Journal of Research in Education Sciences
2011, 56(1), 91-127

# Differential Item Functioning Analyses in Large-Scale Educational Surveys: Key Concepts and Modeling Approaches for Secondary Analysts

Xiao-Shu Zhu
Department of Measurement, Statistics & Evaluation, University of Maryland
Graduate Student

André A. Rupp
Department of Measurement, Statistics & Evaluation, University of Maryland
Associate Professor

Jing Gao
Office of Assessment & Evaluation, University College University of Maryland
Psychometrician

## Abstract

Many educational surveys employ a multi-stage sampling design for students, which makes use of stratification and/or clustering of population units, as well as a complex booklet design for items from an item pool. In these surveys, the reliable detection of item bias or differential item functioning (DIF) across student groups is a key component for ensuring fair representations of different student groups. In this paper, we describe several modeling approaches that can be useful for detecting DIF in educational surveys. We illustrate the key ideas by investigating the performance of six hierarchical generalized linear models (HGLMs) using a small simulation study and by applying them to real data from the Trends in Mathematics and Science Study (TIMSS) study where we use them to investigate potential uniform gender DIF.

**Keywords:** complex booklet design, DIF, HGLMs, multi-stage sampling design

Corresponding Author: Xiao-Shu Zhu, E-mail: xzhu16@umd.edu

National and international educational surveys are important empirical tools for monitoring student achievement in particular content domains such as mathematics, science, and reading. For instance, the Program for International Student Assessment (PISA) (http://nces.ed.-gov/assessments/pisa/) and the Trends in International Mathematics and Science Study (TIMSS) (http://nces.ed.gov/timss/) are well-established international surveys while the National Assessment of Educational Progress (NAEP) (http://nces.ed.gov/nationsreportcard/) is a well-established national survey in the United States whose technical developments have driven many of the current standards in like surveys around the world (e.g., Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992). There also exist numerous independent educational surveys conducted by provinces or states that are implemented at regular intervals.

These educational surveys are designed to support system-wide accountability systems, which requires that any inferences about mean performance differences across groups of students can be made reliably and validly (Rutkowski, Gonzalez, Joncas, & von Davier, 2010; von Davier, Sinharay, Oranje, & Beaton, 2006). Statistically, the examination of differential item functioning (DIF) is regarded as an important component in this overall process (Mapuranga, Dorans, & Middleton, 2008) and is reflected in various quality assurance statements in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Associating, & National Council on Measurement in Education, 1999).

Our goal for this paper is to introduce key concepts for DIF detection as well as six model specifications under a unified estimation framework for DIF detection within educational surveys, which is the framework of hierarchical generalized linear models (HGLMs) (e.g., Kamata, 2001; Prowker & Camilli, 2007; Raudenbush & Bryk, 2002). We investigate the practical utility of the six models using a small scale simulation study and demonstrate their use with real data from TIMSS 2007.

We have organized the paper as follows. In the first section, we provide an overview of key concepts for DIF detection, whether in educational surveys or other application contexts. In the second section we discuss important features of data collection procedures in educational surveys and the way in which they impact parameter and standard error / variance estimation in parametric statistical models. We then provide a rationale for a designed-based estimation process of DIF effects as implemented in the HGLM framework. In the third section, we conduct a small simulation study to investigate the DIF detection ability of six different HGLMs. In the fourth section we apply these models to a subset of the 2007 TIMSS data. We close the paper with a summary and brief discussion of key findings.

## Basic Concepts in DIF Detection

The literature on the theory and practice of DIF is vast and we recommend the following sources for further reading. For a general overview of DIF methods, we recommend, for example, Ferne and Rupp (2007), Mapuranga, Dorans, and Middleton (2008), Osterlind (2009), and Zumbo (1999). For an overview of key implications of the complex sampling designs for educational surveys for secondary analyses generally, we recommend the articles by Rutkowski et al. (2010), von Davier, Gonzalez, and Mislevy (2010), and von Davier et al. (2006). For examples of how DIF analyses can be conducted within unified estimation frameworks for parametric statistical models, which are most appropriate for educational survey data, we recommend Binici (2008), Kamata (2001), Kamata and Binici (2003), Kim (2003), and Prowker and Camilli (2007).

## DIF vs. Impact

DIF occurs when different *item response probabilities* are observed for students with identical levels of proficiency (i.e., equal values on the observed or latent variables in the statistical model) who belong to at least two distinct groups. Importantly, DIF is not the same as true mean differences proficiency, which are known as *impact* in the literature (Camilli & Shepard, 1994; Hauger & Sireci, 2008). Put differently, DIF reflects *conditional performance differences* whereas impact reflects *unconditional / marginal performance differences*. DIF can be viewed as caused by distributional differences on a variable that reflects a secondary construct that an instrument is not intended to measure but that the items that display DIF appear to require during responding. Hence, biased inferences for items displaying DIF will only result if the students in different groups actually differ in their proficiency distribution on the secondary variables (Shealy & Stout, 1993).

## Uniform vs. Non-uniform DIF

Researchers distinguish between two general types of DIF, which are known as *uniform DIF* and *non-uniform DIF*. Uniform DIF refers to the condition when one of the groups – typically the one denoted as the *reference group* – is predicted to perform either better or worse than the other group(s) – typically denoted as the *focal group(s)* – throughout the entire proficiency range. In contrast, non-uniform DIF exists when there is a point on the proficiency continuum where the predicted difference in performance across the groups reverts. This is important to remember because certain statistical approaches to DIF detection do not allow for the detection of non-uniform DIF. Importantly, this is equally true of some methods that employ *observed-score matching* (e.g., the Mantel-Haenszel method from the area of multivariate statistics / categorical data analysis) and

some methods that employ *latent-variable matching* (e.g., models in the Rasch family from the area of Item Response Theory (IRT)); for an overview of IRT see, for example, de Ayala (2009), Yen and Fitzpatrick (2006), and Embretson and Reise (2000).

## Grouping Structures in DIF Analyses

The reference and focal groups that are used in DIF analyses could be *observed* (e.g., gender groups, ethnicity groups) or *unobserved*, in which case the groups are often believed to represent students who utilize differential strategies for responding. In the first case, statistical models with categorical group indicators can be used while in the second case statistical models with latent classes need to be used, which are also known as *finite mixture models* (e.g., McLachlan & Peel, 2000).

Although two-group comparisons are commonly utilized in DIF studies for legislative or technical reasons (Zhang, Dorans, & Matthews-Lopez, 2005), this can sometimes be problematic because specific performance differences across different subgroups created by combining individual grouping variables (e.g., male Caucasians vs. female non-Caucasians) might not be fully teased apart (Mapuranga et al., 2008). Fortunately, multi-group comparisons with either one or multiple categorical variables can nowadays be carried out with relative ease by borrowing principles from factorial analysis-of-variance models (for a review of the latter see, e.g., Lomax, 2007).

## Matching Criteria for DIF Analyses

The criterion that is used to match the student groups to investigate potential conditional performance differences can technically be *external* or *internal* to an instrument under investigation, even though internal criteria are by far more frequently used in practice. External criteria could be observed total scores or estimated latent-variable scores from supplementary assessment batteries that measure constructs that are similar to the target construct of the instrument of interest; in contrast, observed total scores or estimated latent-variable scores frequently serve as internal matching criteria (Camilli, 1993; Camilli & Shepard, 1994; Clauser, Mazor, & Hambleton, 1993; Dorans & Holland, 1993; Holland & Thayer, 1988).

One potential challenge for using internal criteria is that DIF analyses might suffer from empirical and interpretational circularity (Camilli, 1993; Zenisky, Hambleton, & Rubin, 2003a) because matching scores are computed with information from potential DIF items included. One possible remedy is applying iterative strategies that remove such items and lead to a *purification of the matching criterion / criterion refinement* (Clauser et al., 1993; Holland & Thayer, 1988; Holland

& Wainer, 1993; Zumbo, 1999). In contrast, a confirmatory application of procedures such as SIBTEST ( Shealy & Stout, 1993) requires the a priori specification of a subtest for matching for which it is known that the constituting items do not show DIF.

## Statistical Models and Methods for DIF Detection

As already alluded to, it is helpful to differentiate conceptually between *parametric statistical models* for DIF detection, which assume a particular distributional form for the item response variables (e.g., a Bernoulli or multinomial distribution for each variable) and *non-parametric statistical methods*, which do not. Since parametric latent-variable models are the state-of-the-art for analyzing data from educational surveys in light of the complex sampling design for students and the complex booklet design for items, these models are typically also used exclusively for DIF analyses in operational stages for such surveys.

The difference between observed-score matching and latent-variable score matching has immediate implications for the statistical models that can be used. Nowadays, most parametric modeling approaches are viewed as special cases of more general modeling frameworks. Many parametric models, including the ones formulated within the HGLM framework that is the focus of this paper, can be cast as special cases within a Generalized Linear Mixed Model (GLMM) framework (e.g., Goldstein, 2003; Raudenbush & Bryk, 2002) or an even broader Generalized Linear Latent and Mixed Model (GLLAMM) framework (e.g., Binici, 2008; Skondral & Rabe-Hesketh, 2004).

## Statistical Detection versus Substantive Explanation

It should be noted that DIF analyses are purely empirical exercises in that certain statistical models or methods are applied to data and certain items or groups of items are then flagged. What these basic analyses frequently do not provide is either an explanation of what might have led to the conditional performance differences or a prescription of what should be done with the items that are flagged as displaying DIF.

