ISSN 1303-0485 • **eISSN** 2148-7561
DOI 10.12738/estp.2015.5.2505

Educational Sciences: Theory & Practice • 2015 October • 15(5) • 1229-1246

Received | April 1, 2014
Accepted | May 8, 2015
OnlineFirst | 30 October, 2015

Effect of Differential Item Functioning on Test Equating*

Kübra Atalay Kabasakal[a]
Hacettepe University

Hülya Kelecioğlu[b]
Hacettepe University

Abstract
This study examines the effect of differential item functioning (DIF) items on test equating through multilevel item response models (MIRMs) and traditional IRMs. The performances of three different equating models were investigated under 24 different simulation conditions, and the variables whose effects were examined included sample size, test length, DIF magnitude, and the test type. The MIRMs, in which the DIF factors were added as parameters, were compared with the Stocking–Lord (SL) method (one of the IRM-based calibration methods) and concurrent calibration method. According to the results, differences were found in the performances of the methods under the analyzed conditions. More specifically, the MIRMs were able to identify the DIF items, carry out the equating processes, and eliminate the biases caused by DIF in only one analysis. However, this does not indicate that using MIRMs is the best approach since the increase in sample size and test length generally had a positive effect on IRM-based equating, whereas MIRMs were less affected by these two conditions. Considering the IRM-based methods, it was found that separate calibration methods were more affected by the presence of DIF items compared to concurrent calibration. Moreover, this effect becomes most significant when DIF items are in common test and the magnitude of DIF is C.

Keywords: Test equating • Differential item functioning • Equating error • Equating bias • Multilevel item response models • Hierarchical Rasch Model

* This study is based on a brief summary of the doctoral dissertation entitled "The Effect of Differential Item Functioning on Test Equating" prepared in the Educational Measurement and Evaluation Program, Hacettepe University, Turkey.

a Corresponding author
Kübra Atalay Kabasakal (PhD), Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara Turkey
Research areas: Educational measurement; Differential item functioning; Test equating; Multilevel modeling
Email: katalay@hacettepe.edu.tr

b Prof. Hülya Kelecioğlu (PhD), Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara Turkey
Email: hulyaebb@hacettepe.edu.tr

To ensure security in large-scale and central exams, different questions are presented to students in each term or year. A lot of forms are developed for the tests applied to this end. Although the security of questions is protected by this application, problems of equality and fairness of tests emerge. More specifically, although the tests are similar in terms of content, it is possible that some individuals can take a simpler or more reliable test and become advantageous compared to others (Cook & Eignor, 1991). For example, the Foreign Language Exam (YDS) and the Academic Personnel and Postgraduate Education Entrance Exam (ALES) are held twice a year, and their scores are used for entrance to educational institutions. The same scores from different periods are considered equal without any regard to test equality, which, in turn, can lead to errors when making decisions about the students' capabilities. Thus, such exams that are repeated at regular intervals and conducted for the same purpose should be equated.

Another issue that should be considered in national and international tests is the effect of being a member of different demographic groups with the same ability levels on measurement results. In some cases, other variables can be included into the characteristics of individuals that we would like to measure. The effect of these variables on test scores can threaten the validity of the results and cause test score bias. In this case, measurement bias means systematic error against a particular group on measurement scores and differential item functioning (DIF) is an index of bias (Camili & Shephard, 1994). The reasons of DIF in national tests includes variables such as sex and school type (Bakan Kalaycıoğlu & Kelecioğlu, 2011; Gök, Kelecioğlu, & Doğan, 2010), whereas that in international tests includes translation problems, cultural differences, and differences in education programs (Le, 2009; Yıldırım & Berberoğlu, 2006).

DIF items not only act biased toward groups or individuals with certain qualities, but they may increase equating errors as well as parameter estimation errors. When DIF parameters exist, it can cause undesirable results, one of which is the error in ability parameter estimation. In addition, DIF can affect ability parameter estimation in two ways. First, DIF directly affects such estimation; second, the equating coefficients are indirectly affected because item parameter estimation is affected by DIF. Therefore, when there is DIF in test items, parameter estimation and test equating should be performed by considering both the direct and indirect effects of DIF (Han, 2008). Furthermore, it is important to identify DIF items and exclude these items from the test before test equating and parameter estimation are performed. However, many test equating studies have indicated that all items did not show DIF without conducting any type of DIF analysis (Chu, 2002).

Investigating DIF only through quantitative methods does not provide enough information about the quality and functionality of the test items. In addition, when DIF is detected in the test item with a quantitative method, an expert opinion is required to decide whether this item should remain in the test. According to Crocker and Algina (1986), if the expert decides that the item is biased, it is excluded from the test. Conversely, excluding an item from a test is undesirable since it adversely affects construct and content validity. There are findings in the literature which show that excluding DIF items from a test can result in (*1*) reduced construct validity; (*2*) reduced precision of ability parameter estimation; and/or (*3*) increased cost of test development (Chu, 2002). Thus, deleting DIF items is ideal to prevent biases in tests, and this process is not only relatively simple but also less controversial than using the information obtained from DIF items. However, if a test contains a large number of DIF items, eliminating such items can reduce both test validity and the precision of parameter estimations. For the aforementioned reasons, it is important to determine DIF items during the equating procedure as well as develop and employ methods that can minimize the effect of these items on the equating coefficients (Hidalgo-Montesinos & Lopez-Pina, 2002).

Within the scope of the present study, how test equating and parameter estimation are affected in tests containing DIF items was examined through multilevel item response models (MIRMs) and traditional IRMs. Current IRMs do not have the flexibility to control external variables such as being a member of a group in item parameter estimation (Turhan, 2006). MIRMs, also known as hierarchical models, offer opportunities to examine the effects of various variables on parameter estimation such as membership in a school, region, or group. Therefore, MIRMs can help control DIF effects and eliminate any biases that might result from DIF during test equating. Thus, the present study compares MIRMs with IRMs to determine the extent to which MIRMs can control the biases that result from DIF during item and ability parameter estimations.

Item Response Theory

Item response theory (IRT) applications are commonly used for various purposes similar to classical test theory (CTT) applications. The main purposes include test development, test equating, determining item bias, and scaling. Contrary to CTT, IRT mathematically models the relationship between an individual's ability and his/her opportunity to provide the correct answer to an item (Cook & Eignor, 1991). One of the most important properties of IRT is that ability and item difficulty are in the same scale, which ensures the invariance of item and ability parameters. In this regard, the invariance of item parameters refers to the independence of the parameters from the calibration group (Lord, 1980). In other words, item parameters do not change depending on the group that was used for their calibration. On the other hand, the invariance of ability parameters means that the items that are administered to an examinee to estimate his/her ability do not matter because the administration of different item sets will estimate the same ability scores.

IRT models used for two-category item responses include one-, two-, and three-parameter logistic models. In all three models, there is an item difficulty parameter *(b)* being the location of the logistic curve along the ability scale (θ). In the one- or two-parameter logistic models, this point (b_i) shows that an examiner has a 50% chance of correctly responding to an item (Hambleton & Swaminathan, 1985). Theoretically, parameter b can take values between $-\infty$ and $+\infty$. However, practically, it generally takes a value between −3 and +3 because the scale is scaled with 0 mean value and 1 standard deviation. A high b value indicates that the item is difficult, whereas a low b value indicates that it is easy (Harris, 1989). In the present study, the one-parameter logistic model, also known as the Rasch (1966) model, was employed.

In this model, all items are assumed to have equal discrimination, and the guessing behaviors of the examinees are not parameterized (Crocker & Algina, 1986). In the model, the probability of a randomly selected examinee (with ability level θ) giving a correct answer for item i can be expressed as follows:

$$P_{ij}(y_i = 1 \,|\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

where θ_j is examinee j's ability level, b_i is the difficulty of item i, and $P_{ij}(y_i=1|\theta_j)$ is the probability that examinee j (with ability level θ) answers item i correctly.

