

Random-Effect Differential Item Functioning Across Group Unites by the Hierarchical
Generalized Linear Model

Akihito Kamata

Saengla Chaimongkol

Evrin Genc

Kuzey Bilir

Florida State University

April 2005

Correspondance:

Akihito Kamata, Ph.D.

Department of Educational Psychology & Learning Systems

Florida State University

Tallahassee, FL 32306-4453

kamata@coe.fsu.edu

This paper was presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April 2005.

This research was supported by the NAEP secondary data analysis grant program of the U.S. Department of Education (R902B030025).

Abstract

This study introduces a random-effect differential item functioning (DIF) model that allows one to estimate the variation of DIF across group units. A three-level hierarchical generalized linear model is adopted. A series of data analysis are demonstrated on DIF between the limited English proficiency (LEP) sample who received test accommodations and students who did not receive any test accommodations in the 2003 NAEP 4th grade mathematics assessment data. DIF between accommodated and non-accommodated students are detected, and the variations of the magnitude of DIF across schools are estimated. Then, some school characteristic variables are examined whether they can explain the variation of DIF between schools.

Introduction

There has been a considerable trend in the use of testing accommodations with the aim of including more students with disabilities or limited English skills in statewide and district-wide academic testing. Through the expansion of educational reforms, several initiatives have been accomplished in an effort to provide similar testing experiences and opportunities to those individuals with disabilities as those without disabilities. The incorporation and modifications of the Individuals with Disabilities Education Act (IDEA) and the Americans with Disabilities Act (ADA) provide a strong incentive and motivation to include individuals with disabilities in statewide and district-wide assessment measures, thereby helping these individuals achieve new standards in academic situations (Gordon, Lewandowski, Murphy, & Dempsey, 2002). As a consequence of the reform movement, in 1996, the National Assessment of Educational Progress (NAEP) started the incorporation of these accommodations, such as extended time, large print, transcription, oral reading, and signing of directions, for students whose schools felt it necessary to provide such accommodations during testing situations (Rogers, Kokolis, Stoeckel, & Kline, 2002).

Testing accommodations refer to those adaptations, modifications, and test alternatives in standardized and non-standardized tests that create comparable testing results between individuals with and without disabilities (Schulte, Elliott, & Kratochwill, 2001). These accommodations are assumed to help an individual with a disability in a way that does not alter the meaning and content of the test. In other words, accommodations in testing are not intended to help the individual's score increase, but to provide a similar testing situation for those individuals with disabilities to those without disabilities. However, the advantages and disadvantages of using of such testing accommodations have been widely discussed in academic circles. Therefore, it is important to provide a report examining the effects of accommodations in terms of aiding students during testing situations.

In order to study the effects of accommodations on the test items, differential item functioning (DIF) analyses is of interest due to its indication of a possible bias between accommodated and non-accommodated students. Technically, DIF is present for a test item when respondents from two subpopulations with the same trait level have different probability of answering the item correctly. A consequence of having a DIF item is that the same true trait levels for examinees from different subpopulations could indicate different total test scores or

trait level estimates. Currently, many statistical techniques have been proposed, based upon various theoretical backgrounds and practical purposes. They include contingency table, such as the Mantel-Haenszel procedure as modified by Holland and Thayer (1988), parametric-based Item Response Theory (IRT) methods, such as area measures (Raju, 1988), and likelihood-ratio tests (Thissen, Steinberg, & Wainer, 1988), and non-parametric multidimensional-based IRT approach (Shealy & Stout, 1993).

Once an item is identified as functioning differently from one subpopulation to another, understanding why the item is functioning differently between groups may be useful for many audiences. As one attempt, Gierl et al. (2003) studied gender DIF in mathematics by combining substantive and statistical analyses, as a two-stage process. Three different statistical methods: SIBTEST, DIMTEST, and multiple linear regression, were used to test hypotheses about gender differences and to test whether content and cognitive differences were among items. Bolt (2000), for another example, found that multiple-choice items had more DIF characteristic than constructive-response items between males and females on SAT math pretest items. Also, Walker and Beretvas (2001) found that DIF between proficient writers and non-proficient writers were prominent only for constructed-response items that required writing about their solution. These results can possibly provide suggestions that may be informative to minimize DIF items in future by many different means, including instruction, policy and test construction. These studies were based on multidimensional IRT based approaches.