However, with regards to the former, substantive hypotheses about potential causes of DIF can nowadays be empirically operationalized by including predictor variables for the variation in item parameters in statistical models (e.g., surface-structure item design variables, deep-structure cognitive complexity variables) even though that is not necessarily always done in practice. Substantive hypotheses about potential causes for impact can be simultaneously investigated by including predictor variables for the variation in student parameters (i.e., observed or estimated proficiency scores) in statistical models. The inclusion of covariates at both of these levels is

nowadays relatively straightforward within the unified estimation frameworks mentioned earlier and is known in the latent-variable modeling literature specifically as *explanatory item response modeling* (De Boeck & Wilson, 2004).

## Summary

To conceptually understand and judiciously apply particular methods for DIF detection, it is important to be clear about: 1. whether DIF for two or multiple groups is of concern, 2. whether these groups are observed or unobserved, 3. whether uniform or non-uniform DIF is likely to be present, 4. whether external matching criteria are available and of interest for use, 5. whether single or multiple matching variables are available, 6. whether the matching criterion is / the matching criteria are observed total score(s) or estimated latent-variable score(s), (7) whether parametric or non-parametric statistical models can or should be used for DIF detection, and (8) whether covariates for explanatory purposes at the item or student levels can or should be directly or indirectly included in the analysis.

Since educational surveys contain sampling complexities at both the student and the item level, certain statistical modeling approaches for DIF detection – non-parametric statistical models in particular – are not appropriate if integrated DIF analyses are desired. In the following section we outline some of the key implications of these complexities for DIF analyses and focus specifically on DIF detection with parametric models within a unified estimation framework.

## DIF Detection in Educational Surveys

### *Complex Sampling Designs for Students and Sampling Weights*

As discussed in Rutkowski et al. (2010), educational surveys usually employ a *multistage cluster sampling design* that additionally uses stratification at one or multiple stages. For instance, NAEP utilizes a two-stage sampling design: the first stage involves the selection of schools within strata, and the second stage involves the selection of students within schools (von Davier et al., 2006). Generally speaking, multistage cluster sampling designs are employed because they reduce the cost of data collection and make the surveys practically feasible even though they typically increase the sampling error compared to simpler sampling designs.

Specifically, *stratification* is a process of grouping students into relatively homogenous subgroups before sampling at a particular level. *Cluster sampling* refers to the selection of sets of units (e.g., districts, schools, or classrooms within schools) rather than individual students. Further information about the particularities of sampling designs for educational surveys can be found in

Rutkowski et al. (2010), von Davier et al. (2006), and the technical reports for educational surveys that can be found on the websites we referenced above.

With multi-stage sampling designs, *sampling weights* are used to adjust parameter estimates for differences in the probability of selecting individual sampling units. Each sampling unit is assigned a *base weight*, which is the inverse of the overall initial probability of selection. For example, if a student is selected with probability of 1/100 during sampling, then that student represents 100 students in the target population. The sum of the base weights of all the sampled students is an unbiased estimate of the total number of students in the target population.

Developers of large-scale educational surveys provide various types of sampling weights in publicly available databases. For instance, five main types of sampling weights are available in the TIMSS 2007 database: *total student weight*, *student house weight*, *student senate weight*, *school weight*, and *teacher weight* (Rutkowski et al., 2010). The *total student weight* is the inverse of the overall probability of a student being selected. The *student house weight* is a linear transformation of the total student weight such that the sum of these weights is equal to the observed sample size; it eliminates the inflation of degrees of freedom that would result if the total student weight were used. The *senate weight* is the student total weight rescaled such that all students' senate weights sum to 500 in each country; it is useful when comparing statistics from countries of different population sizes. When senate weights are properly used, the country with a considerably larger population does not dominate any statistical analyses.

Due to the fact that TIMSS also reports on school, teacher, and classroom characteristics aside from student characteristics, a *school weight* and a *teacher weight* are also provided. The *school weight* is the inverse of the probability of selection for a specific school while the *teacher weight* is the student total weight divided by the total number of teachers a specific student has. Depending on the focus of analysis (e.g., capturing the directionality, magnitude, and significance of effects for students, teachers, or schools) researchers need to either select or, potentially, compute weights at appropriate sampling levels to accommodate features of nested data.

Before assigning a weight to a particular level of analysis with a statistical analysis software program for survey data (e.g., STATA, SAS assessment procedures, R 'assessment' package, SUDAAN, WesVar, AM), it is important to check program manuals to ensure the software's definition of weights at different levels agrees with those listed in the data. For example, in a two-level HGLM, the level-1 student weight should be the inverse of the conditional probability of the student being selected given his or her school was selected. However, the total student weight in the TIMSS data is the inverse of the joint probability of a certain student and the school that he or

she attends being selected and is, thus, not directly appropriate. When appropriate weights are not readily available, researchers should be prepared to manually calculate weights for each level (see Rutkowski et al., 2010, for illustrations).

## Complex Sampling / Booklet Designs for Items

### *Correct Estimation of Mean Proficiency Differences via Plausible Values*

The design of any large-scale educational survey requires a fine balance between accommodating theoretical desiderata for reliably and validly assessing student performance and accommodating practical constraints regarding implementation of data-collection schemes at the same time. Ideally, survey developers would like to make very precise inferences about the proficiency of individual students while (1) ensuring a broad coverage of tested domains for each individual student and (2) simultaneously testing a large cross-section of the student population. However, since resources such as testing time or money for administering and scoring responses are finite, there exists a notable tension among these desiderata (Adams & Gonzalez, 1996).

As a solution, complex item sampling designs, which are also known as *matrix sampling designs* or *booklet designs*, are commonly employed in educational surveys. In these designs, subsets of items are selected from the total item pool and are typically arranged into *blocks* corresponding to a certain administration time (e.g., 15 or 20 minutes), which are then assigned to the *survey forms* or *test booklets* using a particular design structure (see Frey, Hartig, & Rupp, 2009, for a didactic introduction to such designs). All educational surveys use multiple booklets (e.g., there were 14 booklets in TIMSS 2007 and 13 booklets in PISA 2009) and each booklet typically contains items from multiple content domain(s) of interest (e.g., reading, mathematics, and science).

Educational surveys are designed to provide reliable inferences at aggregate levels (e.g., school districts, countries, student groups defined otherwise). That is, resulting estimates of proficiency should not be used to make inferences about individual students because these proficiency estimates are far too imprecise / unreliable (Rutkowski et al., 2010; von Davier et al., 2010; von Davier et al., 2006). This is a result of the booklet design, because each individual student only responds to few items from each domain, often targeted at a relatively broad range of ability, resulting in very few well-targeted statistical pieces of information that are available for domain-specific subscore estimation for each student. Put differently, data from booklets typically match the mean proficiency level of groups of students (e.g., students at particular school types or in particular countries) relatively well, but match the proficiency levels of individual students relatively poorly.

Consequently, secondary analyses that focus on mean proficiency differences for groups of

students use alternative estimates of proficiency known as *plausible values* (e.g., Mislevy, 1991; Mislevy & Sheehan, 1987; Rubin, 1987; von Davier et al., 2010; von Davier et al., 2006). The statistical machinery for estimating plausible values was first developed for NAEP in 1983 and is the current state-of-the-art for many educational surveys. Unlike the single latent variable values estimated in a traditional IRT model, plausible values are *multiple imputations* of the single latent variable value that needs to be estimated for a particular reporting dimension for each student. Put even more technically, they are multiple random draws from an appropriate *posterior distribution of latent variable values* for each individual student, which is adjusted for the influence of a wide range of covariates on student performance.

Typically, published data sets for educational surveys contain several – often five – plausible values and proper statistical analyses require an *integrated analysis* that aggregates and synthesizes the point estimates and standard errors from separate analyses, each one run on a different plausible value. For example, Rutkowski et al. (2010) discussed how to combine the results of each analysis into a single set of point estimates and standard errors using Rubin's (1987) multiple imputation formulas. This aggregation and synthesis strategy avoids potential problem of underestimating the standard errors of the statistics of interest if only one of plausible values or the mean point estimates across the plausible values were used.

## *Model-Based Standard Error Estimation*

The assumption of *independent responding* that underlies traditional statistical analyses (i.e., of a response process for individuals that is not influenced by context effects such as a common educational environment) is violated when data are collected using the complex sampling designs that are common to educational surveys; consequently, analytic procedures must be adjusted to appropriately address the hierarchical structure of the sample.

The ratio of the adjusted variance for a parameter estimate and the unadjusted variance that is produced when treating a complex survey sample as a simple random sample is called the *design effect*. When the *design effect* of a survey is not considered in the statistical analysis, the results underestimate variances and consequently inflate type-I errors (Cochran, 1977). For example, in DIF analyses for educational surveys, more items than appropriate may be flagged for DIF if the sampling design is not accounted for by the statistical analysis.

Two approaches exist for variance estimation under a complex sample design, which are grounded in *design-based* and *model-based inference* (Kalton, 1983). In a design-based approach, post-hoc adjustments are made to unadjusted variance estimates to account for the complex sampling structure. For the analyses that follow, we adopt a model-based approach, which directly

decomposes the observed variance of response variables into contributions at various design levels (e.g., classrooms, schools) and, thus, represents the various sampling levels with their stratification variables directly in the statistical model. Consequently, corrected standard errors are produced directly.

A commonly used statistical framework for model-based DIF analysis in general is the HGLM framework that we mentioned earlier. Various general-purpose latent-variable programs can nowadays estimate HGLMs using, for example, *pseudo-likelihood methods* (Binder, 1983; Pfefferman, Skinner, Holmes, Goldstein & Rasbash, 1998), amongst them SAS PROC GLIMMIX in SAS, MPlus (Muthén, L. K. & Muthén, B. O., 2007) and HLM (Raudenbush, Bryk, Cheong, Congdon & du Toit, 2004). Such programs have options for specifying sampling weights at multiple design levels as well as associated item and student covariates.