Multilevel Item Response Models

MIRMs merge hierarchical linear models (Raudenbush & Bryk, 2002) with IRMs. The interest toward multilevel models increased with the Rasch model, which was reformulated by Kamata (1998; 2001). The main advantage of MIRMs is their ability to arrange hierarchical data structures. Such models are especially effective in educational measurements because of the nested data structure. For example, the students are nested in classrooms, the classrooms are nested in schools, the schools are nested in cities, the cities are nested in regions, the regions are nested in countries, and so on. In this regard, the application of one-level models to multilevel data leads to both statistical and conceptual problems (Kreft & Leeuw, 1998 as cited in Pastor, 2003). Another advantage of MIRMs is their ability to add external variables (e.g., sex, level, and so on) into the applications and offer a more flexible and comprehensible model that explains the relationship between the probability of giving a correct answer to an item for an individual and individual's ability. Thus, MIRMs can estimate item and ability parameters in a similar manner to IRMs. In addition, MIRMs can examine the effects of external variables in parameter estimation (Turhan, 2006), and concurrent estimations made by these models (in a single calibration) can remove errors that might result from separate estimations.

Kamata (1998) evaluated the hierarchical Rasch model and found that it was sufficiently sensitive to detect DIF. Following this study, MIRMs were used in DIF studies (Atar, 2007; Binici, 2007; Cho & Cohen, 2010; Luppescu, 2002); test equating (Chu, 2002; Chu & Kamata, 2000, 2005; Park, Kang, & Wollack, 2007), and determining dimensionality (Beretvas & Williams, 2002). Chu and Kamata (2000) investigated the equating property of the hierarchical Rasch model in common-item non-equivalent group design and found that the model performed similarly to single-group concurrent equating. This similar performance of the hierarchical Rasch model supported the development of hierarchical Rasch equating models. Their study revealed that MIRMs can be considered as a usable method to explore possible problems of large-scale assessment programs and external variables related to academic success (Turhan, 2006).

Kamata's Hierarchical Model: 1PL-IRM

Kamata (1998) suggested the multiple-group model through hierarchical linear models (Bryk & Raudenbush, 1992). This reformulated model

was a special form of one-parameter hierarchical generalized linear model (Chu, 2002) that merged logistic regression with multilevel data structures using Bernoulli sampling and linking function. In this model, items are considered as Level-1 units, whereas examinees are Level-2 units (Kamata, 1998). The model also uses the following logit link function that connects probabilities among latent variables:

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$$

where η_{ij} is the Level-1 structural model and p_{ij} is the probability of answering item i correctly by examinee j. When the probability of answering item i correctly by examinee j is equal to 0.5, the logit is zero. In addition, when this probability is less than 0.5, the logit is a negative value, whereas when it is larger than 0.5, the logit is a positive value.

The Level-1 structural model is the item-level model, which is formulated as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij},$$

$$= \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \ldots + \beta_{(k-1)j} X_{(k-1)ij},$$

$$= \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij}.$$

where β_{0j} is the intercept term in the model and X_{qij} is the q. dummy variable of item i for examinee j (when q = i, the value is 1, and when q ≠ i, the value is 0). Dummy variable coding results in a design matrix, in which all diagonal elements are equal to 1. The unit matrix is obtained by the exclusion of the final item in dummy coding. In this case, β_{0j} is interpreted as the difficulty of the reference item excluded from the model or the average effect of all items for examinee j. β_{qj} is the coefficient X_{qij} from i = 1 to (k – 1); when q = i, it is interpreted as the effect of item i. As a result, the Level-1 (item-level) model can be written as follows:

$$\eta_{ij} = \beta_{0j} + \beta_{qj}.$$

Because the Level-1 model is a structural model, there is no error term. In this case, the coefficients in the Level-1 model offer a better understanding of those in the Level-2 (person-level) model. The person-level equations are as follows:

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

$$\beta_{1j} = \gamma_{10},$$

$$\vdots$$

$$\beta_{(k-1)j} = \gamma_{(k-1)0}.$$

β_{0j} (intercept coefficient) is assumed to be a random effect across persons. β_{0j} is the intercept coefficient consists of one fixed component and one random component. The random component γ_{00} is the average value of the overall item effect across all examinees in the sample. The random component u_{0j} is interpreted as the ability of examinee j and is assumed to be distributed normally with a mean of zero and variance of τ.

When Level-1 and Level-2 models are combined, the probability of answering item i correctly is as follows. Where i = q for examinee j and item i,

$$p_{ij} = \frac{1}{1 + \exp\{-[u_{0j} - (\gamma_{q0} - \gamma_{00})]\}}$$

Kamata (1998) showed that this model is equivalent to the Rasch (1966) model as follows:

$$P_{ij}(y_i = 1 \mid \theta_j) = \frac{1}{1 + \exp\left[-(\theta_j - b_i)\right]}$$

In this equation, the ability of a person θ_j is equivalent to u_{oj}, whereas item difficulty parameter b_i is equivalent to $(\gamma_{q0} - \gamma_{00})$.

Differential Item Functioning

To detect bias in an item, initially, the item must contain DIF, which refers to differentiations between the correct answering probabilities of examinees in different groups to the related item in a comparison to be made on ability level which the item intended to measure (Zumbo, 1999). In general, two types of DIF can occur: uniform and non-uniform DIF. Uniform DIF emerges when there is no interaction between ability level and group membership during item performance. This is when an item containing DIF favors a group in all ability levels, and only the item difficulty parameter differs among the group. Non-uniform

DIF emerges when there is an interaction between ability level and group membership during item performance (Camili & Shephard, 1994).

The selection of proper statistics is extremely important in the analysis of nested data in educational measurements. When the data includes a nested or clustered structure, the results will contain errors because the hypothesis regarding the independence of observations has been violated using a classical linear model. Bryk and Raudenbush (1992) emphasized that MIRMs have less measurement errors compared with traditional models for nested data. This feature makes MIRMs advantageous compared with traditional methods of DIF detection. For example, if the source of DIF is both region and sex, then MIRMs can determine sex on the person level and region on the group level. Another advantage of MIRMs is their ability to simultaneously determine different DIF factors in different magnitudes. For example, when the 1st and 2nd items contain sex DIF, the 3rd and 4th items may include race DIF or the same item may contain both DIF factors. DIF magnitudes in these four items can differ from one another. In sum, there are no limitations of DIF factors that can be added to the model in MIRMs (Chu, 2002).

1PL-MIRM and DIF (Kamata's Hierarchical Rasch DIF Model)

Kamata (1998) developed a DIF model by adding group indicator variables into the Level-2 1PL-MIRM and showed that the model was sufficiently sensitive to detect DIF. The equations of this Level-2 model to which DIF parameters were added are as follows:

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

$$\beta_{1j} = \gamma_{10} + \gamma_{11}G_j,$$

$$\vdots$$

$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}G_j.$$

G_j is the group indicator for two-category items, and the values from γ_{11} to $\gamma_{(k-1)}$ are the DIF magnitude of corresponding items. In this case, the group indicator can be coded as "1" for the focal group and "0" for the reference group or vice versa. Significant shows that being a member of a distinctive group affects the probability of a correct response for item 1. In this model, equality of item difficulty for the focal group, in other words; uniform DIF is tested.

Luppescu (2002) compared Karmata's hierarchical Rasch DIF model and the classical *b* parameter difference method for DIF detection. It was found that in general, lower root mean square error (RMSE) values were obtained in Kamata's hierarchical Rasch model. The reasons of this case were found as the fact that classical method required two different calibrations. In addition, the hierarchical method required only one calibration with less estimated parameters.

Test Equating

Generally, in large-scale exams, different test forms that are similar in terms of content and difficulty levels are used. However, these forms are not entirely equivalent regardless of such similarities. The statistical method used to place two or more tests into a common scale is referred to as "test equating," and the results can be used interchangeably (Kolen & Brennan, 2004). Following a successful test equating procedure, examinees are expected to obtain the same scores regardless of the test form that was applied (Holland & Dorans, 2006; Kolen & Brennan, 2004).

Equating methods are most commonly classified as CTT- and IRT-based equating models. The first procedure in IRT-based equating is to ensure model–data fit, after which a suitable equating design is selected. Thereafter, the equating method is selected, and the item and ability parameters are estimated. In the third stage, the item and ability parameters are placed on a common scale. In equating designs (single and random groups) in which parameter estimation is performed with a single calibration, the parameters automatically take place on the same scale. Thus, the third stage is not required, and the scale to report the test scores is selected in the fourth stage.

