Eğitim ve Bilim
2014, Cilt 39, Sayı 172
Geniş Ölçekli Test Uygulamaları Özel Sayısı

Education and Science
2014, Vol. 39, No 172
Large-Scale Assessment Special Issue

Findings from an Empirical Vertical Scaling Study with BILOG-MG

BILOG-MG ile Empirik Bir Dikey Ölçekleme Çalışmasından Bulgular

Hüseyin Hüsnü YILDIRIM[1]
Abant İzzet Baysal University

Abstract

This study compared two procedures for the vertical scaling in the Item Response Theory (IRT) context: fixed estimation, and simultaneous estimation. The results favored the simultaneous estimation procedure to the fixed estimation procedure, especially when there were few anchor items. However, the results also revealed that using expected a posteriori estimates (EAP) of ability scores in 3-parameter IRT model may have a deteriorating effect on the vertically scaled test results through the simultaneous estimation procedure. Overall, the results of this empirical study showed that in the large scale tests which aim to monitor the development across grade levels, the simultaneous estimation procedure with the 2-parameter or the 3-parameter IRT models would be a reasonable choice.

Keywords: Vertical scaling, item response theory, Bilog-MG, simultaneous estimation

Öz

Bu çalışmada, Madde Tepki Kuramına (MTK) dayalı iki dikey ölçekleme yöntemi (sabitlemeye dayalı kestirim ve eşzamanlı kestirim) karşılaştırılmıştır. Sonuçlar, özellikle çapa madde sayısının az olduğu durumlarda, eşzamanlı kestirimin daya iyi sonuçlar verdiğini göstermektedir. Ancak, 3 parametreli MTK modeli kullanılarak öğrenci yeterliklerinin "expected a posteriori" (EAP) yöntemiyle kestirildiği durumlarda dikey ölçeklenmiş değerlerde bozulma görülmektedir. Genel olarak bu çalışmanın sonuçları, sınıf seviyeleri arasındaki gelişimin takip edilmesi amacıyla yürütülebilecek geniş ölçekli test uygulamalarında, 2 veya 3 parametreli MTK modellerine dayalı eşzamanlı kestirimin makul bir seçenek olabileceğini göstermektedir.

Anahtar Sözcükler: Dikey ölçekleme, madde tepki kuramı, Bilog-MG, eşzamanlı kestirim

[1] Assist. Prof. Dr. Hüseyin Hüsnü Yıldırım, Abant İzzet Baysal University, Faculty of Education, Gölköy, BOLU, Turkey; yildirim.huseyin@ibu.edu.tr

Introduction

A measurement scale is a number system with a certain metric on which students' performance on a set of tasks is represented. Accordingly, scores of students in a test can be considered as values of such a scale. In item response theory (IRT) applications these scale values are called ability estimates or scale scores and they form a basis to compare students' test performance.

However, ability estimates are not directly comparable if they are estimated using data from different test administrations. To get comparable estimates in such cases some further process is necessary, which might consist of some complicated steps, especially when the different tests are of different difficulty level and examinees who took the different tests can not be considered as individuals from the same population (Stocking & Lord, 1983). The process to determine equivalent or comparable ability estimates in such a case is called vertical scaling, which is the main concern of this study. In this study vertical scaling is considered in the IRT context. Thus, it might be convenient to go over the main points of the issue from this perspective.

In logistic IRT models, metric of the ability is undetermined (Cook & Eignor, 1991). This means that a suitable linear transformation of ability and item parameter estimates can also equally fit the IRT model used as the initial parameters. This is just a plain consequence of the mathematical structure of the logistic IRT models as described below.

The logistic IRT models identify the probability of a correct (or incorrect) answer to an item i by a person with ability level θ. These models are basically a function of $a(\theta - b)$ where a and b are respectively, the discrimination and the difficulty parameters of item i (Hambleton, Swaminathan, & Rogers, 1991). Thus, if the values a, b and θ fit the model, their linear transformations of a/x, $xb + y$ and $x\theta + y$ will also fit the model equally well for any set of real numbers x and y (provided that $x \neq 0$). This is due to the simple fact that $a(\theta - b)$ is equal to $a/x((x\theta + y) - (xb + y))$.

