shows the location parameter estimates for resolved Items 7 and 4 and their differences. The magnitude of DIF in Item 7 is of the order of 0.5 logits, whereas that of Item 4 is of the order of half of that size. The statistical comparisons of these resolved items can be carried out using the *Z*-test: $Z(location) = (\hat{\delta}_b - \hat{\delta}_g) / \sqrt{SE_{\delta b}^2 + SE_{\delta g}^2}$. The values of this test, also shown in Table 6, indicate that the DIF with girls having higher scores than boys are both statistically significant.

A Simulated Example Based on the Parameters of the Real Example

A simulated example with the parameter estimates of the real example is used to consolidate the demonstration of artificial DIF using the PRM with polytomous items. The simulation is intended to demonstrate the general validity of the using the ANOVA of residuals and resolving items and is not intended to be a demonstration of general factors affecting artificial DIF or of the distributional properties of any of

Figure 5. ICCs for resolved Item 7 and observed means (●) in five class intervals.

Table 6. Parameter Estimates (Standard Errors) for Resolved Items 7 and 4 and Significance Test for Location Difference.

	Location (SE)		Difference	Z	p
	Boys	Girls			
Item 7 Sad	0.108 (0.076)	−0.444 (0.078)	0.552	5.068	.000
Item 4 Stomach	0.552 (0.080)	0.316 (0.073)	0.236	2.179	.024

the indices. The results of this simulation, which do follow the results of the example with real data, are shown in Appendix A.

Implications of Resolving for DIF on Person Measurement

Before proceeding to consider the interpretation of the effect of the resolution of items that confirms and takes account of DIF, it is stressed that in designs in which some items form a link between items that are administered to different groups, it is critical that the link items show no DIF (Andrich & Hagquist, 2012; Looweer & Mulligan, 2009). In this article, we are concerned only with the case in which all persons respond to all items and in which two items have been identified to have real DIF, and both items showing a relatively higher score for one of the two groups.

For completeness, Figure 6 shows the distribution of the locations estimates of the boys and girls following the resolution of Items 7 and 4. It shows that person and threshold distributions are relatively well aligned and that the group means indicate low scores relative to the defined threshold mean of 0.0. The inference is that the students have relatively low frequencies of the maladies represented by the items, which is consistent with the response distributions for the items shown in Table 3.

Figure 6. Person and threshold distributions with Items 7 and 4 resolved.

Table 7. Mean, Standard Deviation, *F* Ratio for an ANOVA Between Means for the Original Items and for Items Showing DIF Resolved.

	Boys, mean (SD)	Girls, mean (SD)	Boys-Girls, means	ANOVA, <i>F</i> (<i>p</i>)
All items	−1.336 (1.23)	−0.512 (0.93)	0.824	85.832 (.000)
2 DIF resolved	−1.267 (1.22)	−0.608 (0.97)	0.559	53.642 (.000)

As shown with the real example, although resolving items with real DIF can overcome the problem of misfit, the relative location estimates of the DIF items that are resolved are no longer invariant across the groups. This section considers briefly some implications of this trade-off between model fit and item invariance on the comparison of means between the two groups. For this purpose, Table 7 shows the mean, standard deviation, and the *F* test for checking the difference between the means of the boys and girls before and after the two items with real DIF were resolved.

For a resolved item with real DIF, the group with the higher scores on the item has a lower location estimate for that group. This is evident in Table 6 in which resolved Items 7 and 4 have lower estimates for girls than boys. Therefore, for the same total score the location estimates for girls will be less than those for boys, and the difference between their means is reduced. Table 7, which shows the means from the original analysis and that when Items 7 and 4 are resolved, confirms this effect. Because the identifying constraint that ensures that the sum of the item parameter estimates sums to zero in every analysis, the origin was adjusted to be the same in both analyses. This was achieved by anchoring the parameter estimates of the six

invariant items to their values obtained when Items 7 and 4 and the data reanalyzed. The results of the fit analysis were essentially the same as in the original analysis of eight items.

The reduction in the difference between the means of girls and boys when resolved items no longer have invariant estimates for the groups raises the question of the validity of the comparison between their means. If the source of DIF can be understood as a result of an aspect *irrelevant* to the content of the variable and therefore deemed dispensable, then resolving the item and accounting for the DIF seem legitimate. However, if the source of DIF involves an *aspect* of the item *relevant* to the content of the variable and thus deemed indispensable, then resolving the item in a way that reduces the difference between the group means may seem dubious. Specifically, in the real example, if the DIF in Items 7 and 4 can be shown from independent research that girls simply have a stronger response set to report the maladies than boys then the source of DIF appears to be an aspect *irrelevant* to the variable and therefore that the original difference is exaggerated. However, if again from independent research it is concluded that differences in frequency of these maladies is greater for girls than boys given the same values on the variable, then resolving the items and thus reducing the difference between boys and girls might be misleading. This concern with the validity of interpretation in the presence of DIF, though not in the context of resolving items, is similar in principle to that expressed by Camilli (1993). He notes that the earlier literature referred to concepts associated with DIF as *bias* (e.g., Ironson, 1983) and that care is required not to ignore the issue of bias when using the more neutral term of DIF.