There are generally three types of data collection designs used in IRT-based equating: single group, random group, and common item. In the common-item design (which is used in the present study), there are two test forms and two different groups. Although there are two test forms, both forms include a number of common items that are used to reveal the equating relationship between the two groups by comparing their performances. When common items are appropriately selected,

problems in the single group and equivalent group design are alleviated. In other words, neither the examinees are required to take both test forms, nor the examinee groups are to be equivalent (Hambleton, Swaminathan, & Rogers, 1991; Holland & Dorans, 2006). In this design, common items play a significant role in determining the equating function. In this regard, properties of common items should be considered in equating studies. Angoff (1971) suggested that all common items should be a mini version (representative) of the entire test in terms of construct, item type, content, and so on. Both properties and the number of common items are important for common-item design. Hambleton et al. (1991) stated that the number of items required for common items must be 20%–25% of the item numbers in the test. Finally, research has shown that the increase in common-item numbers actually decreases equating errors (Kolen & Brennan, 2004).

IRT-based Equating Methods

In IRT-based equating, the ability levels of the examinees are not affected by the items, and item parameters are not affected by the group of examinees (Hambleton et al., 1991). The parameter invariance property of IRT models is one of the main advantages of this model. In addition, having item and ability parameters on the same scale across test forms provides score comparability. However, in practice, these parameters are estimated through various techniques because true parameters are unknown. Many computer programs used to estimate item and ability parameters standardize examinees' ability distribution to a mean of zero and unit variance by default (Baker & Al-Karni, 1991). Thus, item and ability parameters estimated from different tests may not be on the same metric because the standardization of ability distribution will also affect parameter estimation. For this reason, estimations on one test form must be transformed into estimations on another test form. As discussed in the following sections, there are two approaches to transform item parameters (estimated from different groups) into the same scale: separate and concurrent calibrations (Kolen & Brennan, 1995, 2004).

Separate Calibration IRM-SC: When equating is performed through the single group or equivalent group design, extra scaling is not required because test forms are already on the same scale. In an equating method based on common-item design, parameters' estimates from different test forms may not be on the same scale because the groups are different. Therefore, linear transformation should be performed to place the two test forms on the same scale (Kolen & Brennan, 2004). This transformation includes three steps: (*1*) estimate item parameters on scale X (test form X) and scale Y (test form Y); (*2*) determine equating slope (*A*) and equating intercept (*B*); and (*3*) transform the ability estimates (based on the *A* and *B* equating coefficients) from scale X to scale Y (Kolen & Brennan, 1995, 2004).

The relationships between the estimated abilities from the two different test forms can be defined as

$$\theta_{Yi} = A\theta_{Xi} + \mathrm{B},$$

where *A* and *B* represent the equating coefficients and θ_{Xi} and θ_{Yi} represent the ability estimation of person *i*. Similarly, the item parameters of the two tests are transformed, and their relationships are as follows:

$$a_{Yj} = \frac{a_{Xj}}{A},$$

$$b_{Yj} = Ab_{Xj} + \mathrm{B},$$

$$c_{Yj} = c_{Xj},$$

where b_{Yj}, a_{Yj}, and c_{Yj} are the item parameters of item *j* on form Y, whereas b_{Xj}, a_{Xj}, and c_{Xj} represent item parameters of item *j* on form *X* (Kolen & Brennan, 1995).

In separate calibration, one of the mean–mean, mean–sigma, and characteristic curve transformation methods is used. Mean–mean and mean–sigma methods are based on the transformation of item and ability parameters using common items, whereas characteristic curve methods are based on reducing the gap between the item or test characteristic curves of common items. Research has revealed that characteristic curve methods are better than mean–mean and mean–sigma methods and tend to produce more stable results (Baker & Al-Karni, 1991; Gök, 2012; Stocking & Lord, 1983).

Moreover, mean–mean and mean–sigma methods can cause erroneous results for items that include similar item characteristic curves but different parameters. In addition, because *A* and *B* equating coefficients are calculated using descriptive statistics of *b* parameter (or both *a* and *b* parameters) in mean–

mean and mean–sigma methods, the use of three-parameter data causes problems (Han, 2008). As a solution to this problem, Haebara (1980) suggested a method that considers all item parameters at the same time. Thereafter, Stocking and Lord (1983) developed another characteristic curve method that calculates the lost function as follows:

$$L(\theta_i) = \left[\sum_{j=1}^{m} p_{ij}\left(\theta_{i,}, a_{1Lj}, b_{1Lj}, c_{1Lj}\right) - \sum_{j=1}^{m} p_{ij}\left(\theta_{i,}, a^{*}_{2Lj}, b^{*}_{2Lj}, c_{2Lj}\right)\right]^2.$$

The lost function used by Stocking and Lord (1983) is the square of the total difference between the item characteristic curves of each item for examinees at certain ability level. Characteristic curve transformation methods developed to reduce the difference between items or test characteristic curves of common items generally offer similar estimations and provide better results compared with separate calibration methods, especially in the transformation of item discrimination parameters. Research conducted in this field revealed that characteristic curve methods are better than mean-mean and mean-sigma methods and they tend to produce more stable results (Baker & Al-Karni, 1991; Gök, 2012; Stocking & Lord, 1983).

Concurrent Calibration IRM-CC: Another method used to place items on the same scale is concurrent calibration. Concurrent calibration simultaneously estimates item parameters for two test forms. It is assumed that common items have the same item parameters in both test forms. Because differences in ability distribution are considered, the estimated item parameters take place on the same scale (Turhan, 2006). Thus, an extra transformation is not required to obtain *A* and *B* constants. Hanson and Beguin (2002) revealed that concurrent calibration provides more accurate results than separate calibration based on the condition that parametric model assumptions are met.

The Hierarchical Rasch Model as a Concurrent Equating Model: The addition of individual variables into the hierarchical Rasch model allows the model to have two levels while the addition of group variables makes it have three levels. The two-level hierarchical Rasch model containing individual variables is used for horizontal equating, whereas the three-level model containing group variables are used for vertical equating. In the two-level hierarchical Rasch model, The Level-1 model (item-level model) is as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij},$$

$$= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{(k-1)j}X_{(k-1)ij},$$

$$= \beta_{0j} + \sum_{q=1}^{k-1}\beta_{qj}X_{qij}.$$

In this study, person-level DIF was assumed as the focal and reference group and was added into the Level-2 model. When it is assumed that first and second items have DIF in the focal group and the other items do not, the Level-2 equations are as follows:

$$\beta_{0j} = \gamma_{00} + \mu_{0j},$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Focal)_{j},$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(Focal)_{j},$$

$$\vdots$$

$$\beta_{(k-1)j} = \gamma_{(k-1)0}.$$

In equations, γ_{00} to $\gamma_{(k-1)}$ are the intercept coefficients of the terms from β_{0j} to $\beta_{(k-1)j}$. In addition, γ_{11} and γ_{21} are the focal group coefficients for Item 1 and Item 2. The items without DIF only include intercept (γ_{i0}) since the item effect is constant across persons. The DIF variable was added into Item 1 and Item 2 to adjust for focal group effects. In this study, the reference group was coded as 0, and the focal group was coded as 1. Substituting Level-2 γ parameters to β_{ij} provides the following equations:

$$p_{ij} = \frac{1}{1+\exp[-[(\gamma_{00}+\mu_{0j})+\gamma_{i0}]]},$$

$$= \frac{1}{1+\exp[-[\mu_{0j}-(-(\gamma_{00}+\gamma_{i0}))]]},$$

or

$$p_{ij} = \frac{1}{1+\exp[-[(\gamma_{00}+\mu_{0j})+(\gamma_{i0}+\gamma_{i1})]]},$$

$$= \frac{1}{1+\exp[-[\mu_{0j}-(-(\gamma_{00}+\gamma_{i0})-\gamma_{i1})]]}.$$

The first two equations are for the items without DIF and thus, they do not have the term $-\gamma_{i1}$. The last two equations are for DIF items. The error term of β_{0j}, μ_{0j} is the random error for person j. In the hierarchical Rasch model, this random effect is treated as ability parameter estimates and can be seen in the aforementioned equations. Item difficulty is $-(\gamma_{00} + \gamma_{i0})$ and DIF effect size is $-\gamma_{i1}$. For the DIF items in the reference group, item difficulty is $-(\gamma_{00} + \gamma_{i0})$ since $-\gamma_{i1} = 0$. For the DIF items in the focal group, item difficulty is $[-(\gamma_{00} + \gamma_{i0}) - \gamma_{i1}]$. As stated earlier, the last item was used as the reference item. Therefore, item parameters from 1 to (k – 1) need to be adjusted by the reference item parameter, γ_{00} (Chu, 2002).