# HGLM Model for DIF Detection

Statistical models within an HGLM framework are relatively simple to specify and estimate in user-friendly general-purpose latent-variable software such as SAS 9.2, Mplus 5.0 or HLM 6.0. Importantly, these programs can accommodate data that are *missing randomly* and *by design*, both of which are common for educational surveys, as well as plausible value estimation. In the following sections, we present four candidate models that have been proposed for DIF detection within the HGLM framework – two of them in two variations resulting in a total of six models – and assess their efficacy in detecting DIF using a small simulation study. Our key objective is to provide readers with an intuition about the underlying specification principles that they embody, rather than to provide a seemingly complete answer regarding their utility for DIF detection under a wide range of conditions.

## A Primer on the Basic Ideas for DIF Detection: The One-level Logistic Regression Model

Swaminathan and Rogers (1990) proposed the use of logistic regression analysis for DIF analysis, which allows for the simultaneous detection of uniform and non-uniform DIF depending on which model is specified. For a single dichotomously scored item (i.e., a response variable with scores 0 and 1 that follows a Bernoulli distribution), the probability ( $p_{ij}$ ) for an individual $j$ to get a correct response on item $i$ depends on his or her ability level and his or her group membership; these elements are connected via a *log-odds* or *logit link* function ( $\eta_{ij}$ ) as follows:

$$\log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \beta_0 + \beta_{1j}W_j + \beta_{2j}G_j + \beta_{3j}(WG)_j \quad (1)$$

where $W_j$ is the matching variable for student $j$ (e.g., the observed total score or the estimated ability level), $G_j$ denotes the group membership of student $j$ (i.e., 0 for the reference group and 1 for the focal group) and $(WG)_j$ denoted the product / interaction term of these two variables.

In the *linear predictor* on the right-hand side of the equation, the intercept parameter $\beta_0$ corresponds to the conditional log-odds of a correct response for the reference group, which indicates the "difficulty level" of this item; thus, a higher value indicates a greater probability for correctly answering item $j$ for students in the reference group. The parameter $\beta_1$ models a main effect of the matching variable (i.e., the influence of proficiency differences on item performance) whose inclusion as a covariate represents the matching step in DIF analyses.

DIF analyses mainly focus on the coefficients for the grouping variable. The parameter $\beta_2$ corresponds to a group main effect (i.e., the difference in the conditional log-odds of a correct response between the reference and focal groups) so that a significant non-zero value indicates uniform DIF. If non-uniform DIF is of the interest, the $\beta_3$ parameter associated with the interaction effect term is used so that a significantly non-zero value for $\beta_3$, or a significant joint test of $\beta_2$ and $\beta_3$, indicates non-uniform DIF.

The HGLM framework that we describe in the following expands these ideas. It can be used for two distinct purposes as far as DIF detection is concerned, which are to specify and estimate (1) one- and two-parameter hierarchical IRT models for DIF detection, (2) hierarchical logistic regression models for DIF detection, and (3) hybrid IRT and logistic regression models. Since our focus is on data from educational surveys we focus specifically on simple cases of models (1) and (3) in the following.

## Kamata's (2001) Rasch Model Formulation as a HGLM

A traditional unidimensional IRT model can be viewed as the simultaneous estimation of several simple logistic regression models with a single latent predictor variable. In prototypical IRT models, the latent variable values vary randomly across students to model differing proficiency levels of students while the item parameter values are fixed to model that a single fixed form is given to a subset of students. A basic IRT model is, by definition, equivalent to a two-level HGLM. Consequently, DIF analyses can be conducted within a HGLM framework, which possess the additional advantages that more levels can be added to the model to accommodate more complex nested data structures. Moreover, covariates can be included at different model levels to account for

DIF or proficiency differences of students.

The utility of the HGLM framework for DIF analyses was original outlined by Kamata (2001) who introduced a hierarchical Rasch model formulation into the literature. In Kamata's level-1 model, one item serves as *reference item* with its difficulty set to '0' and the other items are represented via dummy variables. For a single dichotomously scored item (i.e., a response variable with scores '0' and '1' that follows a Bernoulli distribution), the probability ( $p_{ij}$ ) for student $j$ to give a correct response to item $i$ is expressed via a logit link function ( $\eta_{ij}$ ) as

$$\begin{aligned} \log\frac{p_{ij}}{1-p_{ij}} = \eta_{ij} &= \beta_{0j} + \beta_{1j}Z_{1ij} + \beta_{2j}Z_{2ij} + \ldots + \beta_{(k-1)j}Z_{(k-1)ij} \\ &= \beta_{0j} + \sum_{q=1}^{k-1}\beta_{qj}Z_{qij} \end{aligned} \tag{2}$$

where $\beta_{0j}$ is the level-1 intercept or student main effect. For $K$ items, $K-1$ dummy variables ( $z_{qij}$ ) are included at the level-1 model, each of which is coded '1' for item $q$ and '0' otherwise to allow for the possibility that not every item is presented to every student. Consequently, $\beta_{qj}$ corresponds to the difficulty parameter for item $q$, which is interpreted relative to the difficulty of the reference item.

Since proficiency varies across students, $\beta_{0j}$ is modeled as a random effect at level 2. In contrast, since item difficulty remains constant across students, the $\beta_{qj}$ parameters are modeled as fixed effects at level 2. When recasting the level-2 model for DIF analyses, a group indicator variable $G_j$ needs to be included similar to the one-level logistic regression model above:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}G_j + u_{0j} \tag{3}$$

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}G_j \tag{4}$$

where $\gamma_{00}$ is the mean proficiency estimate of the reference group, $\gamma_{01}$ denotes the mean difference between the focal and reference group, and $u_{0j}$ is the random effect that follows a normal distribution with a mean of 0 and a variance of $\tau_{00}$.

Similarly, $\gamma_{q0}$ is the relative difficulty parameter of item $q$ for the reference group, and $\gamma_{q1}$ is the difference in relative item difficulty between the reference and focal group for item $q$ (i.e., the effect size of DIF for item $q$ relative to the reference item). Taken together, these equations can be expressed as a single regression equation as follows:

$$\log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \gamma_{00} + \gamma_{01}G_j + \sum_{q=1}^{K-1}(\gamma_{q0} + \gamma_{q1}G_j)Z_{qij} + u_{0j} \quad (5)$$

which shows the simultaneous adjustment for group differences in mean proficiency (i.e., impact) via $\gamma_{01}$ and group differences in item difficulty (i.e., DIF) for item $q$ via $\gamma_{q1}$.

Since the model formulation in equations 2-4 is for the Rasch model, only uniform DIF can be examined. This model is theoretically straightforward to specify and can be easily estimated using general-purpose latent-variable software programs such as SAS 9.2, Mplus 5.0 and HLM 6.0.

## Kim's (2003) Simultaneous Logistic Regression Model within the HGLM Framework

Kim (2003) formulated a hybrid of a logistic regression and an IRT model for DIF detection as a hierarchical model within the HGLM framework. This allowed for the detection of non-uniform DIF while using a Rasch-like modeling basis, which has some potential advantages as HGLM formulations for two-parameter IRT models that could be used to test for non-uniform DIF are rather complicated to set up. The resulting HGLM formulation also allows for a simultaneous estimation of the mean proficiency difference across groups (i.e., impact) as in Kamata's Rasch model, which would have to be estimated separately (e.g., via a two-sample $t$-test on observed total scores) if traditional logistic regression models were used for each item.

The HGLM formulation of this hybrid model thus mimics the formulation of Kamata's model and the one-level logistic regression model from above. The level-1 regression equation is the same as in Equation 3. However, the level-2 equations for the slope parameters differs from Equation 4 and is the same as in the one-level logistic regression model in Equation 1:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}G_j + u_{0j} \quad (6)$$

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}W_j + \gamma_{q2}G_j + \gamma_{q3}(WG)_j \quad (7)$$

Similar to the one-level logistic regression model, the regression coefficient $\gamma_{q2}$ is used to examine uniform DIF while the regression coefficient $\gamma_{q3}$, or both coefficients jointly, are used to test for non-uniform DIF. These equations can be expressed as single model equation as follows:

$$\log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \gamma_{00} + \gamma_{01}G_j + \sum_{q=1}^{K-1}(\gamma_{q0} + \gamma_{q1}W_j + \gamma_{q2}G_j + \gamma_{q3}(WG)_j)Z_{qij} + u_{0j} \quad (8)$$

which shows the simultaneous adjustment for group differences in mean proficiency (i.e., impact) via $\gamma_{01}$, and conditional group differences in item difficulty (i.e., DIF) for item $q$ via $\gamma_{q2}$ and $\gamma_{q3}$. Moreover, it shows that the random effect $u_{0j}$ in Equations 6 and 8 is the conditional proficiency level of an individual student after a matching variable for DIF has been included in the item parameter model at level 2 – arguably not as easily interpretable as in Kamata's model though.

However, just as Kamata's original model, it has been criticized for requiring an arbitrary specification of a reference item which may have an undesirable impact on the DIF detection for other items if the reference item is itself subject to DIF; this is addressed with the following two models.

