In education, the success of students in various areas is examined by achievement tests or by ability tests. The aim of the competitive examinations is to choose the most fitted applicants from different kinds of applicants. The selection of the up-to-grade applicants is related to the specifications of a qualified instrument of measurement. Being biased of the items that form the measuring instrument affects the properties of the measuring instrument. In the measurement results while examining the bias, differential item functioning (DIF) is mostly utilized. DIF is the displaying differences of the probability of answering item correctly according to the subgroups, in every ability level of psychological structure that is aimed to be measured with the item (Embretson & Reise, 2000; Lord, 1980). In DIF studies, the performances of different groups are compared according to the test items related to demographical specifications such as men–women in the same ability level, Asian–European, and so on (Greer, 2004). Uniform-DIF is present if the probability of answering an item of the focused group correctly is higher than the probability of answering an item of the referred group in every ability level. If the probability of answering an item of the focused group correctly differs from the referred group according to its ability level, it is possible to talk about a nonuniform-DIF situation about the item (Zumbo, 2003).

There are many methods present to determine DIF. Some of these methods are related to the classic test theory. The very frequently used Mantel–Haenszel (MH) technique, logistic regression (LR) method, and Simultaneous Bias Test (SIBTEST) can be given as examples for the methods that are related to the classic test theory (Gierl, Khaliq, & Boughton, 1999). Some DIF determining methods are related to item response theory (IRT) for which Lord’s chi-square test, Raju’s area measuring, and likelihood ratio can be shown as examples (Camili & Shepard, 1994; Ogretmen, 1995).

In the area of study, MH method and DIF determining performances were frequently used. However, with the development of the LR and IRT-LR methods, DIF determining techniques were not taken to the scope of this research. The analysis results, which were determined by using the DIF techniques in different sample sizes, can be compared.

Keywords
differential item functioning, hierarchical generalized linear modeling, test bias
performances were given emphasis instead of MH. However, it was observed that the data displayed hierarchical structure in the educational researches, and taking this situation into account, hierarchical generalized linear model (HGLM) technique and DIF determining performances have also become popular (Atar, 2005; Subedi, 2005; Williams, 2003).

Hierarchical Linear Model (HLM) Method

In the recent years in social sciences, the use of proper statistics in the analysis of the nested or hierarchical data has attracted attention (Greer, 2004; National Assessment of Educational Progress [NAEP], 2006). When standard regression equations are applied to hierarchical and nested data, some problems are encountered. Most analysis require independence of observations as a primary assumption for the analysis. But, this assumption is violated in the presence of hierarchical data. Also, hierarchical modeling is similar to that of OLS regression. People or creatures that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population (Bryk & Raudenbush, 1987; Osborne, 2000). For instance, third-grade students in a school are more similar to each other than the students of the other grades. This is because, with the factors such as same teacher, physical environment, similar experiences, and so on, their homogeneity increases.

HLM provides a statistical model that includes multiple-level models. In group researches, Level 1 represents the individual’s level and Level 2 represents the group level. Taking into account that there are different linear regression in every group, the multiple-group factors that have different observation numbers and mixed factors that have multiple specifications can be modeled easily with HLM (Gokiert & Ricker, 2004). HLMs are designed to provide the assumption of independence of the observations from each other, in the conditions where individuals and groups that the individuals belong to are tested together (Raudenbush & Bryk, 1986).

DIF Determining With HGLM

If the outcome variable is measuring results in ordering or classification, HGLM can be used, which is a special form of HLM. Thus, there is no necessity for a conversion process in the outcome variable. In the outcome variables having two categories, binom distribution is taken into account, which is known as Bernoulli distribution, and logit connection function is used (Raudenbush & Bryk, 1986). The logit connection function that is used for the binary outcome variable is used the following way:

\[
\eta_j = \log \left( \frac{\varphi_y}{1 - \varphi_y} \right).
\]

In Equation 1, \( \varphi_y \) shows the probability of “to be” of the outcome variable and the outcome variable takes values between 0 and 1. \( \eta_j \) is the logarithm of probability of “to be” (log-odds).

