Thailand Statistician
January 2008; 6(1) : 65-74
http://statassoc.or.th
Contributed paper

The Functionality of a Hierarchical Generalized Linear Model Used for Detecting Differential Item Functioning with a Fisher-Tippett Ability Distribution

Sutipon Surathanee and Saengla Chaimongkol*
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Phathum Thani, 12121, Thailand.
Author for correspondence; e-mail: saengla@mathstat.sci.tu.ac.th

Received: 1 August 2007
Accepted: 25 October 2007.

Abstract

The purpose of this study is to evaluate the hierarchical generalized linear model (HGLM) used for detecting differential item functioning (DIF) as proposed by Kamata when the distribution of examinees' abilities is a Fisher-Tippett distribution. In the study, 1,000 examinees are divided equally into two groups. In the study, we investigated four factors. The first factor is whether or not there is difference in mean ability between two groups. The second factor is the DIF proportion, which is arbitrary set to 3 different levels, 5%, 15%, and 30%. The third factor is the DIF magnitude, which is set to 3 levels in a logit scale, 0.3, 0.5, and 0.7. The fourth factor is the type of method used for detecting DIF of the HGLM, a simultaneous method or an item-by-item method. This work is a simulation study in which the data are generated by using R Program version 2.5.1. After the data are generated, they are analyzed by using HLM version 6.0, with 100 runs for each set of conditions. The performance of the HGLM used for detecting DIF is evaluated using hit and false alarm rates. The result from this study is that the Fisher-Tippett distribution as an ability distribution does not affect the detection of DIF in a significant manner for the HGLM-DIF model.

Keyword: HLM program, item response theory, multilevel modeling.

1. Introduction

Differential item functioning (DIF) refers to differences in item performance between two or more subpopulations of examinees when their abilities are made equal. This means that there is relationship between the item response and group variables after controlling for the ability level. Non-DIF can be statistically defined as (Chang et al., [1]):

$$E(Y \mid \Theta = \theta, G = R) = E(Y \mid \Theta = \theta, G = F) \quad \forall \theta, \qquad (1)$$

where Y is a dichotomous score of the studied item (Y =0 or Y =1), θ is a trait variable (either latent or observed) measured from the test and used as a matching criterion, G is a group indicator variable where G=R represents a reference group and G=F represents a focal group, and $E(Y \mid \Theta = \theta, G = R)$ and $E(Y \mid \Theta = \theta, G = F)$ are the regressions of Y on the matching variable for the reference group and focal group, respectively. An item is considered to be non-DIF item when the relationship of equation (1) holds.

The assessment of DIF is an essential step in the validation of educational and psychological tests. If an item or the whole test is biased, the inferences and decisions about the true ability of examinees may not be made correctly on the basis of such a test. DIF detecting procedures are designed to detect such differential item validity. Currently, many statistical techniques have been proposed based upon various theoretical backgrounds and practical purposes. Recently, Kamata [4] has proposed a hierarchical generalized linear model used for detecting DIF, which is named the HGLM-DIF model. The HGLM-DIF model considers a two-nested-level structure of data in which the items are level-1 and are nested to the students, who are in level-2. The level-1 model is an item level where the logit link function is utilized to relate the probability of answering correctly to linear predictors of item dummy codes. The level-2 model is a person level where the intercept coefficient and the item coefficients that are suspected to have DIF from level-1 are modeled by adding a group indicator (0: reference, 1: focal groups). The coefficients of the group indicator for a particular item can be interpreted as the difference in the item difficulty parameter between the reference and focal groups, which is the DIF magnitude. Many researchers have studied and extend Kamata's model in order to detect DIF in many focuses. For example, Luppescu [7] has compared the efficiency of detecting the DIF of the Kamata and Rasch methods by using the difference in item difficulty parameters between two groups. Kim [6] has modified level-2 of Kamata's model by adding matching variables

estimated from Kamata's multilevel item analysis model [5], group variables, and their interaction terms. Chaimongkol [2] has extended Kamata's models of detecting DIF for data with two nested levels to data with a three-nested-level structure and has used a Bayesian approach to estimate parameters. Kamata's model was implemented by using HLM 6 software that uses the penalized or predictive quasi-likelihood (PQL) method to approximate maximum likelihood. This estimation method requires a normal distribution of examinees' ability. In a real testing situation, the ability does not necessarily follow a standard normal but follows one of the various other distributions according to how the test was constructed.

Thus, the purpose of this study is a Monte Carlo investigation of the functionality of Kamata's HGLM model of detecting DIF when the distribution of examinees' ability is Fisher-Tippett distribution.