## *Pan's Reparameterizations of Kamata's and Kim's Model Formulations as HGLMs*

Pan (2008) argued that deleting the intercept parameter $\gamma_{00}$ in the level-2 model while keeping dummy variables for all items yields an equivalent formulation of the above models for DIF detection without the need to specify a desirably non-DIF item as a reference item. Pan reparameterized Kamata's and Kim's models with $K$, instead of $K$–1, dummy variables ( $Z_{qij}$ ) in the level-1 model as follows:

$$\begin{aligned}\eta_{ij} &= \beta_{0j} + \beta_{1j}Z_{1ij} + \beta_{2j}Z_{2ij} + \ldots + \beta_{(k-1)j}Z_{(k-1)ij} + \beta_{kj}Z_{kij} \\ &= \beta_{0j} + \sum_{q=1}^{k}\beta_{qj}Z_{qij}\end{aligned} \tag{9}$$

The level-2 model for the intercept was now written as

$$\beta_{0j} = u_{0j} \tag{10}$$

in both models while the slope expressions remained as in Equations 4 and 7 for Kamata's and Kim's models, respectively.

Again, these equations can be expressed as single model equations; Pan's reparameterization of Kamata's model is

$$\log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \sum_{q=1}^{K}(\gamma_{q0} + \gamma_{q1}G_j)Z_{qij} + u_{0j} \tag{11}$$

while Pan's reparameterization of Kim's model is

$$\log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \sum_{q=1}^{K}(\gamma_{q0} + \gamma_{q1}W_j + \gamma_{q2}G_j + \gamma_{q3}(WG)_j)Z_{qij} + u_{0j} \quad (12)$$

Pan's reparameterized version of Kamata's model is still a Rasch model, except that the random effect is a direct estimate of the proficiency level for student $j$ rather than an indirect estimate expressed as a difference from the mean proficiency. Similarly, the $\gamma$ s now directly represent the difficulty parameters for specific items, rather than the difficulty parameters relative to the reference item.

## *Summary and Discussion*

In summary, the four core models for DIF detection within an HGLM framework mentioned above include Kamata's model and Pan's reparameterization of Kamata's model, both of which are Rasch models, and Kim's model and Pan's reparameterization of Kim's model, both of which can be viewed as hybrid versions of logistic regression models and Rasch models.

Importantly, there is an additional methodological question for Kim's model and Pan's reparameterization of it, which is whether total score estimates or latent-variable estimates should be used as internal matching variables. For complex booklet designs, the use of booklet total scores is tedious and even less desirable than the use of plausible values. Of course, both of these methods seem intuitively less desirable than doing a direct DIF analysis within an IRT framework. To investigate the degree to which these intuitions are true for DIF analysis we have designed a small simulation study that we present in the next section.

We also note that all four HGLMs presented in this section treat the DIF effect as a fixed effect across any potential higher-order sampling units. However, as discussed in the previous section, the nature of sampling design in large-scale surveys results in the fact that responses from students within the same school and / or school district and / or country are contextually dependent upon one another even after proficiency differences have been accounted for. As a result, the two-level HGLMs can be extended to three-level HGLMs to investigate whether the magnitude of DIF varies across schools even though this is not the focus of our paper.

# Simulation Study Rationale and Design

## Objective of Simulation Study

No simulation study has been conducted so far to investigate the statistical performance of the four core HGLMs in the previous section – and their variations using booklet total scores and plausible values – in relation to one another under different conditions for DIF. Thus, the three main purposes of our simulation study were: 1. to investigate whether using a non-DIF item or a DIF item as the reference item in Kamata's and Kim's original models impacts their DIF detection ability; 2. whether using booklet total scores or estimated plausible values in Kim's model and Pam's reparameterization of it impacts their DIF detection ability; and 3. which of the six models perform(s) superior overall across a range of conditions.

## Design of Simulation Study

We employed a total of seven models, which were: 1. the Rasch model with a mean item parameter constraint for identification as estimated via marginal maximum likelihood in Conquest 2.0; 2. Kamata's Rasch model formulation as estimated via pseudo maximum likelihood within the HGLM framework with a reference item constraint for identification (Kamata); 3. Kim's model with a reference item constraint for identification as estimated via pseudo maximum likelihood within the HGLM framework and with booklet total scores as matching variables (Kim-BTS); 4. Kim's model with a reference item constraint for identification with plausible values as matching variables as estimated via pseudo maximum likelihood within the HGLM framework and (Kim-PV); 5. Pan's reparameterization of Kamata's Rasch model without a reference item constraint as estimated via pseudo maximum likelihood within the HGLM framework (Pan-Kamata); 6. Pan's reparameterization of Kim's model without a reference item constraint and with booklet total scores as matching variables as estimated via pseudo maximum likelihood within the HGLM framework (Pan-Kim-BTS); and (7) Pan's reparameterization of Kim's model without a reference item constraint and with plausible values as matching variables as estimated via pseudo maximum likelihood within the HGLM framework (Pan-Kim-PV). The six HGLMs with their two-level specifications are summarized in Table 1.

Table 1 Specifications of Six Candidate Models within HGLM Framework

| Models | Reference Item | Reference Item Type | # of Dummy Variables | Matching Variable | Model Specification | |
|---|---|---|---|---|---|---|
| | | | | | Level-one (Item) | Level-two (Student) |
| Kamata | Yes | Non-DIF DIF | $K$-1 | None | $\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} Z_{qij}$ | $\beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j}$<br>$\beta_{qj} = \gamma_{q0} + \gamma_{q1} G_j$ |
| Kim-BTS | Yes | Non-DIF DIF | $K$-1 | Booklet total scores | $\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} Z_{qij}$ | $\beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j}$<br>$\beta_{qj} = \gamma_{q0} + \gamma_{q1} W_j + \gamma_{q2} G_j$ |
| Kim-PV | Yes | Non-DIF DIF | $K$-1 | Plausible values | $\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} Z_{qij}$ | $\beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j}$<br>$\beta_{qj} = \gamma_{q0} + \gamma_{q1} PV_{mj} + \gamma_{q2} G_j \quad (m = 5)$[a] |
| Pan-Kam | No | -- | $K$ | None | $\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k} \beta_{qj} Z_{qij}$ | $\beta_{0j} = u_{0j}$<br>$\beta_{qj} = \gamma_{q0} + \gamma_{q1} G_j$ |
| Pan-Kim-BTS | No | -- | $K$ | Booklet total scores | $\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k} \beta_{qj} Z_{qij}$ | $\beta_{0j} = u_{0j}$<br>$\beta_{qj} = \gamma_{q0} + \gamma_{q1} W_j + \gamma_{q2} G_j$ |
| Pan-Kim-PV | No | -- | $K$ | Plausible values | $\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k} \beta_{qj} Z_{qij}$ | $\beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j}$<br>$\beta_{qj} = \gamma_{q0} + \gamma_{q1} PV_{mj} + \gamma_{q2} G_j \quad (m = 5)$[a] |

[a] indicates that for models with PVs results were aggregated across five runs.

## Fixed Design Factors

We fixed the number of groups for DIF analyses to two to keep the complexity of the study manageable. To mimic the complex sampling design for students in large-scale survey, we specified that all students from the two student groups were clustered within five schools. For each student group, we specifically generated100 students per school resulting in a total sample size of 1,000. Variations in the school and individual weights were not considered in the current simulation study.

As shown in Table 2, we implemented a complex booklet design for items by generating a total of 18 items and constructing two booklets from them; we specified each booklet to contain 12 items with six items shared by the two booklets. We specified the item parameters within each booklet to range from -1.50 to 1.50 with parameters for anchor items ranging from -0.50 to 0.50 in accordance with a discrete approximation to a normal distribution for item parameters within each booklet. Among the 18 items, we selected four items as DIF items; we specifically selected one easy item (Item 3), two moderately difficult items (Item 8 and Item 10) and one difficult item (Item 16) as displaying DIF.

Table 2 Item Difficulty Parameters for Item Generation

| Unique Booklet 1 Items | | Common Items | | Unique Booklet 2 Items | |
|---|---|---|---|---|---|
| Item | $\beta_i$ | Item | $\beta_i$ | Item | $\beta_i$ |
| Item I | -1.50 | Item 7 | -0.50 | Item 13 | -1.50 |
| Item 2 | -1.00 | Item 8[a] | -0.25 | Item 14 | -1.00 |
| Item 3[a] | -0.75 | Item 9 | 0 | Item 15 | -0.75 |
| Item 4 | 0.75 | Item 10[a] | 0 | Item 16[a] | 0.75 |
| Item 5 | 1.00 | Item 11 | 0.25 | Item 17 | 1.00 |
| Item 6 | 1.50 | Item 12 | 0.50 | Item 18 | 1.50 |

[a] indicates items for which DIF was introduced.

We completed a total of 30 replications for each condition. We investigated the DIF detection ability of each model by computing the empirical type-I error rates for non-DIF items as well as the empirical power for DIF items using the 5% cut-offs suggested by the theoretical sampling distributions for the model parameters under the null hypothesis of no DIF. The rates were computed as averages over all items of a particular type (i.e., non-DIF and DIF items) and, for models with plausible values as matching variables, across all five runs for the five plausible values.

## Manipulated Design Factors

We specified two types of reference item conditions, one where the reference item did not display DIF and one where the reference item did display DIF; in both conditions we chose a moderately difficult item as the reference item – Item 9 and Item 10, respectively – to allow for a sufficient amount of response variance. We also manipulated the magnitude of DIF at two levels, a moderate effect size of .50 and a strong effect size of .75, which was applied equally to all DIF items. We also manipulated the proficiency score distribution of the focal group at three levels with $N$(-.5, 1), $N$(0, 1) and $N$( .5, 1) and fixed the distribution of the reference group to $N$(0, 1). The purpose of manipulating the ability distribution was to investigate how well the different models are able to detect DIF in the presence of impact.

Clearly, the above design is not an exhaustive treatment of all possible design factors and their levels that might influence the ability of the models to detect DIF; instead, we wanted to use the design as a consciousness-raising device so that readers can understand some of the basic issues in DIF analyses and interested researchers can extend this simulation to a wider set of conditions.