Predictive variables are added to Level 2 model that reflects the specifications of the student—this is the DIF determining performance on the item—when it is necessary to examine whether the student specifications have impacts on answering the test items correctly. In HGLM, Level 1 and Level 2 equations that are established to determine DIF with conditional modeling are as given below (Williams, 2003):

Level 1 equation (item level) to show \( i (i = 1, 2, \ldots, k) \) item and \( j (j = 1, 2, \ldots, N) \) individual index

\[
\eta_j = \log \left( \frac{P}{1 - P} \right) = \beta_{ij} + \beta_{ij}X_{ij} + \beta_{ij}X_{i2} + \ldots + \beta_{(k-1)j}X_{ik-1j} + \epsilon_j, \quad (2)
\]

where \( \eta_j \) is the estimated outcome variable, in other words, the probability of individual \( j \) to give the correct answer to item \( i \);

\( X_{ij} \) is the indicator variable for item \( i \), when the answer given to an item is on item \( i (q = i) \), the value is 1, otherwise \( (q \neq i) \) the value is 0;

\( \beta_{ij} \) is the intercept of the model. When the all \( X_{ij} \)’s become 0, the affect which is not taken for the model occurs. For this reason, \( \beta_{ij} \) is the effect of the item that is not taken for the model;

\( \beta_{ij} \) is the effect of the first item on the probability (outcome variable) of individual \( j \), to give the correct answer up to \( i = 1, 2, \ldots, (k - 1) \). Parameters from \( \beta_{ij} \) to \( \beta_{(k-1)j} \) are coefficients that show 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, \( j \) in \( \beta_{ij} \) decreases and the item parameters are kept constant between individuals.

Level 2 is formed to see the differences between the probabilities of answering each item correctly according to the gender of the students (Williams, 2003).

Level 2 (student level) equation is as follows:

\[
\beta_{ij} = \gamma_{10} + \gamma_{11}(SES)_j + \nu_{ij},
\]

\[
\beta_{ij} = \gamma_{10} + \gamma_{11}(SES)_j,
\]

\[
\ldots
\]

\[
\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(SES)_j,
\]

where \( \beta_{ij} \) is the effect of item \( i \) on the probability of giving the correct answer for individual \( j \) up to \( i = 1, 2, \ldots, (k - 1) \). Parameters from \( \beta_{ij} \) to \( \beta_{(k-1)j} \) are the effects of the items on the probability of giving the correct answer from the Item 1 to Item \( k \) for individual \( j \).
Acar

\( \gamma_{00} \) is the referred item parameter; 
\( \gamma_{01} \) is the difference between the probabilities of giving the correct answer to the related item of the students with upper and lower socioeconomic status (SES). In other words, it is the effect of the probability of correct answering of item \( i \) on SES variable; and 
\( n_{ij} \) is the effect of random SES variable. It is the random effect of \( \beta_j \), which shows normal distribution that has distribution average 0 and variance \( \tau \).

**DIF Determining With LR**

If the performances of the group members on an item are estimated with LR method, it is possible to talk about a DIF on that item (Swaminathan & Rogers, 1990). For this reason, LR is a method that is used to find out the items containing DIF. With LR method, it is possible to determine both uniform-DIF and nonuniform-DIF. The level of effect can be determined as well. To do this, the standardized regression parameters can be used. Jodoin and Gierl (2001) classified the effect levels of DIF that are determined with LR in the following way.

- **A Level:** If \( R^2 < 0.035 \), a negligible level of DIF is present.
- **B Level:** If \( 0.036 < R^2 < 0.070 \), a medium level of DIF is present.
- **C Level:** If \( R^2 > 0.071 \), a magnitude level of DIF is present.

**DIF Determining With Item Response Theory–Likelihood Ratio (IRT-LR)**

A strong part of DIF determining with IRT is the utility of item response curves and item characteristic curves (Thissen, 2001). If an item functions differently in focused groups and referred groups, in other words, if the item response curves are different for the two groups, presence of DIF is applicable. For both groups, the item parameters are estimated, and the estimated item parameters are compared according to DIF with IRT method. Many softwares have been developed with IRT-LR technique, in determining DIF. In research, results of IRTLRDIF software were used in DIF determining with IRT-LR technique.