Level-1 coefficients with j are the subscripts of persons from β_{0j} to $\beta_{(k-1)j}$. Here, j indicates that different persons are associated with different item-level parameters. When β_{ij}'s are substituted in a higher level, the subscript j is dropped and the item parameters remain constant across persons (Chu, 2002).

The Aim and Significance of the Research

Measurement results obtained through applications ranging from intra-class tests to international large-scale assessments have significant important for students, teachers, families, and politicians. In addition, they provide important information through which countries can develop their respective education systems. Especially, educational studies, such as those by the Program for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS), allow countries to not only monitor their own education systems year-by-year, but they can also compare their systems with other countries. Therefore, the accuracy of the information obtained from these studies will have an effect on subsequent decisions related to testing and assessments.

The proliferation of large-scale assessments has led to the use of different test forms for individuals in the same levels and the application of similar tests for groups with different features. Therefore, it is important to determine the relationship between the scores obtained from different test forms and to make transformations based on the findings. The statistical process that determines this relationship is test equating. However, the test equating process is affected by several factors, one of which being that, if the items show DIF, then it will adversely affect parameter estimation and the test equating process.

In this study, one of the goals was to use the 2 level 1PL-IRM with DIF knowledge for parameter estimation and equating and compare this method with the traditional 1PL-IRM. MIRMs can eliminate the bias resulting from DIF by functioning as model parameters against DIF factors. This can also improve test equating and parameter estimation performances (Chu, 2002; Kamata, 2000; Turhan, 2006). Moreover, this study investigates the results of test equating and parameter estimation using MIRMs and IRMs in case of DIF.

There have been numerous studies on DIF and test equating in the literature. However, the number of studies that have investigated both issues is extremely limited (Chu, 2002; Chu & Kamata, 2005; Han, 2008; Turhan, 2006). Upon closer examination of the existing studies, it was found that the time-consuming MIRM process does not consider different simulation conditions such as large samples, different test lengths, etc. In addition, the considered simulation conditions were used with a limited number of replications (between 5 and 20). Unlike other studies, the present study considers large samples and different test lengths as well as performs 50 replications. Furthermore, it was found that no comparison was made with MIRMs during test equating in the presence of DIF items, based on the fact that the separate calibration method does not provide better results than the concurrent calibration method. Therefore, Stocking–Lord (SL) method, which is one of the separate calibration characteristic curve methods, was used in the present study within the analyzed conditions.

Finally, this study compares IRMs and MIRMs in which DIF items are included in the test forms under different conditions. Since MIRMs are relatively new and developing, it is believed that revealing their strengths and weaknesses through simulation studies is important. In addition, multilevel modeling applications in educational tests are expected to improve and resolve problems related to test development (DIF, test equating, etc.), despite the fact that the number of applications related to multilevel models is limited. Therefore, the present study, it is believed that the present study can be a guide about the use of MIRMs being an alternative to traditional IRMs and this model can be an alternative to current equating methods.

Method

Type of Research

This study compares the efficiency of equating methods in the presence of DIF under different

conditions. To this end, data containing DIF items was generated under certain conditions to determine the method that causes the least errors. Moreover, under different conditions, equating methods were compared with simulation data under controlled conditions. In this sense, this study contributes to the theory and serves as fundamental research (Karasar, 2009).

Simulation Conditions

This study examined the effects of the following conditions on equating errors: sample size, test length, DIF magnitude, and test type containing DIF items. Common-item design in equivalent groups was used to equate the simulated data.

Sample Size: For each test form (with the goal of 500 examinees for a small sample and 2,000 examinees for a large sample), two different sample sizes containing a total of 1,000 and 4,000 examinees, respectively, were analyzed in this study. Although it has been reported that a minimum sample size of 500 is required for successful equating (Spence, 1996), it has been shown that methods provide better results if the sample size is 1,000 or larger (Han, 2008).

Test Length: In this study, the item number of the test was considered on two levels: 20 for a short test and 40 for a long test. The common-item number was determined to be 25% for both conditions. Angoff (1971) suggested that a minimum of 20% of the item number of the whole test should be common item.

DIF Magnitude: Since DIF creates a variance in parameter estimation, as DIF magnitude increases, the stability of ability estimation and equating decreases (Chu, 2002). In the present study, DIF magnitude was analyzed on two levels (B and C), and the difference between the parameters for magnitude B was determined as 0.6 while it was determined as 1 for magnitude C.

Test Type containing DIF Items: DIF items were placed in the test in the following three ways to investigate the effect of their placement on equating: *(1)* DIF on a common test; *(2)* DIF on a non-common test; and *(3)* DIF on both types of tests.

The four simulation factors examined in this study and the 24 conditions of these factors are presented in Table 1.

Table 1
Conditions Considered in Equating

Equating Conditions					
Test Length	Test Type containing DIF Items	Sample Size			
		500-500		2000-2000	
		DIF Magnitude			
		B	C	B	C
20	NCT	S1	S4	S13	S16
	BT	S2	S5	S14	S17
	CT	S3	S6	S15	S18
40	NCT	S7	S10	S19	S22
	BT	S8	S11	S20	S23
	CT	S9	S12	S21	S24

**S: Simulation; NCT: Non-common test; BT: Both tests; CT: Common test.*

Data Generation

The two-category item responses in the study were generated using WinGen 3.0 software (Han, 2007). The data generation procedure was conducted by the following three steps:

1. Ability Parameters: Ability distribution was sampled from standard normal distribution ($\theta \sim N(0,1)$) for each group. A total of four sets of ability parameters were generated for the focal and reference groups under the conditions of both sample sizes. The sizes of the focal and reference groups were equally formed.

2. Item Parameters: Two test forms were generated for equating. Both tests contained specific items and common items, and three item sets were generated. Zimowski, Muraki, Mislevy and Bock (1999) suggested that common items should have high discrimination and medium difficulty. Thus, *b* parameters of common items were selected between −1 and +1. The variance of item difficulty among test forms can affect DIF. For this reason, similar item difficulties were selected for both test forms, and the item difficulties in non-common items differed between −2.5 and +2.5.

Table 2
Numbers of DIF Items

Location of DIF Items		Simulation Conditions	Item Number
NCT	N = 20	S1, S4, S13, S16	19, 20
	N = 40	S7, S10, S19, S22	37, 38, 39, 40
BT	N = 20	S2, S5, S14, S17	1, 20
	N = 40	S8, S11, S20, S23	1, 2, 39, 40
CT	N = 20	S3, S6, S15, S18	1, 2
	N = 40	S9, S12, S21, S24	1, 2, 3, 4

**S: Simulation; NCT: Non-common test; BT: Both tests; CT: Common test.*

Table 3
Item Difficulties for the 20-Item Form

Item No.*	Common Test	Item No.	Form X	Form Y	Item No.	Form X	Form Y	Item No.	Form X	Form Y
1	0	6	−2.5	−2.45	11	−1	−1.05	16	−1.5	−1.55
2	0.5	7	2	2.05	12	2.5	2.45	17	0.75	0.8
3	−1	8	−0.75	−0.8	13	−0.25	−0.3	18	−0.5	−0.55
4	−0.5	9	1	1.05	14	1.5	1.55	19	0	0.05
5	1	10	−2	−2.05	15	0.5	0.55	20	0.25	0.3

**1–5 numbered items are common item.*

In this study, *b* parameters were created to favor the focal group during the generation of DIF item parameters in Form X. The numbers of DIF items are shown in Table 2.