In addition to resolving items with real DIF, an option is simply to remove them. The direction of the effect in the comparisons between means is identical (though the values will be somewhat different) to that from resolving items. The advantage in resolving items over eliminating them is that one of the reasons for having multiple items in an instrument, that of increasing precision, is retained. Resolving rather than removing items means that each person's estimate is based on the same number of items as originally designed for the instrument, eight in the above example. However, as seen above, retaining an item but resolving it does not necessarily overcome the other reason for having multiple items, that of assessing multiple, relevant, aspects of the variable. On the other hand, eliminating items defeats both purposes. First, having less items reduces the precision of the estimates. Second, eliminating items with DIF removes totally the aspect of the variable that is present in the item that shows DIF and can have the same distorting effect on the variable that resolving the item can have.

Although not dwelt on in this article, resolving an item increases the number of parameters in the estimation by a factor of the number of groups for which the item has been resolved, and with an increase in the number of parameters estimated, fit is improved. In principle, achieving model fit by increasing the number of parameters in any other way and in any model will also improve fit, but it will also have the same substantive implications as those exposed above. The interpretations of differences

cannot be based solely on statistical grounds and how to interpret the differences can only be assessed by research outside the analysis of the responses themselves.

Summary

This article generalizes the formalization of artificial DIF induced by real DIF in the MH method of detecting DIF in dichotomous items (Andrich & Hagquist, 2012). In particular, in parallel to the exploitation of the dichotomous Rasch model to both distinguish between real and artificial DIF and to quantify the real DIF in dichotomous items, the PRM is used for the same two purposes with polytomous items. The basis for exploiting the Rasch model is that the total score of a person is the sufficient statistic for the person parameter, and this sufficiency justifies classifying persons by their total scores, or class intervals formed from their total scores.

To distinguish between real and artificial DIF, the article illustrates the logical implication of using a sequential resolution of items, beginning with the item with the largest magnitude of DIF as the item which is most likely to have real DIF. The resolution of an item with potential real DIF creates a new item for each group and eliminates any artificial DIF that was created by the original item. An example of a mental health questionnaire, in which from an initial analysis two items showed DIF with girls having higher scores and two items showed DIF with boys having higher scores, was shown to have just two items with real DIF, both having higher scores for girls.

A consequence of the sequential resolution of items, which creates an item for each group, is that the parameter estimates of these resolved items can be used to quantify the magnitude of the real DIF of the resolved item in the metric of the items with no DIF. Moreover, if DIF is the only source of violation of the responses to the model, misfit of the responses to the model because of DIF is eliminated. Although that kind of misfit is then eliminated, because the resolved item has different parameter estimates for each group, it does not have invariant scale values for the groups. These noninvariant scale values compensate for the effects of real DIF and reduce mean differences between groups.

It was suggested that if the real DIF of an item arises from an aspect irrelevant to the variable, then resolving it does not invalidate the comparison between means. However, it was also suggested that if the real DIF in an item arises from an aspect relevant to the variable, then resolving it can compromise the validity of the comparisons between group means. Of course both types of analyses can be made in reaching a decision as to which analysis is more valid, including studying the magnitude of the DIF, the effect resolving items has on the means and standard deviations of the groups, and the degree of improvement in the tests of fit. However, it was concluded that the choice of interpretation in any given situation cannot be determined by the statistical analyses alone, and that the full context of the data, and possible further substantive research, needs to be considered. It was noted that if items with real DIF are simply eliminated, then the same effect appears in the group means and on the validity of a comparison between groups. It was stressed that this choice of

interpretation is relevant when all persons are expected to respond to all items and that in the case that that the design does not require all persons to respond to all items, as when a relatively small number of common, link items are with a larger number of unique items administered to each group, then it is critical that the link items show no DIF.

Appendix A

This appendix shows the results of a simulation study with parameters equivalent to the estimates from the real example. Table A1 shows all item parameter estimates of the real example following the resolution of Items 7 and 4. In the simulation, the parameters for all items were the same for the boys and girls except for Items 7 and 4. In addition to the differences in the location parameters, there were differences in the threshold parameters, but the dominant differences between the boys and girls were the location differences in Items 7 and 4. Other than the differences between the parameters of Items 7 and 4 for the boys and girls, the responses were simulated to fit the PRM.

The same number of boys and girls, 301, with means and standard deviations of those obtained from the estimates in Table 7, and normally distributed, were used. These means and standard deviations were considered the correct ones statistically and ideal to demonstrate the validity of the assessment of artificial DIF.

Table A2 shows the ANOVA of the residuals in the simulated example. It is evident that the Item 7 has the greatest *F* value, followed by Item 4: this is consistent with the simulation parameters in Table A1. Both items show higher scores for girls. Item 8 shows a significant gender DIF at the .01 level, which is artificial DIF with boys having the higher scores. The other item that shows artificial DIF, though not quite at the significance level of .01, is Item 6. It will be recalled that the artificial DIF is distributed across all items. Therefore, which items show significant DIF will

Table A1. Location and threshold values of the real example and simulated examples.