In determination of DIF with likelihood ratio, IRTLRDIF program, which was more practical than multilog program, was setup by Thissen (2001). The hypothesis of absence, which is built while analyzing DIF determining with likelihood ratio, is “there is no significant difference between the item parameters that are calculated from focused and referred groups.” In IRTLRDIF program, the results of the compact model (CM) for the test of absence hypothesis and the augmented model (AM) are compared. In the CM, the parameters of all items in focused and referred groups are supposed to be equal; in other words, none of the items are assumed as DIF. In the AM, it is supposed that parameters of item \( i \) for the focused and referred group can differ, and for the other items, the parameters are supposed to be equal. While a likelihood function can be obtained from CM, as many likelihood functions as the number of items can be obtained from the AM. \( G^2 \) value is obtained by taking the logarithms of the likelihood function of the CM and AM (Thissen, 2001).

\[ G^2 = -2LL_C - (-2LL_A). \]

\( G^2 \) shows the chi-square distribution. The number of item parameters is the degree of independence of the distribution. In the condition that the value of \( G^2 \) exceeds 3.84 \( (G^2_{df=1, \alpha=.05}) \), the null hypothesis is denied and the presence of DIF is possible for the related item (Thissen, 2001). The quantitative value of \( G^2 \) appoints the effect degree of DIF. Taking into account Cohen’s \( G^2 \) statistics, the classification made for the degree of effect is as seen below (Greer, 2004):

- **A Level:** If \( 3.84 < G^2 < 9.4 \), a negligible level of DIF is present.
- **B Level:** If \( 9.4 < G^2 < 41.9 \), a medium level of DIF is present.
- **C Level:** If \( G^2 > 41.9 \), a magnitude level of DIF is present.

**Purpose of the Study**

The research is the study of determining systematic errors in the test items. In this study, the data display a nested structure. The aim of this research is to compare the results of the DIF determining with HGLM technique and the results of the DIF determining with LR and IRT-LR techniques on the test items. With the reason of frequently having been encountered with the nested data, the comparison of the evaluation of DIF determining stages with HGLM and the results of DIF determining with HGLM with the other methods have been given much emphasis.

For this reason, first in this research, it is determined whether the students encounter DIF with HGLM, LR, and IRT-LR techniques according to SES, in the Turkish, Social Sciences, and Science subtest items of the Secondary School Institutions Examination (SSIE) conducted in Turkey in 2006. With these methods, it was examined whether there is accordance between the items that have been designated as DIF according to the subtests.

**Method**

**Sample**

A total of 798,307 students, who took the 2006-SSIE, cover the scope of research; 6,016 students cover the sample part of the research having been chosen with randomly exemplifying technique. Subgroups are formed according to SES. In all, 2,249 (38%) of the sample students participating in the
exam were from the lower socioeconomic parts and 3,722 (62%) were from the upper socioeconomic parts. The focused group consisted of the lower socioeconomic parts and the referred group consisted of the upper socioeconomic parts.

Instrument
In this research, 2006-SSIE results have been used as data to inspect the DIF determining techniques. For this reason, there is no interpretation concerning the contents of the items that show DIF. SSIE consists of 25-item Turkish, Social Sciences, Maths, and Science subtests. It has been designated that the reliability coefficient of Maths subtest was \( \alpha = .688 \) low; according to the factor analysis technique, the test was not single dimensional and \( G^2 \) designated with IRT-LR technique has taken excessive values. For this reason, Maths test was exempted from the analysis. Cronbach’s alpha reliability coefficient for Turkish was .849, Social Sciences .873, and Science .792.

Software
The data were analyzed with HGLM, LR, IRT-LR techniques. For HGLM, LR, and IRT-LR analyses, HLM-6.04 (Raudenbush, Bryk, Cheong, & Congdon, 2001) program, script that was prepared in Zumbo’s SPSS program (Zumbo, 1999), and IRTLRDIF (Thissen, 2001) programs have been used, respectively.