The item parameters of the three item sets (Form X, Form Y, and common test) are presented in Table 3 and Table 4 for the 20-item form and the 40-item form, respectively.

3. Generation of Item Responses: This analysis required two different tests for equating design and different groups for taking the same test. Thus, two different tests with common items (Form X and Form Y) were administered to the focal and reference groups. The data was arranged so that half of the focal group had Form X and the other half had Form Y. For the reference group, half had Form X and the other half had Form Y.

Both groups had the same ability level and normal distribution. Group level (Level-3) was not considered since the groups were at the same ability level. The focal and reference groups were selected as the person-level variables. The item parameters were determined for each form while the ability parameters were determined for each group. Afterwards, the data generation procedure was conducted according to the 1PL model.

Implementation of Equating Procedure

In this study, the performances of three different equating methods were investigated under 24 different simulation conditions (two different sample sizes, two different test lengths, three different tests containing DIF items, and two different DIF magnitudes). To understand the potential of MIRMs in which DIF factors were added into the model as parameters, the MIRM was compared with the IRM-based concurrent calibration and the SL method, which is one of the separate calibration methods.

To conduct estimations, the following programs were employed: HLM for Windows 6.8 for the MIRM (Raudenbush, Bryk, Cheong, & Congdon, 2005); BILOG-MG for concurrent equating (Zimowski, Muraki, Mislevy, & Bock, 2003); and PARSCALE 4.1 for the SL method (Muraki & Bock, 2003). Then, the IRTEQ (Han, 2009) program was used to place them on the same scale. In this study, to analyze all 50 datasets generated for each condition, these four programs were operated using R software with batch files.

Evaluation Criteria and Data Analysis

Two different equating errors were calculated for item and ability parameters across replications for the purpose of investigating the stability of

Table 4
Item Difficulties for the 40-Item Form

Item No.*	Common test	Item No.	Form X	Form Y	Item No.	Form X	Form Y	Item No.	Form X	Form Y
1	−0.3	11	−0.3	−0.35	21	0.6	0.65	31	−1.9	−1.95
2	−0.15	12	0.4	0.45	22	0.9	0.95	32	0.8	0.85
3	0.15	13	1.6	1.65	23	1.9	1.95	33	1.3	1.35
4	0.3	14	−0.9	−0.95	24	−1.3	−1.35	34	1	1.05
5	−1	15	2.5	2.55	25	−1	−1.05	35	−0.7	−0.75
6	−0.5	16	2.2	2.25	26	−0.6	−0.65	36	−2.5	−2.55
7	0.7	17	−0.8	−0.85	27	0.5	0.55	37	−0.2	−0.25
8	0.5	18	−1.6	−1.65	28	0.7	0.75	38	−0.1	−0.15
9	1	19	−0.4	−0.45	29	0.3	0.35	39	0.1	0.15
10	−0.7	20	−2.2	−2.25	30	−0.5	−0.55	40	0.2	0.25

**1–10 numbered items are common item.*

parameter estimation: RMSE (equating error) and BIAS (equating bias). In this study, the square of BIAS was calculated to prevent any confusion that may arise from the interpretation of negative and positive BIAS.

First, RMSE and BIAS values in all item and ability parameters were computed for each simulation condition through the following formulas (only the means of RMSE and BIAS were reported to obtain brief conclusions from the obtained values):

$$\text{RMSE}\ (\hat{\tau}_j) = \sqrt{\frac{1}{R}\sum_{j=1}^{r}(\hat{\tau}_{jr} - \tau_j)^2}.$$

$$\text{Squared BIAS}\ (\hat{\tau}_j) = (\frac{1}{R}\sum_{i=1}^{r}\hat{\tau}_{jr} - \tau_j)^2,$$

where τ_j is the true value of parameter j, $\hat{\tau}_{jr}$ is the estimate value of parameter j among replications (r = 1, 2, …R), and R is the number of replications.

The mean RMSE for the estimation of ability parameter is $\overline{X}_{RMSE} = \frac{1}{J}\sum_{j=1}^{J}\overline{X}RMSE(\hat{\theta}_j)$, where J is the total number of examinees. The mean RMSE for the estimation of item parameter is $\overline{X}_{RMSE} = \frac{1}{J}\sum_{j=1}^{J}\overline{X}RMSE(\hat{b}_j)$, where J is the total number of items. The mean BIAS values were computed similarly.

In this study, first the RMSE calculated for each parameter value was translated into graphics by selecting an example among the different sample sizes and test lengths. Furthermore, the RMSE and BIAS values of the mean item and ability parameters were presented in graphics for each condition. In these graphics, the Level-2 1PL-IRM was used with the MIRM; traditional single-group concurrent equating was used with the IRM-CC, and the SL was used with IRM-SC abbreviations.

Variance analysis was conducted to test the effects of the conditions analyzed in this study (i.e., sample size, test length, test type containing DIF, and DIF magnitude) on the equating methods. In this case, when a model was built with conditions and all related interactions, analysis could not be performed since the error degrees of freedom was zero. Therefore, only the main effects were examined through two-way interactions. In addition, since too many significance tests were performed, the Bonferroni correction was used to control the error (significance level .002). Furthermore, the eta-squared value was reported to illustrate the effects of the variables on the methods, and post-hoc analyses were conducted to test the significant ANOVA results.

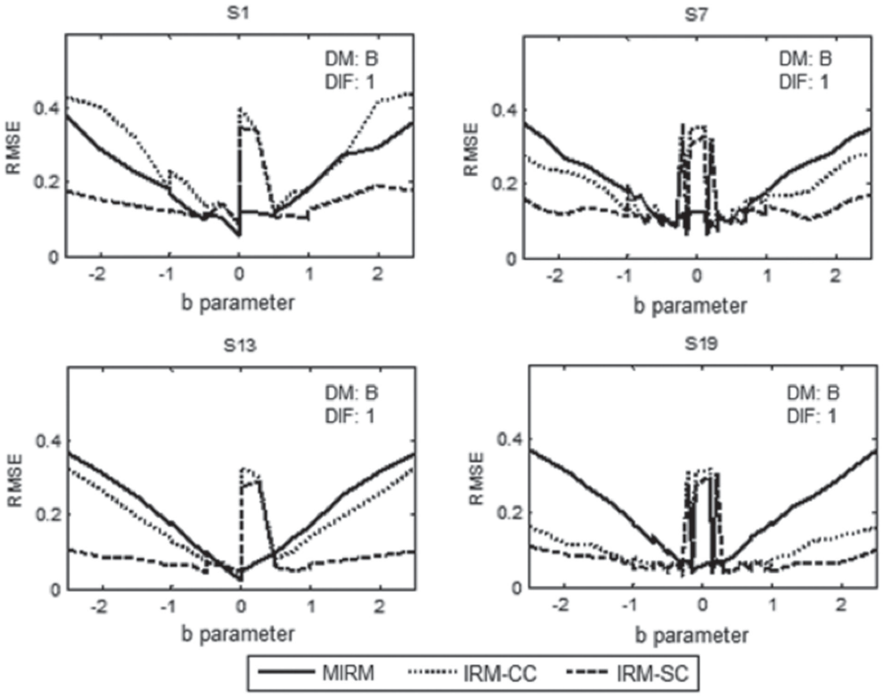

Figure 1: Equating errors of *b* parameter values in S1, S7, S13, and S19 conditions.
*DM: DIF Magnitude; DIF: 1; DIF items are in the non-common test.

Findings

When tests containing DIF items were equated with the MIRM, IRM-CC, and IRM-SC methods, the error values of *b* parameters were calculated. Figure 1 presents the RMSE values of the small sample short test (S1), the small sample long test (S7), the large sample short test (S13), and the large sample long test (S19).

When the equating errors of the item parameters in Figure 1 were broadly examined, it was found that errors increased in extreme values of the *b* parameter for all three methods, and the method that was least affected by the extreme values was the IRM-SC. According to the graphics of the four conditions, as sample size and test length increased, whereas the errors in extreme values of the *b* parameter in the IRM-based methods decreased. Meanwhile, the MIRM was not affected by these conditions.