2. Theoretical Framework

2.1 Hierarchical Generalized Linear Model for Detecting DIF (HGLM-DIF)

The hierarchical generalized linear model for detecting DIF proposed by Kamata [4] extends the hierarchical Rasch model by including person characteristic variables in the model. Let y_{ij} = 1 if person *j* responds to item *i* correctly and y_{ij} = 0 if otherwise, and let p_{ij} be the probability that y_{ij} = 1. This probability varies according to the person. However, as a condition of this probability, we have $Y_{ij} \mid p_{ij} \sim$ Bernoulli, with $E(Y_{ij} \mid p_{ij}) = p_{ij}$ and $V(Y_{ij} \mid p_{ij}) = p_{ij}(1 - p_{ij})$. Therefore, the level-1 model (item-level) is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij}, \qquad (2)$$

where X_{qij} is the item dummy variable for item $i\,(i = 1, 2, \ldots, k)$ responded to by person $j\,(j = 1, 2, \ldots, n)$, with values -1 when $q = i$ and 0 when $q \neq i$. Thus, β_{qj} is the item effect for the *q*th item, and β_{oj} is the intercept for the model which indicates the overall effect to all items.

For the level-2 model, β_{oj} is assumed to have a random effect across the population of people. Thus, a latent trait that is common to all items but varies across people can be modeled. The level-2 model is a person model specified as

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

and (3)

$$\begin{cases} \beta_{qj} = \gamma_{q0} & \text{if no DIF,} \\ \beta_{qj} = \gamma_{q0} + \gamma_{q1} G_j & \text{otherwise.} \end{cases}$$

Here, G_j is the group of person who was coded as 1 if the person belongs to focal group and coded as 0 if the person belongs to reference group. γ_{00} is the mean ability of reference group, while γ_{01} is the mean ability difference between the focal and reference groups. γ_{q1} is the difference of difficulty between focal and reference groups for the *q*th item or DIF magnitude. Thus, the item *q*th presents DIF when the hypotheses of $H_0 : \gamma_{q1} = 0$ are rejected by the chi-square tests with one degree of freedom.

However, this model is assumed that u_{0j} is normally distributed with a mean of 0 and variance of τ, $(N(0,\tau))$. This study investigates the performance of the HGLM-DIF with a violation of the ability distribution assumption.

2.2 Fisher-Tippett Distribution

The Fisher-Tippett distribution is also known as log-Weibull distribution (http://en.wikipedia.org/wiki/Fisher-Tippett_distribution). Fisher-Tippett distribution is used as an alternative to the normal distribution in the case of skewed data. The probability density function is

$$f(x) = \frac{Z \exp(-Z)}{\beta} \text{ , where } Z = \exp\left[-\frac{x-\mu}{\beta} \right], \quad (4)$$

$-\infty \le x \le \infty$, μ is location parameter, and $\beta > 0$ is scale parameter. For generating the Fisher-Tippett variate, a random U variate is generated first from a uniform distribution in the interval (0, 1]. Then the random U variate is transformed to the Fisher-Tippett variate (X) by using the relationship of $X = \mu - \beta \ln(-\ln(U))$. This study uses $\mu = -0.5772$ and $\beta = 1$ in order to get the mean of 0. However, the choice of these values does not affect the results of the simulations. The generated random numbers are trimmed to interval [-3, 3] and treated as person abilities in which the mean is 0, the standard deviation is 1.283, and the skewness is $\frac{12\sqrt{6}\Gamma(3)}{\pi^3} \approx 1.14$.

3. Simulation Study

The simulated dichotomous data is based on the model:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \gamma_{01}G_j + \gamma_{q0} + \gamma_{q1}G_j + u_{0j},$$

which is the combined level-1 and level-2 models. The data sets were generated for 1,000 examinees and divided equally into two groups (reference and focal) in which each examinee answers all the items in the test. The item difficulty parameter of the reference group (γ_{q0}) for DIF items was fixed to 0, and for non-DIF items it was set to -1, 0, or 1. The other conditions were set as follows:

1. The number of items in the test was fixed at 20 and proportion of DIF-items was set to 3 different levels, 5%, 10%, and 15%. Thus, the total number of DIF items in the test was 1, 3, and 6, respectively.
2. The difference of the item difficulty parameter between the reference and focal groups or DIF magnitude (γ_{q1}) was set to 3 different levels in the logit scale, 0.3, 0.5, and 0.7.
3. The true values of the examinee ability parameter (u_{0j}) for both the reference and focal groups were randomly generated from normal and Fisher-Tippett distributions in which the mean ability of the reference group

(γ_{00}) was fixed to 0. The mean ability of the focal group compared to the reference group, or the mean difference between the focal and reference groups (γ_{01}), was set to either 0 or-1. In the case of $\gamma_{01} = 0$, the mean ability parameter for the reference and focal groups was 0. In other words, there was no difference in mean ability between two groups. For the case of $\gamma_{01} = -1$, the mean ability of the reference and focal groups was 0 and -1, respectively. This implies that the focal group has a lower ability than the reference group.