Results
Is There Any Accordance Between the Items That Are DIF, According to SES With HGLM, LR, IRT-LRT Techniques in the SSIE-Turkish Subtest?

In terms of SES variable in Turkish subtest, uniform-DIF was observed in 9 items with HGLM and in 1 item with LR, and nonuniform-DIF was observed in 9 items; in total DIF was observed in 10 items of A Level. With IRT-LR, nonuniform-DIF was designated in 6 items totally, 4 of them in A Level and 2 of them in B Level. In terms of SES variable in SSIE-Turkish subtest, the presence of an accordance among the items having DIF that have been designated with HGLM, LR, and IRT-LR techniques was researched, and the results are summarized in Table 1.

When the accordance between LR and IRT-LR techniques was observed, DIF was found in five common items. The ratio of the number of designated items with DIF to the total number of items is 8%. Between the item numbers with or without DIF that were determined with these techniques, a medium-level correlation (.58) was observed (\( p < .05 \)).

When the accordance between LR and HGLM techniques was observed, DIF was found in two common items. The ratio of the number of designated items with DIF to the total number of items is 8%. Between the item numbers with or without DIF that were determined with these techniques, a lower level correlation (.25) was observed (\( p < .05 \)).

Is There Any Accordance Between the Items That Are DIF, According to Gender With HGLM, LR, IRT-LRT Techniques in the SSIE–Social Sciences Subtest?

In terms of SES variable in Social Sciences subtest, nonuniform-DIF was observed in five items with HGLM and in eight items with LR in A Level. With IRT-LR, nonuniform-DIF was designated in nine items totally, four of them in A Level and five of them in B Level. In terms of SES variable in SSIE–Social Sciences subtest, the presence of an accordance among the items having DIF that have been designated with HGLM, LR, and IRT-LR techniques was researched, and the results are summarized in Table 2.

In terms of SES variable, according to the results when the similarity between LR and IRT-LR techniques were examined in Social Sciences subtest, DIF was found in six items. The ratio of the number of designated items with DIF to the total number of items is 24%. Between the item numbers with or without DIF that were determined with these techniques, a medium-level correlation (.58) was observed (\( p < .05 \)).

When the accordance between LR and HGLM techniques was observed, DIF was found in two common items. The ratio of the number of designated items with DIF to the total number of items is 8%. Between the item numbers with or without DIF that were determined with these techniques, a lower level correlation (.09) was observed (\( p < .05 \)).

When the accordance between IRT-LR and HGLM techniques was observed, DIF was found in three common items. The ratio of the number of designated items with DIF to the total number of items is 24%. Between the item numbers with or without DIF that were determined with these techniques, a very low–level correlation (.09) was observed (\( p < .05 \)).

When the accordance between IRT-LR and HGLM techniques was observed, DIF was found in three common items. The ratio of the number of designated items with DIF to the total number of items is 24%. Between the item numbers with or without DIF that were determined with these techniques, a lower level correlation (.25) was observed (\( p < .05 \)).

DIF was designated in two items (8% of all items) with both HGLM and IRT-LR and LR techniques.

Is There Any Accordance Between the Items That are DIF, According to Gender With HGLM, LR, IRT-LRT Techniques in the SSIE–Science Subtest?

In terms of SES variable in Science subtest, nonuniform-DIF was observed in four items with HGLM and in six items with...
### Table 1. Number of DIF Items in Turkish Subtest, in Terms of SES Variable

|        | LR       | HGLM     | IRT-LR   | HGLM     | IRT-LR   |
|--------|----------|----------|----------|----------|----------|
|        | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF |
| Total  | 19       | 6        | 19       | 6        | 19       | 6          | 19       | 6          | 19       | 6          | 19       | 6          |
| Phi correlation coefficient | 0.497 |              | 0.238 |              | 0.031 |              | 0.031 |              | 0.250 |              | 0.086 |              |
| p      | .013*    |          | .234     |          | .876     | .668       | .876     | .668       | .211     | .876       | .211     | .876       |

Note: DIF = differential item functioning; LR = logistic regression; SES = socioeconomic status; IRT-LR = item response theory–likelihood ratio; HGLM = hierarchical generalized linear model.