In Figure 1, to compare the equating errors of the items with and without DIF, items having the same item parameter in both item types are compared. Since DIF items have medium difficulty, a sudden increase in the graphics of the item parameters with medium difficulty indicates DIF items. The lack of this increase in the MIRM and its presence in IRM-based methods is an indicator that the MIRM method is not affected by DIF items. In addition, the MIRM decreased equating bias and standard error by eliminating DIF variance and as a result, total error also decreased in the RMSE. This finding is consistent with the results obtained from the research conducted by Chu (2002).

When the tests containing DIF items were equated through the MIRM, IRM-CC, and IRM-SC methods, the mean values of the errors in the *b* parameter were calculated. Variance analyses were conducted separately for each of the two error values to determine the effects of the item parameters on equating errors. Significant ANOVA results for error values of the item parameters are presented in Table 5. To obtain a clearer understanding about how significant conditions obtained from the ANOVA results change, graphics of the mean values of errors for the small sample and the large sample are presented in Figure 2 and Figure 3, respectively.

Based on the findings regarding RSME in Table 5, it can be inferred that the main effects of sample size and test length were significant in each of three methods. However, the effects of test type containing DIF and DIF magnitudes were only significant in the IRM-based methods, while no difference was found in the MIRM. Regarding the interaction effects, sample size and test length were significant in the IRM-CC, while sample size and DIF magnitude interaction of test type containing DIF were significant in the IRM-SC method.

According to the results concerning BIAS in Table 5, the main effects of sample size and test length were significant in the MIRM and the IRM-based concurrent equating methods, whereas they were non-significant in the IRM-SC method. In addition, the effects of test type containing DIF and DIF magnitude were significant in the IRM-based methods, and non-significant in the MIRM. Regarding the interaction effects, sample size and test length were significant

Table 5
Significant ANOVA Results for Error Values of the Item Parameters

			Methods					
			IRM-CC		IRM-SC		MIRM	
Errors	Effects	df	F	η^2	F	η^2	F	η^2
RMSE	Sample Size (SS)	1	4705.02*	.99	2151.21*	.99	381.403*	.98
	Test Length (TL)	1	4645.33*	.99	53.122*	.85	515.521*	.98
	Test Type containing DIF Items (TT)	2	19.190*	.81	1650.31*	.99	-	-
	DIF Magnitude (DM)	1	99.307*	.92	1606.93*	.99	-	-
	TT * DM	2	-	-	304.704*	.98	-	-
	SS * TT	2	-	-	15.606*	.78	-	-
	SS * TL	1	23.751*	.73	-	-	-	-
BIAS	Sample Size (SS)	1	1188.37*	.99	-	-	586.973*	.98
	Test Length (TL)	1	3565.51*	.99	-	-	553.438*	.98
	Test Type containing DIF Items (TT)	2	59.968*	.93	175.727*	.97	-	-
	DIF Magnitude (DM)	1	175.168*	.95	657.468*	.99	-	-
	TT * DM	2	-	-	56.922*	.93	-	-
	SS * TL	1	266.116*	.97	-	-	-	-

*$p < .002$.

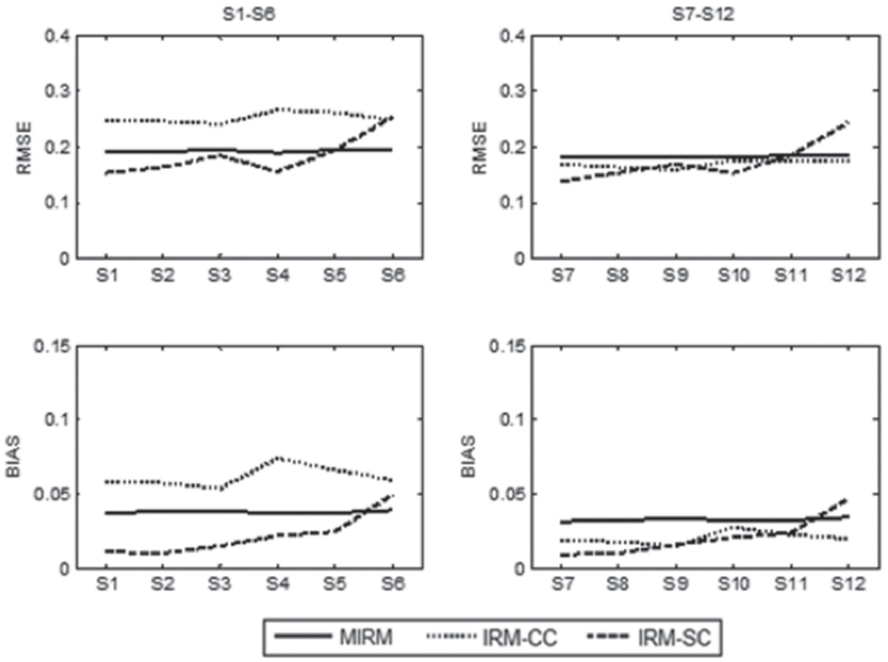

Figure 2: Mean of the item parameters' errors in the small sample.

in the IRM-CC while the interaction between DIF magnitude and test type containing DIF was significant in the IRM-SC method.

In regard to the effect of sample size on the error, a comparison of the graphics in Figure 2 and Figure 3 shows that an increase in the sample size in the MIRM led to a 0.10 decrease in RMSE values and a 0.10 increase in BIAS values. Conversely, in the IRM-CC method, both error types apparently decreased when sample size increased. In the IRM-SC method, while the increase in sample size led to a decrease in RMSE values, it did not change the BIAS values. These results presented in Figure 2 and Figure 3 are parallel to the ANOVA results. Thus, it is clear that the interaction effect of sample size and test length was significant in the IRM-CC method. It is clearly seen in the graphics that the interaction effect of sample size and test length found significant in IRM-CC method in ANOVA is the method in which the error reduced the most as the sample size and test length increased.

After comparing the 20-item and 40-item conditions in Figure 2 and Figure 3, it is shown that the increase in item number leads to a small decline in MIRM values for each of the two error values. In addition, the second method that was least affected from the increase in item number was the IRM-SC method. Furthermore, the BIAS value was not affected by test length in the IRM-SC method, the method in which errors decreased the most as item number increased was the IRM-CC.

The mean RMSE and BIAS values were investigated to understand how the error affected the test type containing DIF and DIF magnitude under the conditions given in Figure 2 and Figure 3. Accordingly, the MIRM method was not affected from either of the two conditions. Meanwhile, in the IRM-based methods, errors increased as DIF magnitude increased and this increase was greater in the IRM-SC. Moreover, as long as the DIF items took place in the common test, errors decreased in the IRM-CC, whereas they increased in the IRM-SC.

The differences among the four main effects given in Figure 2 and Figure 3 were also examined. The interaction of the test type containing DIF and DIF magnitude was significant in the IRM-SC method in ANOVA results. This case attracts attention with a sudden increase in RMSE and BIAS values under S6, S12, S18, and S24 conditions in which DIF took place in the common test and DIF magnitude was Type C.

According to the graphics in Figure 2 and Figure 3, the MIRM had the highest RMSE and BIAS values under all conditions, except for the 20-item conditions in the small sample size.