## Data Generation

The model that we used to generate data was the three-level model presented in Binici (2008). In this model, the log-odds of a correct answer to item $i$ by student $j$ in school $k$ is given by

$$\eta_{ijk} = b_i + aG_{jk} + c_i G_{jk} + u_{1jk}^{(2)} + u_{1k}^{(3)} + u_{2k}^{(3)} G_{jk} \tag{13}$$

where $b_i$ represents the difficulty of item $i$; $G_{jk}$ is the dummy variable for group indicator with value '1' when student $j$ in school $k$ belongs to focal group and '0' otherwise; $a$ is the main group effect representing the mean proficiency difference between the focal and reference groups; $c_i$ is the interaction effect between item and student variables, which is the difference in item difficulty between focal and reference group or the DIF effect size for item $i$.

The last three terms are the random effects corresponding to the three model levels. Specifically, $u_{1k}^{(3)}$ represents the random effect associated with school $k$ and can be interpreted as the mean ability of individuals in school $k$; $u_{1jk}^{(2)}$ represents random effect of students in schools and can be interpreted as the proficiency of student $j$ in school $k$ expressed as a difference from the average ability of students in that school; and $u_{2k}^{(3)}$ represents the random DIF effect, which can be interpreted as the difference of the item difficulty parameter value for school $k$ from the average item difficulty parameter value across all schools.

We assume that the $u_{1k}^{(3)}$ and $u_{2k}^{(3)}$ are sampled from a bivariate normal distribution with the following mean vector and variance-covariance matrix:

$$MVN\left(\begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}\sigma_{1k}^2 & \\ \sigma_{1k2k} & \sigma_{2k}^2\end{bmatrix}\right)$$

We specified the variance-covariance matrix as

$$\Sigma=\begin{bmatrix}0.010 & \\ 0.004 & 0.010\end{bmatrix}$$

which results in a moderate correlation coefficient between the random school and DIF effects ($\rho=\frac{0.004}{\sqrt{0.01\times 0.01}}=0.4$). Finally, we independently sampled the proficiency estimates for student $j$ in school $k$ from a standard normal distribution $N=(u_{1k},\ 1-\sigma_{1k}^2)$. We simulated the data in R and obtained the plausible value estimated via ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) using the standard Rasch model specification in this software, which constrains the mean item parameter value to '0'. We then used SAS PROC GLIMMIX, which uses pseudo maximum-likelihood estimation, to specify and estimate the six proposed HGLMs. For Kim's model with plausible values and Pan's modification of it, we estimated the model five times, each time with one plausible value, and then aggregated all counts for power and type-I error rates across the five runs.

# Simulation Study Results

## Reference Item Effect

The first issue of interest was whether DIF in the reference item had an effect on the power and type-I error rates for Kamata's and Kim's original model formulations that included a reference item; the resulting values are shown in Table 3 for each combination of the DIF effect size and ability distribution of the focal group.

In short, as one would expect, using a DIF item as a reference item has a negative impact on DIF detection overall. When DIF is present for the reference item, Kamata's and Kim's models were not only unable to effectively detect other DIF items but also incorrectly identify most non-DIF items as displaying DIF. This pattern existed largely independent of proficiency distribution differences and DIF effect sizes; some differences existed in places but those were largely

Table 3 Power and Type-I Error as a Function of DIF Effect Size, Proficiency Distribution Difference, and DIF in Reference Item

| DIF Effect Size | Reference Item | Model | Proficiency Distribution of Focal Group (Reference Group Distribution is $N(0, 1)$) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $N(-0.5, 1)$ | | $N(0, 1)$ | | $N(0.5, 1)$ | |
| | | | Power | Type-I Error | Power | Type-I Error | Power | Type-I Error |
| .50 | None | Pan | .82 | .64 | .61 | .23 | .19 | .59 |
| | | Rasch | .69 | .16 | .67 | .17 | .78 | .18 |
| | Non-DIF | Kamata | .53 | .05 | .60 | .05 | .65 | .04 |
| | | Kim | .28 | .30 | .50 | .08 | .74 | .16 |
| | DIF | Kamata | .08 | .56 | .01 | .54 | .04 | .55 |
| | | Kim | .28 | .51 | .06 | .30 | .18 | .08 |
| .75 | None | Pan | .92 | .69 | .86 | .31 | .45 | .73 |
| | | Rasch | .91 | .25 | .91 | .26 | .94 | .28 |
| | Non-DIF | Kamata | .85 | .05 | .84 | .03 | .87 | .03 |
| | | Kim | .47 | .38 | .71 | .16 | .91 | .15 |
| | DIF | Kamata | .07 | .85 | .06 | .81 | .07 | .85 |
| | | Kim | .23 | .52 | .07 | .42 | .22 | .21 |

*Note.* Power and type-I error are computed over all items displaying or not displaying DIF, respectively, excluding the reference item, as well as all replications for plausible values. The row for Pan's model contains aggregated information from Pan's reparameterized versions of Kamata's model and both matching variable versions of Kim's model. The row for Kim's model contains aggregated information for both matching variable versions for Kim's model.

overshadowed by the overwhelmingly high type-I error rates and relatively low power. This pattern was even worse for Kamata's model than for Kim's model as the former had a less than 10% of chance to detect DIF in other items with even more severely inflated type-I error rates.

In contrast, when the reference item did not display DIF, Kamata's model performed overall best in that it controlled the nominal type-I error rate while showing moderate power across all proficiency distribution conditions. As expected, the power increased as the DIF effect size increased as well. Given the distinct adverse impact of using a DIF item as a reference item, the following analyses exclude models with reference items that contain DIF and only focus on the models in which either the reference item does not display DIF or no reference item is required in the first place.

## Matching Variable Effect

There was a rather complex impact of the type of matching variable on DIF detection in Kim's original model and Pan's reparameterization of it as shown in Table 4, which we compared also to basic Rasch model formulations.

Table 4 Power and Type-I Error as a Function of DIF Effect Size, Proficiency Distribution Difference, and Matching Variable Type

| DIF Effect Size | Matching Variable | Model | Proficiency Distribution of Focal Group (Reference Group Distribution is $N(0, 1)$) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $N(-0.5, 1)$ | | $N(0, 1)$ | | $N(0.5, 1)$ | |
| | | | Power | Type-I Error | Power | Type-I Error | Power | Type-I Error |
| .50 | None | Kamata | .53 | .05 | .60 | .05 | .65 | .04 |
| | | Pan | 1.00 | .70 | .68 | .04 | .05 | .64 |
| | | Rasch | .69 | .16 | .67 | .17 | .78 | .18 |
| | BTS | Kim | .06 | .50 | .38 | .07 | .88 | .19 |
| | | Pan | .48 | .57 | .47 | .57 | .43 | .52 |
| | PV | Kim | .50 | .09 | .63 | .10 | .60 | .12 |
| | | Pan | .98 | .66 | .68 | .07 | .08 | .62 |
| .75 | None | Kamata | .85 | .05 | .84 | .03 | .87 | .03 |
| | | Pan | 1.00 | .66 | .91 | .07 | .28 | .67 |
| | | Rasch | .91 | .25 | .91 | .26 | .94 | .28 |
| | BTS | Kim | .13 | .58 | .56 | .15 | .94 | .13 |
| | | Pan | .76 | .83 | .77 | .80 | .77 | .85 |
| | PV | Kim | .81 | .17 | .86 | .17 | .87 | .18 |
| | | Pan | 1.00 | .59 | .90 | .07 | .28 | .67 |

*Note.* Power and type-I error are computed over all items displaying or not displaying DIF, respectively, excluding the reference item, as well as all replications for plausible values. BTV = Booklet total scores, PV = plausible values.

Overall, Kamata's original Rasch model specification and the Rasch model specification in Conquest 2.0 performed comparatively better than Kim's model or Pan's reparameterization of it. They had consistent power largely independent of proficiency distribution differences and power differences were consistent with the magnitude differences of the DIF effect sizes; however, the Conquest 2.0 model estimation for the Rasch model showed some noticeable type-I error rate inflation. Pan's reparameterization of Kamata's model performed relatively more poorly among the

set of Rasch models whenever the proficiency distribution of the focal group did not match the proficiency distribution of the reference group.

Looking across the models that use matching variables specifically, Kim's model performed best when plausible values were used with moderate to high power and relatively mildly inflated type-I error rates across all proficiency distribution and DIF effect size conditions. In contrast, Pan's reparameterization of Kim's model with plausible values only worked well when the proficiency distribution of the focal group matched that of the reference group but performed poorly otherwise. The use of booklet total scores cannot be recommended for Pan's reparameterization of Kim's model at all and did not fare much better for Kim's original model even though, amongst all conditions for it, the best condition was the one where the proficiency distributions of the reference and focal group matched.

## General Model Comparison

To summarize the patterns in Tables 3 and 4 above, the type-I error rates and power for the models we investigated are plotted in Figure1 and Figure 2, respectively.

