Kuram ve Uygulamada Eğitim Bilimleri • Educational Sciences: Theory & Practice - 11(1) • Winter • 284-288

Sample Size in Differential Item Functioning: An Application of Hierarchical Linear Modeling

Tülin ACAR[a]
Parantez Education Research Publisher

Abstract

The purpose of this study is to examine the number of DIF items detected by HGLM at different sample sizes. Eight different sized data files have been composed. The population of the study is 798307 students who had taken the 2006 OKS Examination. 10727 students of 798307 are chosen by random sampling method as the sample of the study. Turkish, science, and social studies subtests, all composed of 25 items and applied in the OKS-2006, are used as data gathering instruments in this study. It has been concluded that varieties in sample sizes have a great effect on DIF detection in test items.

Key Words

Differential Item Functioning, Hierarchical Generalized Linear Model, Sample Size.

In the studies of social sciences having accurate, highly reliable and acceptable measurements and discussions is really hard and very important partly because of the nature of variables examined. Because social sciences based on human beings, it is sometimes technically insufficient to measure the nature of human beings as they have such a complex structure. In physical sciences, with the availability of direct measures, the determination of the direction and magnitude of the systematic and fixed errors which have effects on the measurement results is much easier. However; in social sciences, it is not easy to determine the direction and magnitude of systematic and fixed errors in measurement results, as the measurements are commonly indirect. In educational studies, psychological constructs of individuals such as achievement, ability, and personality are often measured. It is important to answer the questions of how to measure psychological constructs of individuals and what decisions to be made according to measurement results. As these two questions are so critical, the size of systematic and fixed errors affecting measurement results becomes more important for the validity of measurement instruments and results.

With the validity of test items and measurement instruments used in education, the validity of measurement is one of the main problems of bias measuring. As it is known, one of the main objectives of measuring applications in education is to obtain information about individuals and test items. Highly valid and accurate measurement instruments and results are needed to achieve this objective. However; one of the factors which have a negative effect on validity is biased items. The existence of biased items in a test decreases the reliability of the discussions made.

Item bias is said to be a result of "systematic errors" which have an effect on measurement results. It does not affect all the results equally owing to the description of systematic errors. The existence of items including systematic errors is a problem strongly related to the validity of the test. In validity analysis, it is important to detect biased items among the test items. This is about detection of "Differential Item Functioning" which can be determined by statistical methods.

In recent studies, differential item functioning (DIF) typically refers to item bias (Ellis & Raju, 2003). In the late 1980s, the term "DIF" have changed place with the term "item bias." DIF re-

a *Correspondence:* Tülin ACAR, PhD., Measurement and Evaluation specialist, Parantez Education Research Publisher, Selanik Street No:46/4 Kızılay-Çankaya, Ankara/TURKEY. E-mail:totbicer@gmail.com. Phone: +90 312 425 1995 Fax: +90 312 425 1995.

veals the differences in the probability of answering the item correctly according to the subgroups at every ability level of the psychological structure that is intended to be measured with the item (Embretson & Reise, 2000; Lord, 1980). In studies on DIF, there is a requirement of performance comparison on test items of groups in the same capability level but having different demographic characteristics such as male-female or Asian-European (Greer, 2004).

In the case of existence of DIF in the test items, this may be caused by real differences (item impact) or item bias in the subgroups (Zumbo, 1999). There are lots of methods for DIF detection. Some of these methods are based on classical test theory. Mantel-Haenszel (M-H), LR and SIBTEST are the examples of the methods based on classical test theory (Gierl, Khaliq, & Boughton, 1999). Some DIF detection methods such as Lord's chi square test, Raju's area measures and likelihood ratio are the samples of DIF detection methods based on item response theory (Öğretmen, 1995; Zwick, Donoghue, & Grima, 1993). Most of these methods provide similar information about DIF. There are lots of DIF detection studies made by M-H technique in the literature (Allalouf, 2003; Duncan, 2006; Gondal, 2001; Hamzeh & Johanson, 2003; Öğretmen, 2006; Randall, 2001; Yıldırım, 2006; Yurdugül, 2003). LR method and likelihood ratio based on the item response theory gained importance against M-H method in DIF detection studies by the late developing methods. However; in educational research, it has been discovered that data are in a hierarchical structure. As a result, HGLM method became remarkable in DIF detection studies (Chaimongkol, Huffer, & Kamata 2007; Kamata, Chaimongkol, Genç, & Bilir 2005; Luppescu 2002; Vaughn 2006; Williams 2003). HGLM, M-H and logistic regression methods are similar to each other as they are based on observed scores (Binici, 2007). This study focuses on the HGLM method. HGLM is a method that derives linear equations which explains individuals' characteristics and characteristics of group members as a function of the group formed by individuals and group members. Estimator variables of students' characteristics are added to level2 model in order to detect whether the characteristics of students have an effect on the possibility of giving answer correctly to test items or not- which is a DIF detection study on item. In HGLM, level 1 (item level) and level 2 (individual level) modeling in which item scores (result) have two categories are set (Kamata, 2002).