As another statistical approach, Swanson et al. (2002) proposed a two-level logistic regression model to evaluate sources of DIF. This approach explicitly accounts for the nested structure of the data and combines results of logistic regression analyses across individual items to investigate the variation of DIF. Their level-1 model is a logistic regression model for DIF detection proposed by Swaminathan and Rogers (1990). In the level-2 models, the coefficients from level-1 model are treated as random variables and allow one to incorporate item characteristic variables to the models in order to explain the variation of DIF across items.

Meulders and Xie (2004) also proposed a DIF model by parameterizing an interaction between group indicator and item properties. Alternatively, their parameterization was by item facets instead of interaction between group indicator and item indicator, consequently referred as the differential facet functioning (DFF). DFF can be used to identify and explain that the effect of item properties on item difficulty depends on the group. They also proposed a random-weight

DIF (RW-DIF) and a random-weight DFF (RW-DFF) by making some of DIF and DFF parameters random across persons to find evidence of individual difference on the DIF/DFF within groups. Each random DIF/DFF can be considered as an extra person dimension added to the model that may be assumed to correlate with the person ability distribution for each group.

There is also a possibility that the magnitude of the DIF varies across group units, such as schools, and communities. Kamata and Binici (2003) extended a two-level DIF model (Kamata, 1998) to three-level DIF model using the hierarchical generalized linear model (HGLM) framework. Their three-level model approach can be used to model variation of DIF across schools as well as applied to identify the school characteristic variables that explain such variation. Chaimongkol (2005) demonstrated different parameterizations from Kamata and Binici by implementing a fully Bayesian approach. Cheong (in press) also demonstrated the use of HGLM to investigate the effect of school contexts on DIF.

This research demonstrates a random-effect DIF analysis that allows one to incorporate group characteristic variables in the framework of hierarchical generalized linear model (HGLM). More specifically, this study will demonstrate (a) a model for detecting DIF for dichotomously scored items that takes into account the three nested level structure of data, (b) a model in such a way that DIF of a particular item may vary among level-3 units, and (c) a model to identify level-3 unit characteristic variables that explain such DIF variation. The model will be applied to the National Assessment of Educational Progress (NAEP) 2003 data in order to study the effects of testing accommodations for LEP students.

Methods

NAEP Data

In this study, dichotomous test items in one item block from the 2003 NAEP 4th grade mathematics assessment were analyzed. In this item block, there were 21 dichotomously scored test items, including 16 multiple choice items and 5 short constructed response items. The characteristics of the 21 items are summarized in Table 1. Under the standard NAEP design, each student is randomly assigned one of the different booklets containing different combinations of blocks of items. This particular item block was chosen for this study primarily for the purpose of maximizing the available sample size of LEP students.

Table 1. Dichotomously scored items in the studied item block

Item	Sub Domain	Type	Description
1	numbers and operations	MC	Add whole numbers
2	numbers and operations	MC	The 3rd picture shows 3/4 shaded
3	numbers and operations	MC	Identify solution procedure
4	numbers and operations	MC	Solve story problem (division)
5	numbers and operations	MC	Solve multi-step story problem
6	numbers and operations	MC	Solve multi-step story problem
7	numbers and operations	SCR	Use a number line graph
8	numbers and operations	MC	Solve story problem (fractions)
9	numbers and operations	MC	Read a scale diagram
10	measurement	MC	Compare weights
11	measurement	MC	Apply concept of perimeter
12	measurement	SCR	Read a ruler
13	geometry	MC	Apply transformational geometry
14	geometry	MC	Apply properties of rectangles
15	geometry	MC	Apply properties of a cube
16	geometry	SCR	Draw an obtuse angle
17	data analysis	MC	Interpret pie chart data
18	data analysis	SCR	Complete a bar graph
19	algebra	SCR	Complete a letter pattern
20	algebra	MC	Apply concept of equality
21	algebra	MC	Solve an inequality

Note. MC = multiple-choice items, and SCR = short constructed-response items.

According to the NAEP sampling design in 2003 assessments, the schools in the national sample included students with and without disability (SD) and/or limited English proficiency (LEP). A variety of assessment accommodations and adaptations were offered to students in the SD/LEP sample. In this study, DIF analysis was conducted between LEP students who received at least one type of test accommodation and students who did not receive any test accommodations.