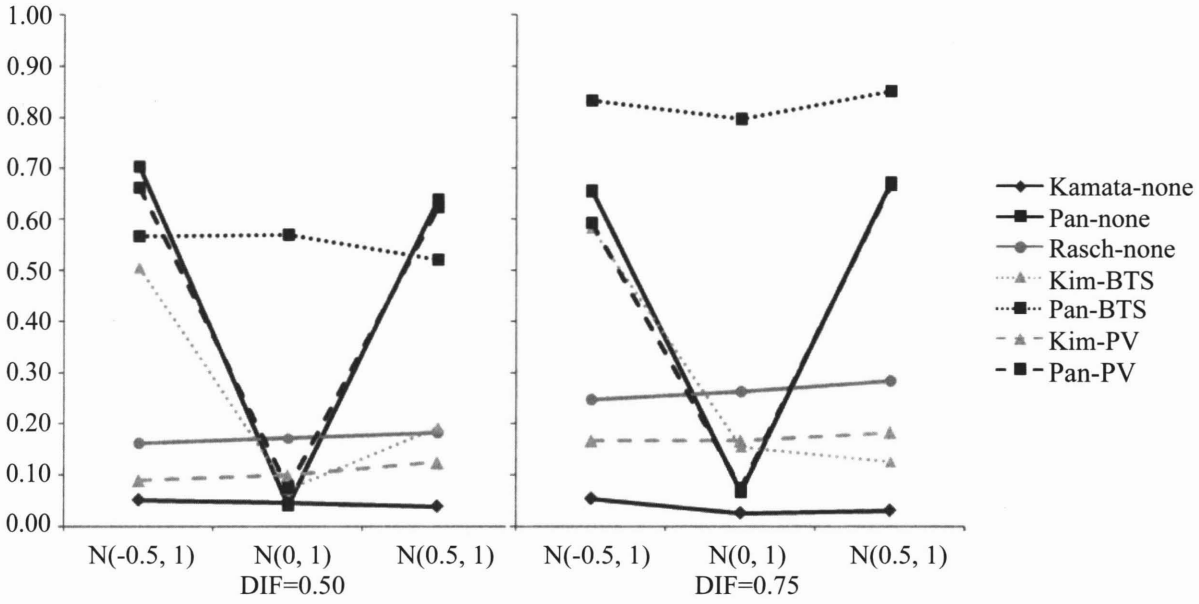



*Figure 1.* Type-I error rates across different models as a function of DIF effect size and distributional differences

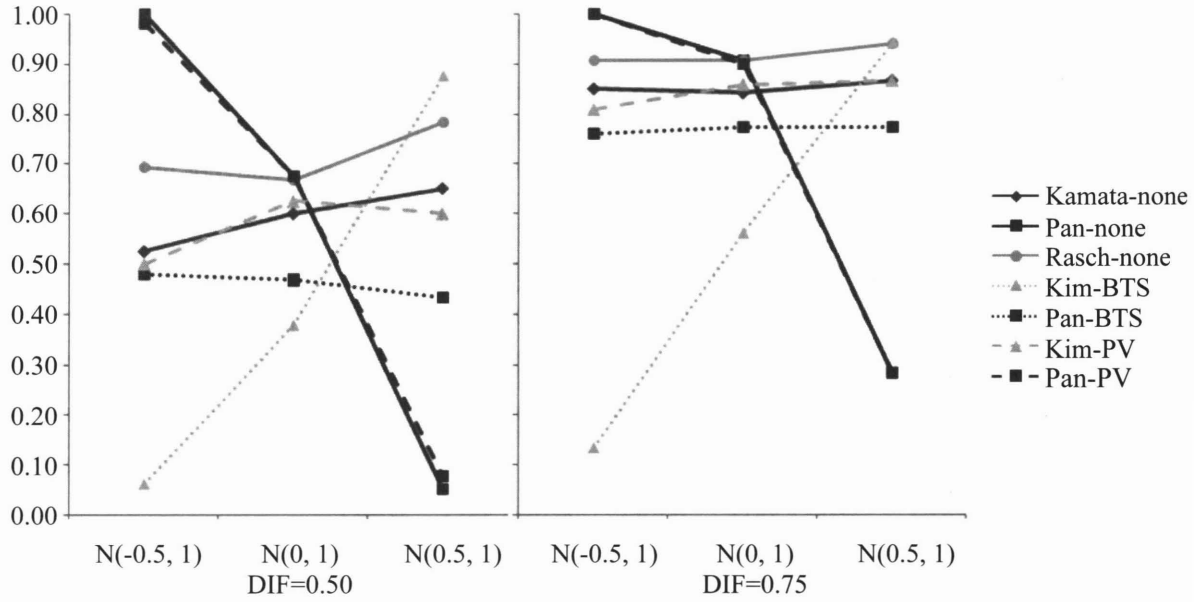



*Figure 2.* Power across different models as a function of DIF effect size and distributional differences computed using theoretical cut-offs

Overall, only in Kamata's model did the empirical sampling distributions match the theoretically expected ones closely, followed by Kim's original model formulation with plausible values as matching variables and the Rasch model as implemented in Conquest 2.0, both of which appear to require slight adjustments to the cut-off values.

More specifically, it can be seen that the test statistics followed the theoretical sampling distributions best when the proficiency distributions of the focal and reference group matched. The one exception was Pan's reparameterized version of Kim's model with booklet total scores, which showed uniformly poor performance. Pan's modifications of models also showed strikingly large increases in type-I error rates overall whenever the proficiency distributions for the focal and reference group were different independent of the magnitude of the DIF effect size.

With respect to power, one has to be careful with the interpretations of the patterns in Figure 2 as the inflated type-I error rates in Figure 1 indicate that the cut-off values for decision-making should be adjusted upward and, thus, effectively result in decreased power rates. As one would expect, then, the Rasch model as estimated via maximum likelihood in Coquest 2.0, Kamata's HGLM version of it as estimated with pseudo maximum likelihood, and Kim's original hybrid model with plausible values as matching variables appear to perform similarly well considering some slight downward adjustments for the latter two models given the observed inflated type-I error rates.

As stated before, Pan's versions of the models without reference items or Kim's model with booklet total scores are only attractive alternatives when proficiency distribution differences between focal and reference group can be assumed to be nonexistent. Given the highly inflated type-I error rates, we are hesitant to make other global statements about the remaining models across conditions.

# Real-Data Analyses Using TIMSS 2007 Data

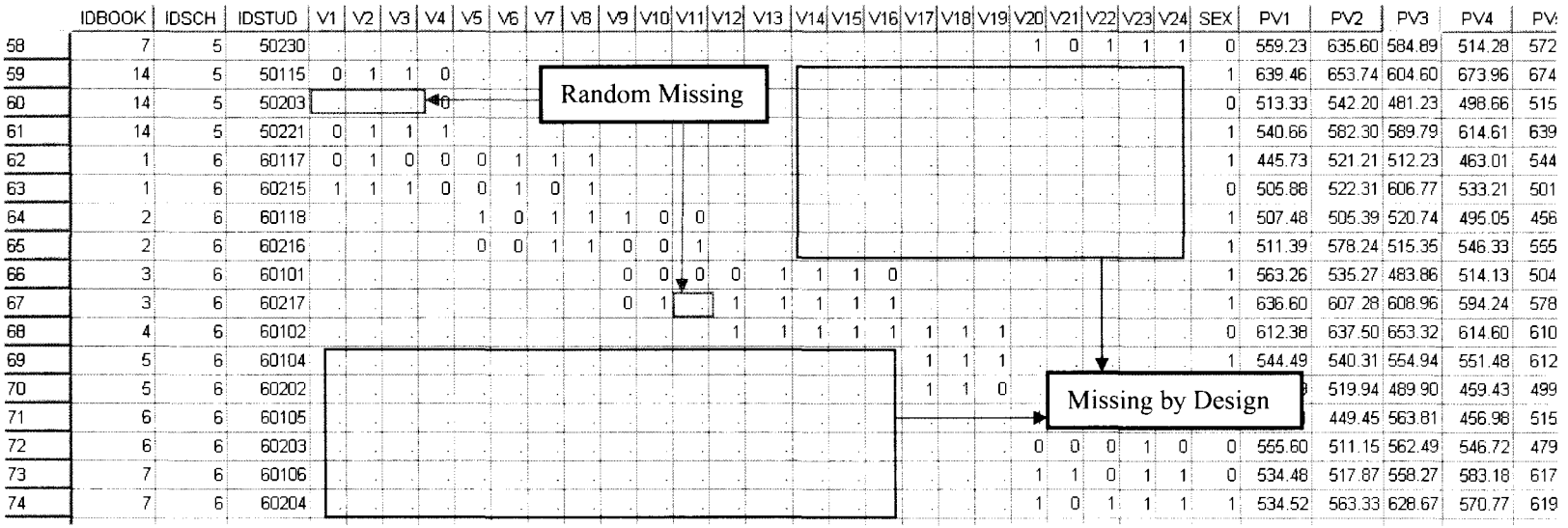
## The Data

The example items for this analysis come from the publicly released TIMSS 2007 science items for 4th grade (TIMSS & PIRLS International Study Center, 2009). To keep the illustration conceptually simple, we use a unidimensional subtest with items from the life science content domain only and use all 24 dichotomously scored life science items that can be found in eight out of the 14 booklets in the database. The subsample from the overall U.S. sample that we used consisted of 4,483 students who provided responses to at least one of these eight booklets.

A section of the raw data file is shown in Figure 3 to illustrate the data structure in TIMSS; in particular, the distinction between randomly missing data and data missing by design is apparent (see also Enders, 2010, for more detailed discussions of missing data analysis in general).

In TIMSS, as in our simulation study, two booklets usually share several common items. For example, booklet 1 and booklet 2 (the first column is the booklet ID) contain eight and seven items, respectively, and share four common anchor items. Consequently, anchor items have more responses from the selected 4,483 students than booklet-specific items and yield more precise parameter estimates.