4. Two methods of detecting DIF were studied, both the item-by-item and simultaneous methods. Only the level-2 models of these two detecting methods were different. For the item-by-item detecting method, only the item coefficient of testing item was modeled by adding the group indicator. On the other hand, in the simultaneous method, the group indicator was added to all *k-1* item coefficients in the level-2 model except for the reference item.

For each simulation condition, data were generated for 100 sets by using the R 2.5.1 program. Once the data were generated, the HGLM-DIF model for each detecting method was fitted to estimate and test the statistical significance of their parameters. HLM 6 software (Raudenbush, Bryk, Cheong, and Congdon [9]) was used to estimate the HGLM-DIF models' parameters. In HLM 6, parameters were estimated by a sixth order approximation of the likelihood for the model based on a Laplace transformation. In this study, only the DIF parameters were of interest, and these were examined to investigate the functioning of the HGLM-DIF model by evaluating the hit and false alarm rates. The hit rate means a number of times that a DIF item resulted in a significant DIF flag from 100 runs, whereas the false alarm rate indicates a number of times that a non-DIF item showed a significant DIF flag. The results of the hit and false alarm rates under all the study conditions are presented in Table 1 and Table 2, respectively.

4. Results

Table 1 shows the results of the hit rates for the DIF items under all study conditions. It can be concluded from this table that the HGLM-DIF model with a Fisher-Tippett distribution performed in a similar fashion to the HGLM-DIF model with a normal distribution. The hit rates did not depend on whether or not there is difference in mean ability between two groups and the number of DIF items in the test, but they varied

across the detecting methods and the DIF magnitudes. The item-by-item detecting method showed higher hit rates across different DIF magnitudes than those of the simultaneous method. The hit rates were proportional to the DIF magnitude. When the DIF magnitude in logit is 0.7, the item-by-item method had almost perfect hits, whereas the simultaneous detecting method revealed an approximate average of 90% for hits. In contrast, when the DIF magnitudes were decreased to 0.3 in logit, the hits for the item-by-item and the simultaneous methods were, on average, less than 45% and 30%, respectively.

Table 1. The hit rates for the DIF items under all conditions.

#DIF-Item	DIF Size	Normal distribution				Fisher-Tippett Distribution			
		$\gamma_{01}=0$		$\gamma_{01}=-1$		$\gamma_{01}=0$		$\gamma_{01}=-1$	
		A	B	A	B	A	B	A	B
1	0.3	32	47	33	51	24	56	27	42
	0.5	68	87	70	86	61	95	75	91
	0.7	89	98	90	100	94	100	95	100
3	0.3	24	29	28	31	22	22	33	51
	0.5	60	85	62	72	65	82	70	86
	0.7	91	99	85	98	91	100	90	100
6	0.3	30	19	31	31	25	12	20	22
	0.5	75	84	70	70	66	67	63	83
	0.7	93	98	94	94	90	98	96	100

"$\gamma_{01}=0$" means there is no difference in mean ability between reference and focal groups.

"$\gamma_{01}=-1$" means there is difference in mean ability between reference and focal groups.

"A" means the Simultaneous detecting DIF method and "B" means the item-by-item detecting method.

The means of false alarm rates computed from the non-DIF items that have the same item difficulty as the reference group are shown in Table 2. The observations of false alarm findings showed that the HGLM-DIF model worked equally well regardless of its latent ability distribution. The false alarms did not depend on whether or not there was a difference in mean ability between two groups, but they did depend on the detection method. The false alarms of the simultaneous method were close to the 5% nominal error level across the number of DIF items in the test and item difficulty parameters. In contrast, the item-by-item method inflated the false alarms when there are 3 or 6 DIF items in the test. Especially when there are 6 DIF items in the test, the false alarms deviated much more clearly from the 5% nominal level and they were also apparently higher for the Fisher-Tippett distribution than normal distribution.