First, students who took a particular test booklet were chosen, because a large sample of LEP students was available for this booklet. This booklet was consisted of two item blocks. For this analysis, one particular item block was chosen because there are more items in this item block than the other one in the test booklet. Then, students who took this item block other than in the selected test booklet were also included in the sample, primarily for the purpose of increasing the sample size. (There were 9 other test booklets that contained the target item block.) However, only students from the schools that were already identified in the target item block in the selected test booklet were included, in order to avoid increasing the number of schools with very few students. As a result, a total of 2,243 students from 236 schools were included in the sample.

Among the 2,243 students, 1,243 students were in the standard administration group with no test accommodation and 1,200 students were in the accommodated group. It was also ensured that each school have at least 2 students in each of the standard administration and accommodated groups.

Analyses

As an initial step, the Mantel-Haenszel (M-H) procedure was used to detect items that show DIF. Results were confirmed by the DIF detection HGLM model. Then, the random DIF model to detect the variation of DIF between schools was fitted. Finally, the model to examine the relationship between a school characteristic variable and the DIF magnitude was fitted to explain a source variation of DIF. All of the parameters are estimated by penalized quasi-likelihood (PQL) method available in the HLM 6 software (Raudenbush, Bryk, Cheong, & Congdon, 2004).

DIF detection. Initial DIF detections were conducted by the Mantel-Haenszel (M-H) procedure (Holland & Thayer, 1988). The Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) is a general statistical approach to test for the dependency of two variables in a three-way contingency table. Holland and Thayer (1988) proposed the use of the MH procedure to detect DIF between two sub-samples of examinees for a dichotomously scored test item, by summarizing test data in a form of a (2 scoring categories) \times (2 sub-populations) \times (k categories in matching criterion) contingency table.

The MH method of DIF analysis involves the construction of a contingency table, which gives the counts of correct (1) and incorrect (0) responses. These counts are broken up by the group indicator (focal and reference groups) and the matching criterion (k categories). It computes a common odds ration for item i ($\hat{\alpha}_{MH_i}$) by

$$\hat{\alpha}_{MH_i} = \frac{\sum_j^k a_j d_j / N_j}{\sum_j^k b_j c_j / N_j},$$

where j is the j th category in the matching criterion ($j = 0, \dots, k$), a_j , and c_j are the frequencies of correct answers by focal group and reference group in the j th category, respectively, b_j , and d_j are the frequencies of correct answers by focal group and reference group in the j th category, respectively, and N_j is the number of examinees in the j th category. If there is no difference

between the reference and focal groups by controlling for the level of matching criterion, then $\hat{\alpha}_{MH_i}$ will be equal to 1. If the reference group performs better on the item, then $\hat{\alpha}_{MH_i}$ will be smaller than 1, an indication of possible bias against the focal group. On the other hand, if the common odds-ratio is greater than 1, it is an indication of possible bias against the reference group. We evaluated the 95% confidence interval of $\hat{\alpha}_{MH_i}$ whether it is significantly different from 1.0.

DIF detection and Random DIF model. DIF detection and Random DIF model were formulated by extending the hierarchical Rasch model framework demonstrated by Kamata (2001). Let $Y_{ips} = 1$ if the i th response is correct for student p of school s and $Y_{ips} = 0$ otherwise, and μ_{ips} be the probability of $Y_{ips} = 1$. This probability varies randomly over students. However, conditioning on this probability, we have $Y_{ips} | \mu_{ips} \sim \text{Bernoulli}$, $E(Y_{ips} | \mu_{ips}) = \mu_{ips}$, and $\text{Var}(Y_{ips} | \mu_{ips}) = \mu_{ips}(1 - \mu_{ips})$. Therefore, the level-1 model is

$$\log\left(\frac{\mu_{ips}}{1 - \mu_{ips}}\right) = \eta_{ips} = b_{0ps} + \sum_{i=1}^{I-1} b_{ips} D_{ips},$$

where D_{ips} is the item dummy variable for item i ($i = 1, \dots, I$), student p ($p = 1, \dots, P$), and school s ($s = 1, \dots, S$). b_{ips} is the item effect for the i th item, and b_{0ps} is the intercept for the model, which indicates the effect of the reference item. The level-2 is the person level model and specified as

$$b_{0ps} = \gamma_{00s} + \gamma_{01s} W_{ps} + u_{0ps},$$

and

$$\begin{cases} b_{ips} = \gamma_{i0s} & \text{if no DIF,} \\ b_{ips} = \gamma_{i0s} + \gamma_{i1s} W_{ps} & \text{otherwise.} \end{cases}$$