Purpose of the Study

The purpose of this study is to examine the number of DIF items detected by HGLM at different sample sizes. In tests which measures different skills, examination of effects of sample size on DIF is important as HGLM is a new method.

Method

This study is a descriptive research which examines whether the DIF results determined by the HGLM Method vary with the sample size or not.

Sample

The population of the study is 798307 students who took the 2006 OKS Examination. 10727 students of 798307 are chosen by random sampling method as sample.

Instrument

Turkish, science and social studies subtests, all composed of 25 items and applied in the OKS-2006, are used as data gathering instruments in this study.

Data Analysis

As the DIF detecting study is made according to gender, subgroups were made according to variety of gender. Female students were chosen to be the focus group and male students were chosen to be the reference group. HLM-6.04 (Raudenbush, Bryk, Cheong & Congdon, 2001) program was used in DIF detection study by HGLM. In HGLM, level-1 and level-2 equations are established as follows, to determine the DIF with conditional modeling (Kamata, 2002):

Level-1 Equation (Item Level): To show the i (i=1,2,….k) item and j (j=1,2,….N) individual index.

$$\eta_{ij} = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{(k-1)j}X_{(k-1)ij} + r_{ij}$$

η_{ij}: Estimated outcome variable, i.e., the probability of the individual j in giving the correct answer to the item i.

X_{qij}: Indicator variable for item i. When the answer given to an item is on item i (q=i), the value is 1, and in other condition ($q \neq i$), the value is 0.

β_{0j}: It is the breakpoint. When all X_{qij} become 0,

the effect of the item that is not considered for the model occurs. Hence, β_{0j} is the effect of the item that is not considered for the model.

β_{1j}: It is the effect of item 1 on the probability (outcome variable) of individual *j* to give the correct answer up to *i*=1,2,…(*k*–1). The parameters from β_{1j} to $\beta_{(k-1)j}$ is a coefficient that shows the effects of the items on the probabilities of giving the correct answer for the individual from item 1 to item *k*. Individual *j* is associated with different individuals and different item-level parameters. If the level increases, then *j* in B_{ij} decreases, and the item parameters are kept constant among the individuals.

Level 2 is employed to determine the differences between the probabilities of answering each item correctly according to the genders of the students.

Level 2 (Student Level) Equation:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Gender})_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Gender})_j$$

...

$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(\text{Gender})_j$$

β_{ij}: It is the effect of item *i* on the probability of giving the correct answer for individual *j* up to *i*=1,2,... (*k*–1). The parameters from β_{1j} to $\beta_{(k-1)j}$ are the effects of the items on the probability of giving the correct answer from item 1 to item *k* for the individual *j*.

γ_{00}: is the referred item parameter.

γ_{01}: is the difference in the probabilities of giving the correct answer to the related item of the students under the conditions of male and female (gender). In other words, it is the effect of the probability of giving the correct answer to item *i* with respect to the gender variable.

u_{01} is the effect of random gender variable. It is the random effect of b0j , which shows normal distribution that has a distribution average of 0 and variance of τ.

As the purpose of this study is to examine variety in number of DIF items obtained by HGLM according to different sample sizes, 8 different sized data files have been composed. Sample sizes have been defined again like; 1%, 2%, 5%, 10%, 25%, 50%, 75%, 100% of 10727 students. Observation numbers related to the 8 different samples are shown in Table 1.