Table 2. The mean of false alarm rates for the non-DIF items under all conditions.

#DIF-Item	γ_{q0}	Normal distribution				Fisher-Tippett Distribution			
		$\gamma_{01}=0$		$\gamma_{01}=-1$		$\gamma_{01}=0$		$\gamma_{01}=-1$	
		A	B	A	B	A	B	A	B
1	-1.0	4.7	4.4	4.4	5.0	5.0	6.4	5.4	5.5
	0.0	4.0	4.7	4.2	5.1	3.7	4.4	5.2	5.4
	1.0	4.4	5.1	4.7	4.6	4.4	4.7	4.7	4.4
3	-1.0	4.3	6.0	7.0	7.1	5.0	7.3	5.3	7.8
	0.0	3.6	8.0	5.7	7.0	4.8	7.8	6.4	7.6
	1.0	4.5	7.7	6.3	8.3	5.6	9.0	5.6	8.2
6	-1.0	6.2	13.0	4.3	12.2	5.0	19.2	6.0	18.6
	0.0	5.0	12.0	5.5	12.1	4.6	18.2	6.6	18.4
	1.0	4.3	13.1	4.4	12.3	4.7	18.7	5.3	19.0

"γ_{q0}" means the item difficulty for the reference group.

"$\gamma_{01}=0$" means there is no difference in mean ability between reference and focal groups.

"$\gamma_{01}=-1$" means there is difference in mean ability between reference and focal groups.

"A" means the simultaneous method of detecting DIF and "B" means the item-by-item detecting method.

The mean of false alarm rates is computed from all items in the test which have the same values of γ_{q0}.

5. Conclusion and Discussion

The HGLM-DIF model where the ability of examinees follows a normal distribution is anticipated to behave in a superior way to the model with a Fisher-Tippett distribution. However, the findings from this study indicate that there was not a distinction of behavior for the HGLM-DIF model with a normal and a Fisher-Tippett distribution. Shin and Wall [10] also found that a violation of the ability distribution assumption showed little impact on the behavior of the IRT test for DIF detection. A more interesting and practically valuable finding from this study was that the performance of the HGLM-DIF model did not depend on the existence of a difference in mean ability between two groups. This result implied that Kamata's HGLM-DIF model was a good and suitable model for detecting DIF in a real test situation. This is because the HGLM-DIF model can make a distinction between item impact and DIF. Item impact refers to a group difference

in measured performance on a test or test items (Dorans & Holland [3]). In addition, one more benefit of the HGLM-DIF model was that it can detect all items except the reference item in the test simultaneously, and its performance was more prevalent and reliable than item-by-item detection. In summary, the HGLM-DIF model appears to be a good device to detect DIF in most testing and measurement data regardless of its latent ability assumption.

References

[1] Chang, H. Mazzeo, J., and Roussos, L. Detecting DIF for Polytomously Scored Items: An Adaptation of the SIBTEST Procedure. Journal of Educational Measurement, 1996; **33**: 333-353.

[2] Chaimongkol S. Modeling Differential Item Function (DIF) using Multilevel Logistic Regression Models: A Bayesian perspective. A Dissertation submitted to the Department of Statistics, The Florida state University College of Arts and Sciences, USA, 2005.

[3] Dorans, N. J., and Holland, P.W. DIF Detection and Description: Mentel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.), Differential Item Functioning, Hillsdale, NJ: Erlbaum.1993: 35-66.

[4] Kamata, A. Some generalization of the Rasch Model: An application of the Hierarchical Generalized Linear Model. A Dissertation submitted to Michigan State University, U.S.A., 1998.

[5] Kamata, A. Item Analysis by the Hierarchical Generalized Linear Model. Journal of Educational Measurement, **38**, 2001:79-93.

[6] Kim, W. Development of a Differential Item Function (DIF) procedure using the Hierarchical Generalized Linear Model: A comparison study with Logistic Regression procedure. A Thesis submitted to the Pennsylvania State University, USA, 2003.

[7] Luppescu, S. DIF Detection in HLM Item Analysis. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, 2002.

[8] R Development Core Team. R: Language and Environment for Statistical and Computing. http://www.R-project.org. 2007.

[9] Raudenbush, S. W., Bryk, A.S., Cheong, Y.F., and Congdon, R. HLM6: Hierarchical Linear and Nonlinear Modeling [Computer Program].

2004, Chicago: Scientifics Software International. 405 p.

[10] Shin S.H. and Wall, N.L. Three Differential Item Function Methods with Three Different Ability Distributions. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, 2006.