Here, W_{ps} is the DIF factor that was coded as 1 if the student was accommodated LEP student, and coded 0 otherwise. It was assumed that $u_{0s} \sim N(0, \sigma_{u_s}^2)$, but mean around each e_{00s} , where e_{00s} is defined below. γ_{01s} is the mean ability difference between the focal and reference groups, while γ_{i1s} is the DIF magnitude for the i th item. If all effects, including DIF parameters are fixed in the level-3 model, except the intercept γ_{00s} , the level-3 model are written as

$$\begin{aligned}\gamma_{00s} &= \pi_{000} + e_{00s} \\ \gamma_{01s} &= \pi_{010} \\ \gamma_{i0s} &= \pi_{i00} \\ \gamma_{i1s} &= \pi_{i10} .\end{aligned}$$

This is considered as the DIF detection model, because the model parameterizes DIF as fixed effects (π_{i10} for item i) and can be used for detecting DIF items. On the other hand, the level-3 model can be specified as

$$\begin{aligned}\gamma_{00s} &= \pi_{000} + e_{00s} \\ \gamma_{01s} &= \pi_{010} \\ \gamma_{i0s} &= \pi_{i00} \\ \gamma_{i1s} &= \pi_{i10} + e_{i1s} ,\end{aligned}$$

where

$$\begin{bmatrix} e_{00} \\ e_{i1} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e_{00}}^2 & \sigma_{e_{00} e_{i1}} \\ \sigma_{e_{00} e_{i1}} & \sigma_{e_{i1}}^2 \end{bmatrix} \right).$$

This is a specification for a random-effect DIF model. Here, our interest is to estimate $\sigma_{e_{i1}}^2$, which is the variance of DIF magnitude across schools.

Exploratory DIF model. The last equation in the level-3 model is further expanded to

$$\gamma_{i1s} = \pi_{i10} + \pi_{i11} X_{1s} + e_{i1s} ,$$

where X_{1s} is the first school characteristic variable for the s th school. In this model, our interest is the magnitude of π_{i11} , which indicates the effect of the school characteristic variable on the DIF magnitude. The effect is a three-way interaction of (item difficulty) \times (group) \times (school characteristic).

Results

The results of M-H analyses for the 21 items, based on the total number of correct answers for the 21 items, are summarized in Table 1. A total of 8 items were identified as displaying DIF characteristic, where the common-odds ratio was significantly different from 1.0. Out of the 8 DIF items, 3 items were identified as their DIF was against the focal group (LEP accommodated student). They were items 12 ($\hat{\alpha}_{MH}=0.66$), 18 ($\hat{\alpha}_{MH}=0.76$), and 19 ($\hat{\alpha}_{MH}=0.52$). Item 12 was a measurement item that required students to read a ruler, while item 18 was a data

analysis item to complete a bar graph. Item 19 was an algebra problem to complete a letter pattern. Interestingly, all these three items were short constructed response items. On the other hand, all 5 items that showed DIF against the reference group (students without accommodations) were multiple-choice items. Among the 5 items that had DIF against the reference group, 3 items (items 1, 3, and 6) were “number and operation” items. Two other items (items 13 and 15) that showed DIF against the reference groups were geometry items. Interestingly, no items in the “number and operations” and geometry sub domains showed significant DIF against the reference groups, while no item in the measurement, data analysis, and algebra sub domains showed significant DIF against the focal group.