* p < .05.

LR in A Level. With IRT-LR, nonuniform-DIF was designated in eight items totally, seven of them in A Level and one of them in B Level. In terms of SES variable in SSIE-Science subtest, the presence of an accordance among the items having DIF that have been designated with HGLM, LR, and IRT-LR techniques was researched, and the results are summarized in Table 3.

### Table 2. Number of DIF Items in Social Sciences Subtest, in Terms of SES Variable

|        | LR       | HGLM     | IRT-LR   | HGLM     | IRT-LR   |
|--------|----------|----------|----------|----------|----------|
|        | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF | No DIF   | Having DIF |
| Total  | 16       | 8        | 16       | 8        | 16       | 8          | 16       | 8          | 16       | 8          | 16       | 8          |
| Phi correlation coefficient | 0.557 |              | 0.086 |              | 0.250 |              | 0.086 |              | 0.250 |              | 0.086 |              |
| p      | .005*    |          | .668     |          | .211     | .668       | .668     | .211       | .211     | .668       | .211     | .668       |

Note: DIF = differential item functioning; LR = logistic regression; SES = socioeconomic status; IRT-LR = item response theory–likelihood ratio; HGLM = hierarchical generalized linear model.

* p < .05.
When the accordance between LR and IRT-LR techniques was observed, DIF was found in five common items. The ratio of the number of designated items with DIF to the total number of items is 20%. Between the item numbers with or without DIF that were determined with these techniques, a medium-level correlation (.62) was observed \( (p < .05) \).

When the accordance between LR and HGLM techniques was observed, DIF was found in two common items. The ratio of the number of designated items with DIF to the total number of items is 8%. Between the item numbers with or without DIF that were determined with these techniques, a lower level correlation (.27) was observed \( (p < .05) \).

When the accordance between IRT-LR and HGLM techniques was observed, DIF was found in three common items. The ratio of the number of designated items with DIF to the total number of items is 12%. Between the item numbers with or without DIF that were determined with these techniques, a lower level correlation (.40) was observed \( (p < .05) \).

When the accuracy between LR and IRT-LR techniques was observed, DIF was found in five common items. The ratio of the number of designated items with DIF to the total number of items is 20%. Between the item numbers with or without DIF that were determined with these techniques, a medium-level correlation (.62) was observed \( (p < .05) \).

DIF was designated in two items (8% of all items) with both HGLM and IRT-LR and LR techniques. The specification that was asked to be measured in the items of the Turkish, Science, and Social Sciences subtests, showing DIF in terms of SES, made a detailed research on the items necessary. The reason for the DIF displaying of the items was not part of the research scope, and as a result, this issue is not discussed. However, the observation of DIF with three different techniques in five, two, and three (by an order) items in Turkish, Social Sciences, and Science subtests, respectively, can be evaluated as an important finding.

**Summary**

DIF was determined in Turkish, Social Sciences, and Science subtests by an order, in 9, 5, and 4 items with HGLM technique; 10, 8, and 6 items with LR technique; and 6, 9, and 8 items with IRT-LR technique. Shen (1999) emphasized in his research that the determination of DIF item with HGLM is more binding in terms of LR. In this research, in all subtests, according to SES variable, the DIF item numbers determined with HGLM was found to be less than DIF item numbers determined with LR. However, the DIF determined with HGLM and LR was almost equal in Kim’s (2003) research, and Kim emphasized in her research that HGLM was more practical in terms of model flexibility.

It can be seen that the DIF showing items are greater in number in the Turkish and Social Sciences subtests than the items in Science subtest. In all subtests, although all the items determined with LR technique show a negligible level of DIF, a negligible level of DIF was found in half/more than half of the items determined with IRT-LR technique. With HGLM technique, while a low number of DIF was observed in the items in Social Sciences and Science subtests, almost half of the number of the items in Turkish subtest has DIF. It is quite significant that the most number of items with DIF determined was in the Turkish subtest.