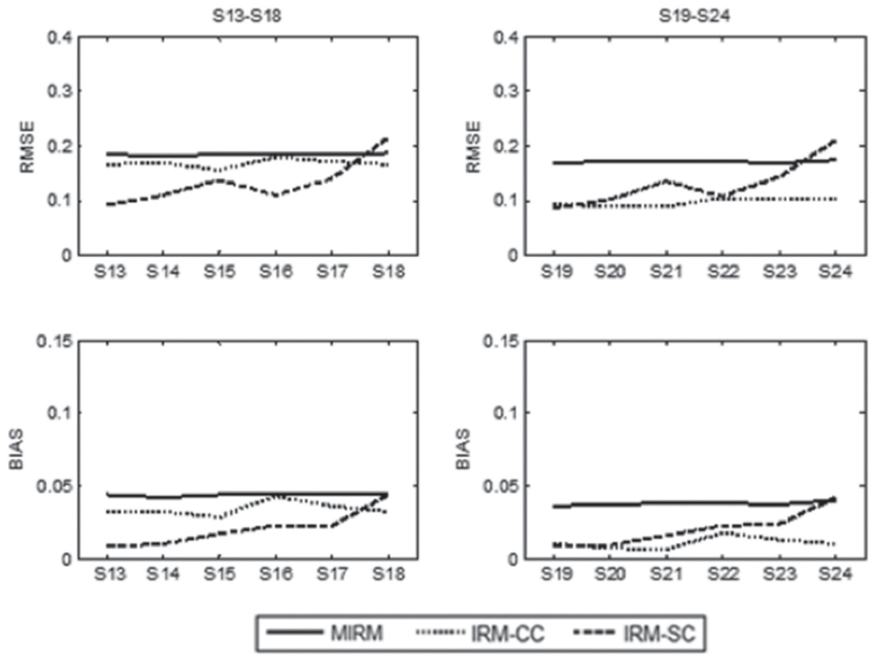

Figure 3: Means of the item parameters' errors in large sample.

When the tests containing DIF items were equated with the MIRM, IRM-CC, and IRM-SC methods, the means of the errors of ability parameters were calculated. Variance analyses were conducted for each of the two error values to determine the effects of the item parameters (i.e., sample size, test length, test type containing DIF, and DIF magnitude) on equating errors. Effect sizes and F values of the effects according to the methods are presented in Table 6. To determine how significant conditions obtained from the ANOVA results change, the graphics of the means of the errors are presented in Figure 4 for the small samples and in Figure 5 for the large samples.

Based on the ANOVA results of the two error types in the ability parameters given in Table 6, it was found that only the main effects of sample size and test length were significant in all three methods. Moreover, the RMSE values of the test type containing DIF was only significant in the IRM-SC method. Furthermore, according to the ANOVA results of the BIAS values in Table 6, the interaction of the test type containing DIF and DIF magnitude was significant in all three methods, while the interaction of sampling size and test length was only significant in the IRM-CC method.

After investigating the RMSEs of ability parameters, as presented in Figure 4 and Figure 5, it was found that the MIRM and the IRM-CC produced similar results under the 20-item conditions and these two methods produced a higher RMSE than the IRM-

Table 6
Significant ANOVA Results for Error Values of Ability Parameters

			Methods					
			IRM-CC		IRM-SC		MIRM	
Errors	Effects	Df	F	η^2	F	η^2	F	η^2
RMSE	Test Length (TL)	1	14075.27*	1	1298.19*	.99	8842.11*	1
	Test Type containing DIF Items (TT)	2	-	-	125.763*	.96	-	-
BIAS	Test Length (TL)	1	57.712*	.98	879.585*	.99	254.522*	.99
	TT * DM	2	17.063*	.79	72.289*	.94	35.449*	0.89
	SS * TL	1	75.000*	.89	-	-	-	-

*$p < .002$.

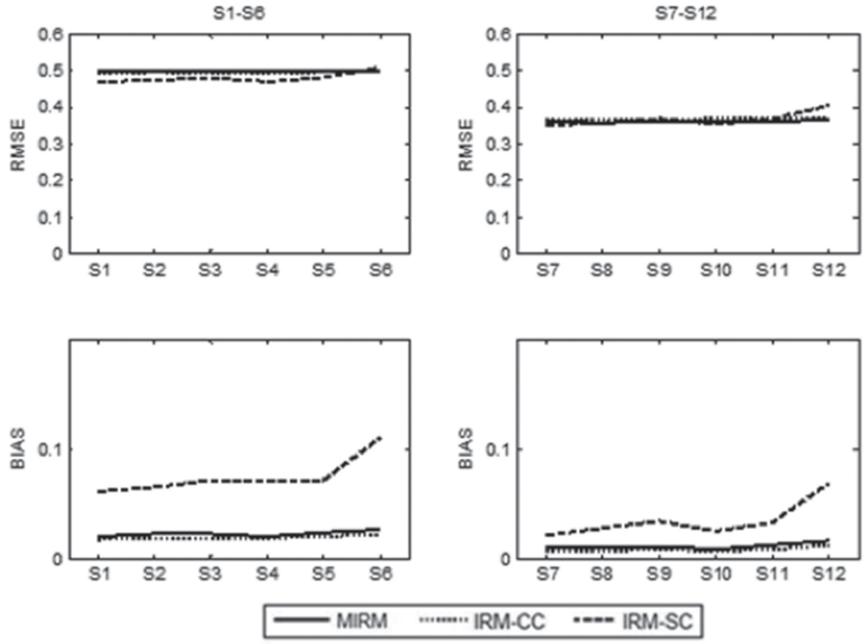

Figure 4: Means of ability parameters' errors in small samples.

SC. This finding is consistent with the results of Chu and Kamata (2000). Under the 40-item conditions, all three methods produced similar RMSE values. Moreover, the BIAS values were found to be higher

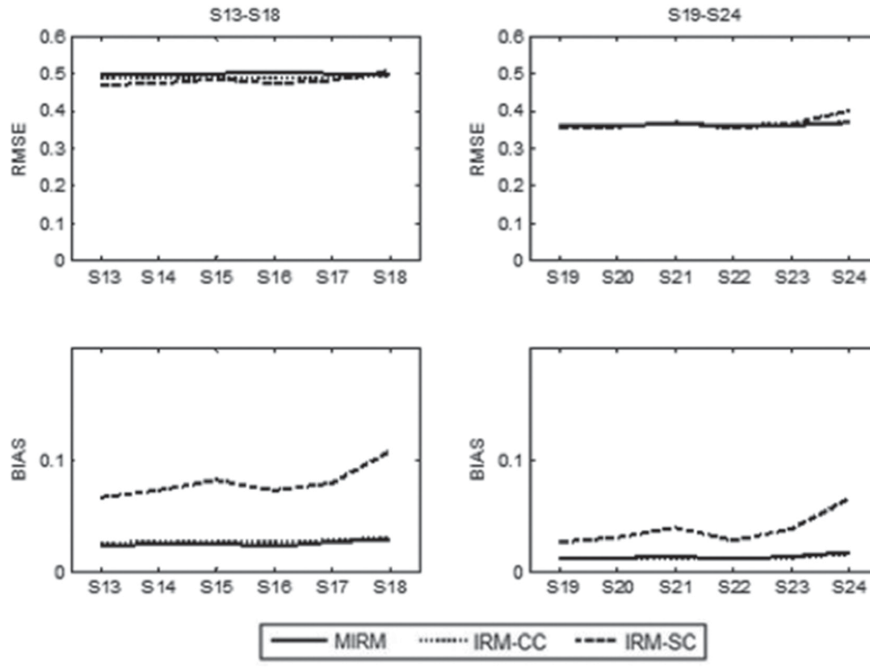

Figure 5: M'ean values of the errors of item parameters in the large samples.

in the IRM-SC method compared to concurrent equating methods.

A comparison of the 20-item and 40-item conditions in Figure 4 and Figure 5 shows that the increase in item number led to a decline in error values in each of two error types, and the method in which the error decreased the most by the increase in item number was the IRM-SC, especially in the BIAS values.

The mean RMSE values were investigated to understand how the error affected the test type containing DIF and DIF effect size, as presented in Figure 4 and Figure 5. Accordingly, it was found that the values were not affected from the aforementioned conditions in any of the three methods, and the BIAS values increased in the IRM-SC method as long as the DIF item took place in the common test and the DIF effect size increased.

The main effect of test length was significant in the ANOVA results was revealed by the differences as presented in Figure 4 and Figure 5. In addition, the interaction of the test type containing DIF and DIF magnitude was significant in the BIAS values of all three methods in ANOVA results can be seen the most clearly in the IRM-SC. This finding is especially important due to the sudden increase in the BIAS values under S6, S12, S18, and S24 conditions where the DIF magnitude is Type C and DIF items took place in the common test.