DIF analysis by HGLM, have been made on different sample sizes which have had varying observation numbers between 97 and 10727. While examining the reliability coefficients of estimations, it has been observed that especially in Turkish and social studies there have been sufficient reliability despite smaller samples in subtests.

Results and Discussion

DIF analysis by HGLM method according to gender, have been obtained from 8 different sized samples for subtests of Turkish, science and social studies separately. Numbers of DIF items obtained by HGLM method at different sample sizes are given in Table 2.

In detection of DIF items, two levels of significance have been considered: 0.05 and 0.01. When the

Table 1.
Sample Sizes and Reliability of Estimations

Representative Sample rate	Sample sizes	Reliability of Estimations		
		Turkish	Science	Social Studies
1%	97	0.850	0,661	0,820
2%	207	0.829	0,728	0,828
5%	532	0.810	0,733	0,832
10%	1055	0.815	0,762	0,837
25%	2681	0,815	0,754	0,840
50%	5320	0,815	0,750	0,839
75%	8037	0,816	0,751	0,838
100%	10727	0,818	0,752	0,840

Table 2.
Number of Observed DIF Items Related to Sample Sizes

Sample Sizes	Turkish		Science		Social Studies	
	<0.05	*<0.01*	*<0.05*	*<0.01*	*<0.05*	*<0.01*
97	2	1	1	0	0	0
207	2	1	1	1	0	0
532	0	0	2	0	0	0
1055	0	0	3	2	3	0
2681	11	5	4	1	6	3
5320	11	6	7	6	9	7
8037	12	11	10	7	13	10
10727	12	12	10	8	15	13

representation ratio of the sample is 25% (n=2681), a remarkable differentiation in the number of DIF items have been obtained at different significance levels. As the numbers of individuals in samples has increased, the number of DIF items has also increased. The number of DIF items obtained in all subtests at 99% confidence level has been nearly the half of the number of DIF items obtained at 95% confidence level. It has been concluded that as the confidence level increases, the number of DIF items decreases in all subtests and at different sample sizes. Another observation obtained in this study is that varieties in sample sizes have a great effect on DIF detection in test items. Vaughn (2006) has applied DIF analysis by the HGLM method on polytomous items in very small sized samples and has determined that the number of estimated DIF items is lower than in bigger samples.

Miller and Spray (1993) applied on the multiple scorable mathematics test of 27 items, have implied that the size of samples have a great effect on DIF item detection, especially if a method based on likelihood ratio is used. As the HGLM method is based on the possibility of 'answering items correctly', the result that Miller and Spray obtained in their study is acceptable in this study also.

In the subtests which measures different abilities, different numbers of DIF items have been obtained by HGLM. Various studies have shown that the presence of multidimensionality may cause DIF (Snow & Oshima, 2009). The undimensionality of the tests, used in the studies, have been examined and great values of DIF items related to gender have attracted notice despite undimentional tests.

According to Roussos and Stout's (1996) simulation studies, no ostensible differences between DIF detection results obtained by the M-H and SIBTEST methods have been seen in small-sized samples. French and Miller (1996) have applied DIF analysis in the samples that they have attributed as small sample (n=500) and large sample (n=2000) by using M-H and logistic regression and have determined that logistic regression method is strongly capable of achieving more accurate results in larger sample sizes. Structurally, DIF detection methods by HGLM and logistic regression techniques are similar to each other. Hence, it can be said that HGLM method is a powerful method in DIF detection studies. In the DIF detection study by HGLM method made on data obtained from a mathematics test which is composed of 39 multiple choice items, it has been emphasized that good estimations can be obtained despite larger sample sizes (Binici, 2007). Luppescu (2002), have discovered that the results obtained by Rasch method and HGLM method are similar to each other, when the ratio of individuals in the focus group and the sample size is small. Kamata (2001), have proved the equality of Rasch method and HGLM method technically in his studies.

Recommendations

By considering the results obtained from the study and the literature, the following recommendations can be listed:

1- The ratio of focus groups and reference groups considered in DIF detection analysis can be examined and discussions can be made on the results.

2- DIF items detected by the HGLM method can be examined in the test which measures different learning fields.

3- The cause of the existence of DIF items detected

by the HGLM method (item bias or item impact) can be determined with the opinions of professionals.