When we inspect the correlations among the techniques in terms of determining the items having DIF, it was discovered that there is significant correlation between the results of IRT-LR and LR techniques in all subtests; merely in Science subtest, the results of the correlation between HGLM and IRT-LR techniques were found significant.
Recommendation

The quality of the items can be increased in the test banks by designating which technique is suitable for which test in determining DIF, by performing simulation to the parameters of the exams such as SSIE, which are prepared with the aim of selecting and placement. DIF examinations can be made on the test items by forming different subgroups such as school type and gender on similar data.

DIF examinations of the large-scale placement exams have to be held every year. Besides, the examinations of the items showing DIF have to be made in terms of bias and the tests have to be organized according to the results of DIF applications.

In this research, DIF analysis has been carried out on the test items by using HGLM, IRT-LR, and LR techniques. DIF applications can be made on test items with the other DIF analysis techniques that were not taken to the scope of this research. The analysis results, which were determined by using the DIF techniques in different sample sizes, can be compared.

DIF results can be compared by taking sample sizes in different ratios from the focused and referred groups.

In this research, DIF examinations have been made on the subtests consisting of 25 items. The effect of test duration to DIF inspections can be researched for the tests that consist of different number of items.

DIF analysis have been made on the binary category rated measurement results, research outcome variable (dependant variable). DIF inspections can be made on the multiple-rated measurement results by using similar techniques.

DIF determining performances with HGLM technique can be compared with the results of DIF determining techniques by experiencing on the exams serving for different aims such as Trends in International Mathematics and Science Study (TIMMS) and The Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA).

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Acar, B. (2005). Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures (Unpublished doctoral dissertation). Florida State University, Tallahassee.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. Psychological Bulletin, 101, 147-158. Retrieved from http://www.deepdyve.com/lp/psycarticles-reg/application-of-hierarchical-linear-models-to-assessing-change-eSQA99YBuU

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items (Vol. 4). Thousand Oaks, CA: SAGE.

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

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 symposium titled “Improving Large-Scale Assessment in Education” at the Annual Meeting of the Canadian Society for the Study of Education, Canada.

Gokiert, R. J., & Ricker, K. L. (2004). Gender differential item functioning on the WISC-II: Analysis of the Canadian standardization sample (Centre for research in applied measurement and evaluation). Edmonton, Canada: University of Alberta.

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.

Jodoim, G. M., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education, 14, 329-349.

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

Lord, M. F. (1980). Applications of item response theory to practical testing problems. Mahwah, NJ: Lawrence Erlbaum.

National Assessment of Educational Progress. (2006). Comparing private schools and public schools using hierarchical linear modeling. Retrieved from http://www.ncces.ed.gov/NAEP/pdf/studies/2006461.pdf

Ogretmen, T. (1995). Differential item functioning analysis of the verbal ability section of the first stage of the university entrance examination in Turkey (Unpublished doctoral dissertation). Middle East Tecnic University, Ankara, Turkey.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. Practical Assessment, Research and Evaluation, 7(1). Retrieved from http://pareonline.net/getvn.asp?v=7&n=1

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. Sociology of Education, 59, 1-17.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2001). HLM 5 Hierarchical Linear and Nonlinear Modelling. [Computer software]. Scientific Software International, Inc. USA.

Shen, L. (1999, April). A multilevel assessment of differential item functioning. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec,Canada.

Subedi, B. R. (2005). Demonstration of the three-level hierarchical generalized linear model applied to educational research (Unpublished doctoral dissertation).The Florida State University, Tallahassee.
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Retrieved from http://www.unc.edu/~dthissen/dl.html

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

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 from http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf

**Bio**

Tülin Acar is an educational measurement and evaluation specialist. The author’s research interests include hierarchical linear models, differential item functioning, psychometric properties of tests, educational statistics, and multivariate statistical analysis.