Table 2. Results of M-H DIF analysis

Item	α_{MH}	95% CI
1	1.61*	1.23-2.09
2	0.97	0.78-1.19
3	1.24*	1.03-1.48
4	1.13	0.93-1.36
5	1.00	0.82-1.22
6	1.26*	1.02-1.57
7	0.94	0.76-1.17
8	0.92	0.75-1.13
9	1.05	0.84-1.32
10	0.87	0.71-1.07
11	1.09	0.90-1.33
12	0.66*	0.45-0.96
13	1.21*	1.01-1.46
14	0.86	0.72-1.04
15	1.57*	1.30-1.90
16	0.91	0.71-1.17
17	0.84	0.70-1.01
18	0.76*	0.63-0.92
19	0.52*	0.43-0.63
20	1.07	0.86-1.33
21	1.21	0.95-1.5

*significantly different from 1.0 at $\alpha = .05$ level

Next, the DIF detection model by the HGLM was fitted in order to test the consistency between the M-H results presented above. The results are summarized in Table 3. The scale of DIF coefficients in this table is in log-odds-ratio, which is different from the M-H results in Table 2. In the log-odds-ratio scale, odds-ratio of 1 is transformed to 0. Because of coding convenience, a positive log-odds-ratio is the same as odds ratio of smaller than 1.0, while a negative log-odds-ratio is the same as odds ratio of larger than 1.0.

Based on the HGLM DIF detecting model, 6 items were identified as their log-odds-ratios were significantly different from 0. These 6 items were consistent with the M-H results. However, 2 items (items 3 and 6) were not identified as displaying DIF based on the 2-level HGLM model. Since the subsequent analyses will be based on the variant of 2-level HGLM DIF detecting model, we decided to concentrate on the 6 items for the subsequent analyses.

Table 3. Results of three-level DIF detecting model followed MH-DIF item

Item	DIF coeff.	SE	p-value
1	-0.394	0.144	0.007
12	0.462	0.193	0.017
13	-0.541	0.108	0.000
15	-0.409	0.150	0.000
18	0.240	0.104	0.021
19	0.517	0.111	0.000

Next, the random effect DIF model was fitted for the 6 items, simultaneously. Estimates of the variation of the DIF (in the scale of standard deviation) for the 6 items are summarized in Table 4. Among the 6 items, 2 items (items 13 and 15) with p -values smaller than .05. The estimate of SD of DIF between schools for items 13 and 15 were .819 and .815, respectively. These are two items that had DIF in such a way that accommodated students had higher odds of correct answer than non-accommodated students. They can be interpreted that 95% of school-level DIF magnitudes are captured approximately in the range of (overall DIF) \pm 2SD, assuming the normality of the DIF distribution across schools. Since the overall DIF magnitudes were estimated as .541 and .409 for these two items (see Table 3), such ranges are $.541 \pm 2 \times .819 = [-1.097, 2.179]$ and $.409 \pm 2 \times .815 = [-1.221, 2.039]$. For other 4 items, the SD estimates were as large as items 13 and 15, such as 1.266 for items 12. However, the data did not show statistical evidence that they are different from zero.

Table 4. Results of random effects for only the items presenting DIF

Random effects	SD	χ^2	df	p-value
Student	0.592	4575.827	2198	0.000
School	0.366	799.783	234	0.000
Item1	1.012	253.253	234	0.185
Item12	1.266	221.848	234	>.500
Item13	0.819	272.435	234	0.043
Item15	0.815	277.963	234	0.026
Item 18	0.702	239.541	234	0.388
Item 19	0.826	265.927	234	0.074

Finally, the explanatory DIF model was fitted. Five school-level variables were considered. Descriptions of the 5 dichotomously coded variables and basic descriptive statistics are summarized in Table 5. All 5 variables had reasonably well balanced distributions of two response categories.

Table 5. Descriptions and descriptive statistics of school characteristic variables

label	description	Coding	N	Mean
LEPENROLL	Less than 50% enrollment identified as LEP	0=No, 1=Yes	236	.674
EXTMATH	Fourth grade extracurricular for math	0=No, 1=Yes	236	.335
ESL	Over 10% students with ESL instruction	0=No, 1=Yes	236	.551
REPEAT	No fourth graders held back or repeat	0=No, 1=Yes	236	.479

Four school-level variables were included in the explanatory DIF model separately. However, the effect of each school characteristic variable was estimated for all 6 suspected items simultaneously. Results are summarized in Table 6a-6e. Two cases resulted in a significant effect of a school characteristic variable on DIF variation. Again, it is equivalent to a three-way interaction effect between (item difficulty) \times (group) \times (school characteristic).