Item	Location	Thresholds (sum to zero)			
		1	2	3	4
1. Concentrating?	−0.915	−2.991	−0.870	0.992	2.868
2. Sleeping?	−0.162	−1.989	−0.517	0.510	1.996
3. Headaches?	−0.306	−1.822	−0.594	0.401	2.016
4 B. Stomach aches?	0.552	−1.816	−0.344	0.361	1.800
4 G. Stomach aches?	0.316	−2.141	−0.935	0.508	2.568
5. Tense?	−0.045	−2.211	−0.553	0.746	2.018
6. Appetite?	0.313	−1.446	−0.429	0.616	1.259
7 B. Sad?	0.108	−1.711	−0.192	−0.056	1.958
7 G. Sad?	−0.444	−2.819	−0.785	0.638	2.966
8. Giddy?	0.584	−1.664	−0.727	−0.115	2.506

Table A2. ANOVA of Standardized Residuals for the Test of Fit Using Five CIs for the Simulated Example.

Item	Gender		Gender-by-CI		CI	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
1. Concentrating?	1.584	.209	0.605	.659	0.726	.575
2. Sleeping?	1.155	.283	3.170	.014	1.129	.342
3. Headaches?	0.001	.977	1.261	.284	0.362	.836
4. Stomachaches? (G > B)	15.276	.000	3.157	.014	2.874	.022
5. Tense?	2.064	.151	0.552	.698	3.884	.004
6. Appetite? (B > G)	3.293	.070	1.501	.200	0.853	.492
7. Sad? (G > B)	41.957	.000	3.044	.017	1.837	.120
8. Giddy? (B > G)	6.891	.009	1.267	.282	0.910	.458
	df: 1, 602		df: 4, 602		df: 4, 602	

Note. *p* < .01 taken as significant (shown in bold).

depend on other factors. In the case of the simulated items, where all items other than Items 7 and 4 have no DIF and fit the PRM, these are chance factors. The resolution of Items 7 and 4 and the effects on the person locations are also parallel to the real example and therefore are not repeated.

Auhtors’ Note

Barry Sheridan made constructive comments on an earlier draft of this article.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported in this article was supported in part by Australian Research Council Linkage grants with the School Curriculum and Standards Authority of Western Australia and the Australian Curriculum, Assessment and Reporting Authority as industry partners, and Pearson Plc and The Swedish Foundation for International Cooperation in Research and Higher Education.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69-81.
Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-574.

- Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37, 387-416.
- Andrich, D., & Luo, G. (2003). Conditional estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, 4, 205-221.
- Andrich, D., Sheridan, B., & Luo, G. (2013). RUMM2030: A Windows interactive program for analysing data with Rasch unidimensional model for measurement. Perth, Western Australia, Australia: RUMM Laboratory.
- Bland, J. M., & Altman D. G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, 3, 170.
- Brodersen, J., Meads, D., Kreiner, S., Thorsen, H., Doward, L., & McKenna, S. (2007). Methodological aspects of differential item functioning in the Rasch model. *Journal of Medical Economics*, 10, 309-324.
- Budgel, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, 19, 309-321.
- Camilli, G. (1993). The case against item bias techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum.
- Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Hillsdale, NJ: Prentice-Hall.
- Guttman, L. (1950). The problem of attitude and opinion measurement. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 46-59). New York, NY: Wiley.
- Hagquist, C. (2008). Psychometric properties of the PsychoSomatic Problems scale—A Rasch analysis on adolescent data. *Social Indicators Research*, 86, 511-523.
- Hagquist, C., & Andrich, D. (2004). Is the Sense of Coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. *Personality and Individual Differences*, 36, 955-968.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum.
- Ironson, G. H. (1983). Using item response theory to measure bias. In R. Hambleton (Ed.), *Applications of item response theory*. Vancouver, Canada: Education Research Institute, British Columbia.
- Karabatsos, G. (2000). A critique to Rasch residual fit statistics. *Journal of Applied Measurement*, 1, 152-176.
- Kreiner, S., & Christensen, K. B. (2011). Item screening in graphical loglinear Rasch models. *Psychometrika*, 76, 228-256.
- Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education*. Boston, MA: Houghton Mifflin.
- Looveer, J., & Mulligan, J. (2009). The efficacy of link items in the construction of a numeracy achievement scale—From kindergarten to year 6. *Journal of Applied Measurement*, 10, 247-265.

- Mantel, N., & Haenszel, M. W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.; Quantitative Applications in the Social Sciences, Vol. 161). Thousand Oaks, CA: Sage.
- Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19(3), 5-15.
- Potenza, M. T., & Dorens, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23-37.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. IV, pp. 321-334). Berkeley: University of California Press.
- Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalised formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational Measurement*, 24, 293-322.
- Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological measurement*, 48, 657-667.
- Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, 20, 1082-1084.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.
- Wang, W-C. (2000). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement*, 1, 63-82.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233-251.