The TIMSS 2007 international database also provides various student background variables, including sex, age and testing language of the students. For this paper our focus is on the potential gender bias on the life science items to mimic a two-level grouping variable from the simulation study. Our use of sex as a grouping variable, albeit chosen primarily for its two-level nature, was substantively motivated by research on DIF by Hamilton (1999) and Hamilton and Snow (1998) who found that for those items requiring spatial reasoning or visual content, the male examinees performed better than the female examinees after matching on total test scores. Zenisky, Hambleton, and Robin (2003b) further suggested that gender DIF in science could result from three sources: item content category, reference component, and item type. However, we acknowledge that our intent here is not substantive interpretation but, rather, illustration of consistency of results across a variety




| | IDBOOK | IDSCH | IDSTUD | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 | V14 | V15 | V16 | V17 | V18 | V19 | V20 | V21 | V22 | V23 | V24 | SEX | PV1 | PV2 | PV3 | PV4 | PV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 58 | 7 | 5 | 50230 | | | | | | | | | | | | | | | | | | | | 1 | 0 | 1 | 1 | 1 | 0 | 559.23 | 635.60 | 584.89 | 514.28 | 572 |
| 59 | 14 | 5 | 50115 | 0 | 1 | 1 | 0 | | | | | | | | | | | | | | | | | | | | | 1 | 639.46 | 653.74 | 604.60 | 673.96 | 674 |
| 60 | 14 | 5 | 50203 | | | | 0 | | | | | | | | | | | | | | | | | | | | | 0 | 513.33 | 542.20 | 481.23 | 498.66 | 515 |
| 61 | 14 | 5 | 50221 | 0 | 1 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | 1 | 540.66 | 582.30 | 589.79 | 614.61 | 639 |
| 62 | 1 | 6 | 60117 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | | | | | | | | | | | | | | | | | 1 | 445.73 | 521.21 | 512.23 | 463.01 | 544 |
| 63 | 1 | 6 | 60215 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | | | | | | | | | | | | | | | | | 0 | 505.88 | 522.31 | 606.77 | 533.21 | 501 |
| 64 | 2 | 6 | 60118 | | | | | 1 | 0 | 1 | 1 | 1 | 0 | 0 | | | | | | | | | | | | | | 1 | 507.48 | 505.39 | 520.74 | 495.05 | 456 |
| 65 | 2 | 6 | 60216 | | | | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | | | | | | | | | | | | | | 1 | 511.39 | 578.24 | 515.35 | 546.33 | 555 |
| 66 | 3 | 6 | 60101 | | | | | | | | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | | | | | | | | | 1 | 563.26 | 535.27 | 483.86 | 514.13 | 504 |
| 67 | 3 | 6 | 60217 | | | | | | | | | 0 | 1 | | 1 | 1 | 1 | 1 | 1 | | | | | | | | | 1 | 636.60 | 607.28 | 608.96 | 594.24 | 578 |
| 68 | 4 | 6 | 60102 | | | | | | | | | | | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | | | | | 0 | 612.38 | 637.50 | 653.32 | 614.60 | 610 |
| 69 | 5 | 6 | 60104 | | | | | | | | | | | | | | | | | 1 | 1 | 1 | | | | | | 1 | 544.49 | 540.31 | 554.94 | 551.48 | 612 |
| 70 | 5 | 6 | 60202 | | | | | | | | | | | | | | | | | 1 | 1 | 0 | | | | | | | | 519.94 | 489.90 | 459.43 | 499 |
| 71 | 6 | 6 | 60105 | | | | | | | | | | | | | | | | | | | | | | | | | | | 449.45 | 563.81 | 456.98 | 515 |
| 72 | 6 | 6 | 60203 | | | | | | | | | | | | | | | | | | | | 0 | 0 | 0 | 1 | 0 | 0 | 555.60 | 511.15 | 562.49 | 546.72 | 479 |
| 73 | 7 | 6 | 60106 | | | | | | | | | | | | | | | | | | | | 1 | 1 | 0 | 1 | 1 | 0 | 534.48 | 517.87 | 558.27 | 583.18 | 617 |
| 74 | 7 | 6 | 60204 | | | | | | | | | | | | | | | | | | | | 1 | 0 | 1 | 1 | 1 | 1 | 534.52 | 563.33 | 628.67 | 570.77 | 619 |

*Figure 3.* Section of TIMSS data set with values missing at random and by design

of statistical DIF detection methods.

## Data Preparation

In the following we show the data preparation for DIF analyses in HLM 6.0 (Raudenbush et al., 2004) where each model level requires one separate data file; if SAS PROC GLIMMIX were used then only one file would be needed.

In the original data file the responses and background information for one individual are arranged in one single row as shown in Figure 3. The ID variables for the booklet (IDBOOK), school (IDSCH) and student (IDSTUD) were used to identify each individual. The following 24 variables (V1 to V24) are the response to the 24 selected life science items. The grouping variable (SEX), five plausible values (PVs), as well as the booklet scores for one particular booklet (SUBT) are also included. As we discussed earlier, the use of a complex sampling design requires the use of appropriate sampling weights for secondary analyses such as the DIF analyses we conduct here. Consequently, the two appropriate sampling weight variables for the students and the schools, which are not shown in Figure 3, were also used.

For all the HGLM DIF analyses, dummy variables were specified in the model as item indicators. The 24 items require 23 dummy variables for Kamata's and Kim's original models. Instead of one row for one student as in the original data file, a matrix with 24 rows and 23 columns is specified for each student as shown in Figure 4; in this matrix the response for a particular item for a particular student occupies one row.

In particular, the first row represents the first student's response to the first item. Since the first item is used as reference item in the example data, it is noticeable that all the dummy variables are given '0' values for this item. The exact item scores for the 24 items are arranged in one column while the ID variables, weight variables, plausible values, booklet total score, and sex for a particular student are the same across the 24 rows for each student. The data files for Pan's modified models are very similar except that there is one dummy variable for each item.

As the simulation study showed, whether a reference item displays DIF or not has a strong impact on DIF detection for other items in Kamata's and Kim's original models; unfortunately, in the absence of a priori DIF information, for the empirical data it is hard to decide which item should be selected as reference item. The strategy we used is to select two items (Item 1and Item 24) as reference items for the TIMSS data in a two-step strategy described next.

| | UD | WSCH | WSTU | IDBOOK | SEX | SUBT | PV1 | PV2 | PV3 | PV4 | PV5 | Y | Z2 | Z3 | Z4 | Z5 | Z6 | Z7 | Z8 | Z9 | Z10 | Z11 | Z12 | Z13 | Z14 | Z15 | Z16 | Z17 | Z18 | Z19 | Z20 | Z21 | Z22 | Z23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 19 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 20 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 21 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 22 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 23 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 24 | 101 | 801.69 | 1.45 | 3 | 1 | 4 | 569.682 | 599.416 | 546.224 | 538.28 | 486.261 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 102 | 801.69 | 1.45 | 4 | 0 | 5 | 609.936 | 599.614 | 597.091 | 579.95 | 562.373 | . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | 102 | 801.69 | 1.45 | 4 | 0 | 5 | 609.936 | 599.614 | 597.091 | 579.95 | 562.373 | . | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Missing by Design

Response ( $y_{ij}$ )

Dummy Codes ( $z_{qij}$ )

*Figure 4.* Example of SAS data file for HGLM with Item 1 as reference item

## DIF Analysis Results

Recall that no reference item is required for the Rasch model and Pan's three reparameterized versions of Kamata's model and Kim's model – one with booklet total scores and one with plausible values. Since these models seem to perform better for DIF detection, the results of DIF analyses for these four models are compared first to provide the initial identifications of DIF items; results are summarized in Table 5.

Overall, eight out of 24 items are consistently identified as displaying DIF by at least three models; specifically, five items were declared to favor male students (Item 1, Item 3, Item 5, Item 11 and Item 22) and the other three were declared to favor female students (Item 2, Item 4 and Item 19); due to the consistency across models it is likely that these results are trustworthy. The Rasch model identified four more items as displaying DIF (i.e., Item 7, Item 8, Item 12 and Item 20), which may be due to a slightly inflated type-I error rate relative to the other models assuming the ability distributions are identical across the two groups. We note also that we would expect about 1-2 items to be declared as displaying DIF anyway under a nominal 5% type-I error rate.

The models that require reference items were compared next with Items 1 and 24 chosen as reference items for two separate runs; we selected these items based on their likely DIF characteristics. Specifically, according to the DIF analyses using models without reference items above, it is most likely that Item 1 displays DIF against males while Item 24 is most likely not displaying DIF against either sex group.

In short, the analyses with these two items as reference items echo the findings in the simulation study. Specifically, when the likely non-DIF item, Item 24, is used as a reference item, Kim's model with plausible values as well as Kamata's model agrees with the results from the previous analyses for the remaining items, especially for items where multiple models pointed to DIF. However, when the likely DIF item, Item 1, is used as a reference item, Kamata's model and Kim's model with plausible values agree but they flag many other items as displaying DIF, predominantly for females, which is largely contrary to the directionality of the DIF effects detected previously. Moreover, Kim's model with booklet total scores gives conflicting results with respect to these two models.