In one case, LEPENROLL was significant for the DIF variation of item 12 (see Table 6a). The coefficient for LEPENROLL was estimated as .981 ($se = .453, p = .030$), indicating that schools with less than 50% enrollment of LEP students had higher values of DIF coefficient. Recall that the overall value of DIF coefficient for item 12 was .462 (see Table 3), indicating that the DIF was against accommodation group. Therefore, it should be interpreted that schools with less than 50% LEP enrollment had higher DIF against accommodation group than schools with 50% or more LEP enrollment. Some explanation can be speculated, such as a possibility of more attention to get LEP student accustomed to test accommodations may be associated with large LEP enrollment schools. However, note that this item did not display a significant variation of DIF. Since the effect is a three-way interaction effect between (item difficulty) \times (group) \times (LEP enrollment), the interpretation does not have to be a differential DIF (item difficulty \times group interaction) between high and low LEP enrollment schools. For example, the interpretation could be a differential (item difficulty) \times (LEP enrollment) interaction (DIF based on LEP enrollment) between accommodated and non-accommodated students. More research is needed to fully explain why it is the case.

In another case, REPEAT was significant for the variation of item 13 (see Table 6d). The coefficient for REPEAT was estimated as .515 ($se = .242, p = .033$), indicating that schools with no held-back 4th grade students had higher DIF coefficient values. However, this does not mean a stronger DIF for schools with no held-back 4th graders. Recall that Item 15 had an overall DIF coefficient of $-.409$ that indicated DIF against non-accommodation group (see Table 3). Therefore, the result indicates that the magnitude of DIF is closer to 0 for schools with no held-back 4th grade students.

On the other hand, no significant result was obtained for EXTMATH and ESL, indicating the availability of extracurricular math program or proportion of ESL students did not explain the variation of DIF across schools. However, we attempt to interpret the magnitude and direction of the effect of those two variables on items 13 and 15 (items with significant DIF variation between schools – see Table 4) here. The effect of EXTMATH was near zero ($.087, se = .254, p = .733$) for item 13, while it was substantially large for item 15 ($-.245, se = .248, p = .325$). For item 15, it can be interpreted that the schools with extracurricular math program had lower DIF coefficient. In this case, the DIF magnitude was stronger against accommodated students because the overall DIF coefficient was negative. On the other hand, the effect of ESL was near zero for item 15 ($.025, se = .238, p = .916$), while it was fairly large for item 13 ($.219, se = .244, p = .369$). Therefore, it can be interpreted that the school with over 10% ESL enrollment had higher DIF coefficient, on average, than other schools by .244, indicating smaller DIF against accommodated students.

Table 6. Results of random-effect DIF model with s school characteristic variable

a. School variable = LEPENROLL						
Item	School variable = LEPENROLL			Random DIF effects		
	Coeff.	SE	<i>p</i> value	SD	χ^2	<i>p</i> value
1	0.387	0.326	0.235	0.999	249.550	0.218
12	0.981*	0.453	0.030	1.289	223.073	>.500
13	-0.012	0.251	0.961	0.820	272.624	0.038
15	0.278	0.249	0.265	0.819	277.927	0.023
18	-0.027	0.237	0.911	0.702	239.548	0.368
19	0.347	0.256	0.175	0.834	267.058	0.062

b. School variable = EXTMATH						
Item	School variable = EXTMATH			Random DIF effects		
	Coeff.	SE	<i>p</i> value	SD	χ^2	<i>p</i> value
1	-0.097	0.333	0.772	1.012	253.196	0.174
12	-0.035	0.436	0.937	1.266	221.855	>0.500
13	0.087	0.254	0.733	0.818	272.055	0.040
15	-0.245	0.248	0.325	0.810	276.518	0.026
18	0.321	0.238	0.178	0.698	237.742	0.402
19	0.288	0.259	0.267	0.823	264.940	0.074

c. School variable = ESL						
Item	School variable = ESL			Random DIF effects		
	Coeff.	SE	<i>p</i> value	SD	χ^2	<i>p</i> value
1	0.362	0.322	0.261	1.016	252.852	0.178
12	0.133	0.419	0.751	1.265	221.186	>.500
13	0.219	0.244	0.369	0.819	272.150	0.040
15	0.025	0.238	0.916	0.814	277.853	0.023
18	0.142	0.227	0.533	0.702	239.442	0.372
19	-0.066	0.247	0.791	0.827	266.103	0.067