4- It can be determined that if the number of DIF items detected by the HGLM method varies with test length.

References/Kaynakça

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement In Education*, *16* (1), 55-73

Binici, S. (2007). *Random-Effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods*. Unpublished doctoral dissertation, The Florida State University, Florida.

Chaimongkol, S., Hufferve, F. W., & Kamata, A. (2007). An explanatory differential item functioning (DIF) model by the WinBUG 1.4. *Songklanakarin Journal of Science Technology*, *29* (2). Retrieved February 10, 2008 from http://www.sjst.psu.ac.th/journal/29_2_pdf/19item%20response%20theory_449-458.pdf

Duncan, S. C. (2006). *Improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression dıf and mantel-haenszel dıf methods*. Unpublished doctoral dissertation, Sam Houston State University, Texas.

Ellis, B. B., & Raju, N. S. (2003). *Test and item bias: What they are, what they aren't, and how to the detect them*. Retrieved February 10, 2008 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/contentşstorageş01/0000019b/80/1b/57/a3.pdf

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

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33* (3), 315-332.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999, June). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the Improving Large-Scale Assessment in Education Symposium at the Annual Meeting of the Canadian Society for the Study of Education, Canada. Retrieved February 25, 2008 from http://www.education.ualberta.ca/educ/psych/crame/files/dif_csse99.pdf.

Gondal, M. B.(2001). *Differential item functioning analysis of 4Th graders' science and urdu (national language) achievement test items in Pakistan*. Unpublished doctoral dissertation, Middle East Technical University, Ankara.

Greer, T. G. (2004). *Detection of differential item functioning (DIF) on the SATV: A comparison of four methods: Mantel-Haenszel, logistic regression, simultaneous item bias and likelihood ratio test*. Unpublished doctoral dissertation, University of Houston, Texas.

Hamzeh, D., & Johanson, G. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment & Evaluation in Higher Education, 28* (2), 129-134.

Kamata, A. (2001). Item analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement, 38*, 79-93.

Kamata, A. (2002, April). *Procedure to perform item response analysis by Hierarchical Generalized Linear Model*. Paper presented at the annual meeting of the American Educational Research Association, April, New Orleans.

Kamata, A., Chaimongkol S., Genc, E., & Bilir, K. (2005). *Random-Effect differential item functioning across group unites by the Hierarchical Generalized Linear Model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. Retrieved February 25, 2008 from http://garnet.acns.fsu.edu/~akamata/papers/AERA%202005.pdf

Lord, M. F. (1980). *Applications of item response theory to pratical testing problems*. Broadway, Hillsdile, NJ.: Lawrence Erlbaum Associates.

Luppescu, S. (2002, April). *DIF detection in HLM*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30* (2), 107-122.

Öğretmen, T. (1995). *Differential item functioning analysis of the verbal ability section of the first stage of the university entrance examination in Turkey*. Unpublished maters' thesis, Middle East Technical University, Ankara.

Öğretmen, T. (2006). *Uluslararası okuma becerilerinin gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi: Türkiye Amerika Birleşik Devletleri örneği*. Yayınlanmamış doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Randall, D. P. (2001). Assessing differential item functioning among multiple groups: A Comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education*, *14* (3), 235-259.

Raudenbush, S.W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2001). *HLM 5 hierarchical linear and nonlinear modelling*. Lincolnwood: Scientific Software International, Inc.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied ıtem parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33* (2), 215-230.

Snow, T. K., & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement*, *69* (5), 732-747.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A bayesian multilevel approach*. Unpublished doctoral dissertation, The Florida State University, Florida.

Williams, N. J. (2003). *Item and person parameter estimation using hierarchical generalized linear models and polytomous item response theory models*. Texas: The University of Texas at Austin.

Yıldırım, H. H. (2006). *The differential item functioning (DIF) analysis of mathematics items in the international assessment programs*. Unpublished doctoral dissertation, Middle East Technical University, Graduate School of Social Sciences, Ankara.

Yurdugül, H. (2003). *Ortaöğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının madde yanlılığı açısından incelenmesi*. Yayınlanmamış doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (Ordinal) item scores*. Retrieved February 10, 2008 from http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30* (3), 233-251.