After investigating the effect of sample size on equating errors, as presented in Figure 4 and Figure 5, no effect was found in any of the three methods. Kolen and Brennan (2004) reported that sample size has no effect on bias and the increase in sample size has no effect on reducing bias. The interaction effects of sample size and test length were significant for the BIAS values in the IRM-CC method.

Finally, the graphics in Figure 4 and Figure 5 show that the highest equating bias values were obtained from the IRM-SC method. Studies have shown that concurrent calibration provides better results than separate calibration (Hanson & Beguin, 1999a; Kim & Cohen, 2002).

Discussion

In this study, three different equating methods were used to equate item and ability parameters. The findings show that the performances of the methods differed by the analyzed conditions. For example, the MIRM detected DIF items, conducted the equating process, and eliminated the bias derived from DIF through a single analysis. In the MIRM-based equations, the lack of the effect of the test type containing DIF items and DIF magnitude on error types was due to the fact that DIF items can be determined in the model. However, this does indicate that the MIRM is the best equating method since an increase in sample size and test length generally has a positive effect on IRM-based equating. In addition, the MIRM was less affected by these two conditions compared to the IRM. Furthermore, the MIRM analysis became more time consuming as the model became more complicated and the test length and sample size increased.

Previous studies have revealed that the MIRMs produce smaller or similar errors compared to IRM-based methods (Chu, 2002; Chu & Kamata, 2000, 2005; Luppescu; 2002). Consistent with previous studies, the present study also revealed that MIRMs produce smaller errors in small sample sizes and short test lengths. Low error values produced by MIRMs under these conditions are consistent with the results of other studies. However, the difference in the results of this study was due to its inclusion of large sample sizes and longer test conditions.

It was found that error values increased in extreme values of the *b* parameter for all three methods. In addition, it was shown that the increase in sample size and test length led to a decline in error in extreme values of the *b* parameter in IRM-based methods and it had no effect in the MIRM. In the MIRM, the increase in sample size and test length decreased error estimation in medium-difficult item parameters, whereas extreme values were not affected by these conditions. Thus, it was concluded that this case did not reduce means of error in the MIRM. After comparing the IRM-based methods, it was found that the increase in sample size and test length decreased errors in concurrent calibration more than separate calibration.

Finally, the results show that errors did not differ by sample size in any of three equating methods, and errors decreased as the test length increased. This finding is consistent with the results of Gök (2012), which showed that sample size had no positive effect on methods, while test length had a positive effect. In addition, studies have shown that equating performance is in accordance with equating steps and therefore, a one-step process is better than a two-step process (Chu, 2002; Hanson & Beguin, 1999b; Kim & Cohen, 1998). Consistent with other studies, the present study found the highest bias value in the IRM-SC.

Recommendations

The findings of this study can help test developing experts determine how test equating processes are affected in the presence of DIF. In addition, it is recommended that MIRMs be used in future test equating studies, especially those that focus on small samples, since such models can control DIF items and prevent incorrect decisions. For large samples, the IRM-CC can be used for equating, but DIF items should be determined and extra precautions should be taken so that the equating results are not affected.

In this study, a total of 24 conditions, as two different sample sizes, two different test lengths, three different test types containing DIF, and two different DIF magnitudes, were analyzed. Although the rate of DIF items was fixed in this study (10%), future studies should increase this rate.

Since the common test design was used in equivalent groups in this study, the equating process was conducted through a Level-2 IRM. For future studies, it can be recommended that a Level-3 IRM be applied in non-equivalent groups and Level-3 DIF factors can be added into the model. In addition, since one-parameter MIRMs were used in this study, only uniform DIF was included in the process. Thus, it is recommended that future studies use two-parameter MIRMs and include both uniform and non-uniform DIF in the process.

In this study, two different sample sizes, consisting of 1,000 and 4,000 examinees, respectively, were analyzed. For future equating studies, it is recommended that larger samples be employed. In addition, this study was based on data scored dichotomously. Therefore, equating errors of similar conditions should be researched by studying on the data scored as polytomously based and/or the data with mixed scoring. In other words, different conditions than those investigated in this study should be investigated to compare overall performance.

Finally, simulation data was used to compare equating methods in the presence of DIF items. If real data was used in this study, then it would have been difficult to determine and compare the accuracy of the methods. When real data is used, it is possible to determine the differences between the employed methods. However, similar studies should be conducted using real data along with simulation data so that the results from the two data sets can be compared.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and Gllamm procedures* (Doctoral dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3263842)

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*, 3–13.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measuremet, 28*(2), 147–162. doi:10.1111/j.1745-3984.1991.tb00350.x

Beretvas, S. N., & Williams, N. J. (2002, April). *The use of HGLM as a dimensionality assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Binici, S. (2007). *Random-effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: a comparison of estimation methods* (Doctoral dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3282570)

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Camili, G., & Shephard, L. A. (1994). *Methods for identifying biased test items.* London, UK: Sage.

Cho, S. J., & Cohen, E. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35*(3), 336–370. doi:10.3102/1076998609353111

Chu, K. L. (2002). *Equivalent group test equating with the presence of differential item functioning* (Doctoral dissertation). Available from ProOuest Dissertations and Theses database. (UMI No.3065477)

Chu, K. L., & Kamata, A. (2000, April). *Nonequivalent group equating via 1-P HGLLM.* Paper presented at the Annual Meeting of the American Educational Research, New Orleans, LA.

Chu, K. L., & Kamata, A. (2005). Test equating in the presence of DIF items. *Journal of Applied Measurement, Special Issue: The Multilevel Measurement Model, 6*(3), 342–354.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational measurement: Issues and Practice. 10*(3), 37–45. doi:10.1111/j.1745-3992.1991.tb00207.x

Croker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Holt, Rinehart and Winston.

Gök, B. (2012). *Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Doctoral dissertation, Hacettepe University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Gök, B., Kelecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenzsel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 35,* 3–16.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Han, K. T. (2007). *WinGen3: Windows software that generates IRT parameters and item responses [computer program].* Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3325324)

Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. *Applied Psychological Measurement, 33*(6), 491–493. doi:10.1177/0146621608319513

Hanson, B. A., & Beguin, A. A. (1999a). *Separate versus concurrent estimation of IRT item parameters in the common item equating design.* ACT research report series, Iowa City, IA. (Eric Document ED 438 310).

Hanson, B. A., & Beguin, A. A. (1999b, April). *Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24. doi:10.1177/0146621602026001001

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35–41. doi:10.1111/j.1745-3992.1989.tb00313.x

Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the lord statistic. *Educational and Psychological Measurement, 62*(1), 32–44. doi:10.1177/0013164402062001003

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Westport, CT: Praeger.

Kamata, A. (1998). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 9922331)

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x

Karasar, N. (2009). *Bilimsel araştırma yöntemi.* Ankara: Nobel Yayınevi.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131–143. doi:10.1177/01466216980222003

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26,* 25–41. doi:10.1177/0146621602026001002

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd. ed.). New York, NY: Springer.

Le, L. T. (2009). Investigation gender differential item functioning across countries ABD test languages for PISA science items. *International Journal of Testing, 9*(2), 122–133. doi:10.1080/15305050902880769

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

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

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software].* Skokie, IL: Scientific Software International, Inc.

Park, C., Kang, T., & Wollack, J. A. (2007, April). *Application of multilevel IRT to multiple-form linking when common items are drifted.* Paper presented at the 2007 annual meeting of the National Council on Measurement in Education, Chicago, IL.

Pastor, D. A. (2003). The use of multilevel ıtem response theory modeling in applied research: an illustration. *Applied Measurement in Education, 16*(3), 223–243. doi:10.1207/S15324818AME1603_4

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical & Statistical Psychology, 19*(1), 49–57. doi:10.1111/j.2044-8317.1966.tb00354.x

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2005). *HLM 6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific software.

Spence, P. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 9703612)

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. doi:10.1177/014662168300700208

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3216552)

Yıldırım, H. H., & Berberoğlu, G. (2006). Judgmental and statistical analyses of the PISA 2003 mathematics literacy items. *International Journal of Testing, 9*(2), 108–121. doi:10.1080/15305050902880736

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1999). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*: Chicago, IL: Scientific Software International.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software].* Skokie, IL: Scientific Software International, Inc.

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.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.