d. School variable = REPEAT						
Item	School variable = REPEAT			Random DIF effects		
	Coef	SE	<i>p</i> value	SD	χ^2	<i>p</i> value
1	0.368	0.317	0.247	0.999	249.874	0.214
12	0.768	0.427	0.071	1.291	223.230	>.500
13	0.515*	0.242	0.033	0.804	265.979	0.068
15	0.038	0.237	0.873	0.814	277.883	0.023
18	0.206	0.227	0.364	0.700	238.541	0.388
19	0.331	0.248	0.182	0.829	265.976	0.068

Conclusions

The current study demonstrated a random-effect DIF model by HGLM and its utility using a part of NAEP 2003 mathematics assessment for the 4th graders. Results of this study provides in-depth information about why some schools have large DIF between accommodated and non-accommodated students. Although this study identified only two school characteristic variables that are related to the variation of DIF, there may be other variables that are more informative. If variables related to environment to enhance learning activities for students with limited English skills are identified as sources of DIF variation, findings could suggest better test preparation environment for students who need test accommodations. Also, it is highly expected the combination of item, person, and school characteristic variables are examined in order to explain DIF variations across schools.

From the technical side, it is widely recognized that the number of clusters (schools, in this paper) and the cluster size (the number of students in each school, in this paper) are highly relevant to obtain good estimates of level-3 variance (DIF variation, in this paper). Therefore, it is hoped that a systematic study on the effect of the number of clusters and the cluster size is conducted in the future. Also, this study employed PQL estimation; an estimation method that approximate a marginal likelihood by the first two moments. PQL is very quick in convergence, but very rough in estimation especially with small number of clusters and small cluster size. Even though we attempted to maximize cluster sizes in the data, it is still unknown how well we were able to estimate parameters. Further investigations on this issue for PQL, as well as for alternative estimation/optimization methods, such as Gaussian quadrature, Laplace approximation, and fully Bayesian methods, are expected in the near future.

References

- Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, 37, 307-327.
- Chaimongkol, S. (2005). *Modeling differential item functioning using multilevel logistic regression models: A Bayesian perspective*. Unpublished Doctoral Dissertation, Florida State University.
- Cheong, Y. F. (in press). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*.
- Gordon, M., Lewandowski, L., Murphy, K., & Dempsey, K. (2002). ADA-based accommodations in higher education: A survey of clinicians about documentation requirements and diagnostic standards. *Journal of Learning Disabilities*, 35, 357-363.
- Holland, P.W., & Thayer, D.T. (1998). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.
- Kamata, A. & Binici, S. (2003). *Random Effect DIF Analysis via Hierarchical Generalized Linear Modeling*. Paper presented at the biannual International Meeting of the Psychometric Society, Sardinia, Italy.
- Kamata, A. (1999). *Some generalizations of the Rasch model by the hierarchical generalized linear model*. Unpublished Doctoral Dissertation, Michigan State University.
- Maller, S. J. (2001). Differential Item Functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, 61, 793-817.
- Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer-Verlag.
- Muthen, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Erlbaum.
- Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

- Raudenbush, S. W., Bryk, A.S., Cheong, Y. F., & Congdon, R. (2004). *HLM6: Hierarchical linear and nonlinear modeling* [Computer Program]. Chicago: Scientific Software International.
- Rogers, A. M., Kokolis, G. A., Stoeckel, J. J., & Kline, D. L. (2002). 2000 mathematics assessment secondary-use data files: Data companion. Washington, DC: National Center for Educational Statistics.
- Schulte, A. A., Elliott, S. N., & Kratochwill, T. R. (2001). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities. *School Psychology Review, 30*, 527-547.
- Shealy, R., & Stout, W.F. (1993). An item response theory model for test bias. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.
- Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.
- Swanson, D.B, Clauser, B. E, Case, S. M, Nungester, R. J., & Morrison, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*, 53-75.
- Thissen, D., Steinberg, L. & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Erlbau.
- Walker, C.M., & Beretvas, S.N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38*, 147-163.