We conclude this section by noting that DIF detection under complex item sampling designs is, of course, unnecessarily more complex when booklet total scores are used to match students. To illustrate this point, when the reference item is not contained in a particular booklet, the last item in that particular booklet is identified by SAS as the reference item for that booklet by default. In our

Table 5 DIF Detection Results for TIMSS Items using Different Models

| Item | No Reference | | | | Reference Item 1 | | | Reference Item 24 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rasch | Pan-Kamata | Pan-Kim-PV | Pan-Kim-BTS | Kamata | Kim-PV | Kim-BTS | Kamata | Kim-PV | Kim-BTS |
| Item 1 | M | M | M | M[a] | -- | -- | -- | M | M | M |
| Item 2 | F | F | F | F | F | F | F | F | F | F |
| Item 3 | M | M | M | | | | | M | M | M[a] |
| Item 4 | F | F | F | F | F | F | F | F | F | F[a] |
| Item 5 | M | M | M | M[a] | | | | M | M | M[a] |
| Item 6 | | | | | | | | | | MF[b] |
| Item 7 | M | | | | | | | | | MF[b] |
| Item 8 | F | | | F[a] | F | F | F | F | F | F |
| Item 9 | | | | F[a] | F | F | MF[b] | | | MF[b] |
| Item 10 | | | | | F | F | MF[b] | | | MF[b] |
| Item 11 | M | M | M | | | | | M | M | M |
| Item 12 | M | | | | | | | | | M |
| Item 13 | | | | | F | F | M[a] | | | M[a] |
| Item 14 | | | | | F | F | M[a] | | | M[a] |
| Item 15 | | | | | F | F | | | | |
| Item 16 | | | | MF[b] | F | | M | | | M |
| Item 17 | | | | | | | M | | | M |
| Item 18 | | | | | F | F | M | | | M |
| Item 19 | F | F | F | F[a] | F | F | -- | F | F | -- |
| Item 20 | F | | | | F | F | | | F | |
| Item 21 | | | | F | F | F | F | F | F | F |
| Item 22 | M | M | M | M | | | M | M | M | M |
| Item 23 | | | | | | | | | | |
| Item 24 | | | | | F | F | -- | -- | -- | -- |

[a] one of the two gender tests is significant; [b] both gender tests are significant but the results favor different gender groups. F: the item favors the female students; M: the item favors the male students. Pan-Kamata, Pan-Kamate-PV, and Pan-Kim-BTS are Pan's reparameterized versions of Kamta's model, Kim's model with plausible scores as matching variables, and Kim's model with booklet total scores as matching variables, respectively.

analyses, seven out of the 18 booklets that we used required five or six other reference items in addition to Item 1, which makes strict comparisons difficult.

# Discussion

In this paper we reviewed key ideas for DIF detection generally and their practical implementation for educational surveys using the HGLM framework specifically. The complex sampling design for students coupled with the complex booklet design for items have a significant impact on the ways DIF parameters and their standard errors need to be estimated. By utilizing the HGLM framework, as one member of the larger family of GLLAMMs, explanatory DIF analyses can be conducted accurately and efficiently.

Six general HGLMs within the GLLAMM framework were introduced and a small-scale simulation study was conducted to compare their DIF detection ability under different conditions. Given the technical nuisance and undesirability of obtaining total scores on a booklet-by-booklet basis under a complex booklet design, IRT models are generally preferable. Indeed, the simulation study indicated that the Rasch models perform best across a wide range of conditions with some exceptions that align with theoretical expectations.

Specifically, the DIF detection ability of Kamata's Rasch model formulation is dependent upon whether DIF is present in the reference item. One possible solution is to randomly select several items as reference items and compare results for the other items across all analyses in order to identify potential DIF items. Since both the results of the simulation study and the empirical data indicate that the Rasch model as estimated in Conquest 2.0 may also lead to a few additional false-positives, we suggest combining the results from both model formulations – or multiple comparable model formulations for the same model generally – to arrive at a joint judgment about DIF.

Even though removing the need to identify a reference item through the application of Pan's reparameterized models is conceptually appealing, all of Pan's modified models were sensitive to whether the proficiency distributions between the two groups of interests matched or not. The unacceptably high type-I error rates when the distributions did not match in the simulation study reflected the seeming inability of the modified models to differentiate between impact and DIF. Further study is required to explore this issue comprehensively.

In real-data analyses where true proficiency score distributions are not known the use of models that are not affected by potential distributional differences is key and, if a range of models with

different properties is used, a synthesis of evidence is required. For example, for our data set there were several items for which the same items were flagged for the same sex group – even when Pan's models were used – which is strong evidence of likely DIF for these items.

Interestingly, the use of plausible values in Kim's model – albeit not Pan's modification of it for differing proficiency distributions – did not appear to deteriorate the DIF detection markedly compared to Kamata's Rasch model formulation or the Rasch model formulation as estimated in Conquest 2.0. That is, if a hybrid IRT and logistic regression modeling approach within a HGLM framework is desirable, using plausible values in complex booklet design scenarios seems to be possible if Kim's original model formulation with a non-DIF reference item is used.

Of course, if individual response data are available, which is necessary for DIF analyses, there is no technical reason – and no real statistical advantage that we can think of – to using a hybrid IRT and logistic regression approach as in Kim's model over a direct IRT approach. This would seem to be true for Rasch models, which allow only for the detection of uniform DIF by definition, as well as two-parameter models, which allow for the detection of both uniform and non-uniform DIF. The specification and the resulting DIF detection of two-parameter models can still be somewhat cumbersome in some general-purpose latent-variable estimation programs, but we expect that any perceived specification and estimation advantages of Kim's model will vanish in the near future. Nevertheless, our simulation study provides some evidence that plausible values could be used if desired, especially if the results are properly aggregated across the multiple analyses that they require.

# References

Adams, R. J., & Gonzalez, E. J. (1996). The TIMSS test design. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study technical report, volume I: Design and development*. Chestnut Hill, MA: Boston College.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*(3), 279-292.

Binici, S. (2008). *Random-effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods.* Unpublished doctoral dissertation, Florida State University, Miami, FL.

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

Camilli, G., & Shepard, L. (1994). *MMSS volume 4: Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269-279.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

de Ayala, R. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

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

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly,*

*4*(2), 1-36.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Goldstein, H. (2003). *Multilevel statistical models*. London: Arnold.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(3), 211-235.

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests* (CSE Tech. Rep. No. 483). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing, 8*(3), 237-250.

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

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

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review, 51*(2), 175-188.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kamata, A., & Binici, S. (2003, July). *Random-effect DIF analysis via hierarchical generalized linear model*. Paper presented at the International Meeting of the Psychometric Society (IMPS), Sardinia, Italy.

Kim, W. (2003). *Development of a differential item functioning procedure using the hierarchical generalized linear model: A comparison study with logistic regression procedure*. Unpublished doctoral dissertation, Pennsylvania State University, Pennnsylvania, PA.

Lomax, R. G. (2007). *Statistical concepts: A second course* (3rd ed.). Mahwah, NJ: Erlbaum.

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

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples.

*Psychometrika, 56*(2), 177-196.

Mislevy, R. J., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131-154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84* (Technical Report No. 15-TR-20) (pp. 293-360). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus*. Los Angeles: Author.

Osterlind, S. (2009). *Differential item functioning* (2nd ed.) (Quantitative applications in the social sciences series). Thousand Oaks, CA: Sage.

Pan, T. (2008). *Using the multivariate multilevel logistic regression model to detect DIF: A comparison with HGLM and logistic regression DIF detection methods.* Unpublished doctoral dissertation, Michigan State University, Ann Arbor, MI.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, 60* (Series B), 23-40.

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement, 44*(1), 69-87.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Rubin, D. B. (1987). M*ultiple imputation for nonresponse in sample surveys*. New York: John Wiley.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). Secondary analyses of large-scale assessment data. *Educational Researcher, 39*(2), 142-151.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Skondral, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

TIMSS & PIRLS International Study Center. (2009). *TIMSS 2007 international database and user guide.* Retrieved June 12, 2010, from http://timss.bc.edu/TIMSS2007/idb_ug.html

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2010). *What are plausible values and why are they useful?* Retrieved June 30, 2010, from http://www.ierinstitute.org/IERI_Monograph_Volume-_02_Chapter_-01.pdf

von Davier, M, Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 1039-1055). North Holland: Elsevier.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). Acer Conquest 2.0: Generalised item response modeling software [Software program]. Camberwell, Victoria: Acer Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). *Item response theory.* In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CN: Greenwood.

Zenisky, A., Hambleton, R., & Robin, F. (2003a). Detection of differential item functioning in large scale state tests: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*(1), 51-64.

Zenisky, A., Hambleton, R., & Robin, F. (2003b). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1&2), 61-78.

Zhang, Y., Dorans, N., & Matthews-Lopez, J. (2005). *Using DIF dissection method to assess effects of item deletion* (ETS Research Rep. No. RR-05-23). Princeton, NJ: ETS.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

教育科學研究期刊 第五十六卷第一期
2011 年，56（1），91-127

# 大型教育調查研究中的差別試題功能：次級分析中的核心概念及建模方法

朱小姝
馬里蘭大學
測量統計評估系
研究生

安德魯・儒普
馬里蘭大學
測量統計評估系
副教授

高靜
馬里蘭大學
大學學院
測量評估辦公室
心理測評師

## 摘要

大型教育評量研究常採用多階段抽樣的設計（multi-stage sampling design），透過對母群體之抽樣單位進行分層以抽取受測者。此外，還會採用複雜題本設計（complex booklet design）的方式將題目組成多份測驗題本。在此情況下，欲確保公正測量出不同受測群體的能力，關鍵在於能夠有效偵測所採用的題目是否具差別試題功能（differential item functioning, DIF）。本文旨在介紹探討在大型教育評量複雜設計之下能用以偵測差別試題功能的建模方法，並應用六種可用於偵測 DIF 的多階層廣義線性模式（hierarchical generalized linear models, HGLMs），再透過電腦模擬比較它們偵測 DIF 的效力。接著又將這些模式應用到國際數學與科學教育成就趨勢調查研究（TIMSS）的實證數據上，藉以探測是否存在一致性的性別 DIF（uniform gender DIF）。

關鍵字：複雜題本設計、差別試題功能、多階層廣義線性模式、多階段抽樣設計

通訊作者：朱小姝，E-mail: xzhu16@umd.edu
收稿日期：2010/06/14；修正日期：2010/11/28；接受日期：2010/12/03。