To deal with this indeterminacy, the scale in IRT analyses is to be fixed which can be done in various ways. For example, in analyses using the 3-parameter IRT model the traditional approach is fixing the ability scale such that the mean of the sample θ estimates is 0 and its variance is 1. Another approach, which is usually employed in calibrations (i.e., the process for estimating item parameters) using the Rasch model, is setting the mean of item difficulty parameter estimates to 0.

However this is only the one side of the story. It is crucial to notice that these scale fixing depends on the characteristics of the calibration sample from which the test data comes, and the test itself. In other words, fixing the scales in two separate IRT calibrations using the same approach, does not assure equivalency of the scales unless the samples and the tests in these separate cases can be considered as exactly the same. Otherwise, a further process is required to get comparable item parameters or ability estimates obtained through separate test administrations (Cook & Eignor, 1991). This process can be one of two main types: the horizontal scaling and the vertical scaling.

The process is called horizontal scaling if the test forms administered in separate sessions can be considered as parallel forms of the same test, and the individuals taking the assorted test forms can be considered as random samples from the same population. A language test in which examinees take one of the parallel multiple test forms can be considered as an example. On the other hand, if the test forms are of different difficulty level and individuals replying the different forms can not be considered as samples from the same population then the process is called vertical scaling (Kim & Hanson, 2002).

For example, a mathematics test administration in which students of 6^{th}, 7^{th} and 8^{th} grades replies one of the test forms developed with respect to their grade level can be an example. In such a case results from the grade-specific test forms are to be placed on a common scale if one intends to monitor the growth of individuals across the grade levels. This type of calibration process to provide a common scale across test forms of different difficulty and across nonequivalent respondent groups is known as vertical scaling. As compared to horizontal scaling, the main difference in vertical scaling is that it requires anchor items; that is, common items to be used at both of the consecutive test forms in the series. Further information on vertical scaling is provided in the next section.

Vertical scaling is far from being a straightforward process yet. The research in this area shows that the amount of differences in test difficulty from grade to grade, scaling methods and IRT models used in the analyses, methods of estimating scale scores, the number and the psychometrical properties of anchor items all have an effect on the vertical scaling process. To make the things more complicated some miscellaneous findings in the literature may be found. For example, one can read in the literature both that common item anchor should be at least %20 of the total test, and that when groups are not equivalent accurate linking can be obtained with as few as 2 anchor items (Vale, 1986; Cook and Eignor, 1991). Another finding on this issue comes from simulation studies which show that when the sample sizes are large enough, as few as 4 anchor items can precisely equate the scales between test forms (Lord, 1980).

Thus, one who needs vertical scaling in her endeavor will most likely has to deal with all these complications to decide on her way in the analyses. In such cases empirical studies that present possible impacts of choosing different methods on practical applications would provide a significant guide. To this purpose this current study presents effect of a) different linking methods (i.e., simultaneous estimation and fixed estimation methods) and b) different IRT scale score estimation methods (i.e., maximum likelihood (ML), expected a posteriori (EAP), and maximum a posteriori (MAP) on the vertical scaling using the 2-parameter and 3-parameter IRT models, and the estimation software BILOG-MG. A real data set of a mathematics achievement test administered to 6th, 7th and 8th grade-students was used in the study. The results were evaluated using the estimated proficiency distributions.

Method

For this study, analyses were carried out with the program BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996; du Toit, 2003). The data used in this study consisted of 6th, 7th and 8th grade-students responses to three mathematics achievement test forms which were assigned to the students with respect to their grade levels. The further details are as follows.

Data and Sample

Test forms for grade 6, 7 and 8 consisted of 40, 45 and 45 multiple choice mathematics items, respectively. Tests were constructed by a group of educational experts. Test forms for the 6th and 7th grades contained 6 common items. Similarly, test forms for the 7th and 8th grades contained another 6 common items as well. These common items were used as the anchor items required for the vertical scaling. These common items were determined by the educational experts as to be suitable for both of the grade levels they were administered.

Tests were administered by schools at the beginning of the academic-year 2012-2013 to determine how ready their students were for the mathematics subjects to be taught during the school term. In total, 13002 students in 104 schools sat for the tests. The number of the students who took the tests in grades 6, 7 and 8 were 4791, 4412 and 3799, respectively.

Tests were constructed with respect to the national mathematics curriculum of Turkey (MEB, 2009). Items were written one year before the intended test administration time, and conducted a pilot study. The final version of the tests included only the items that met certain psychometric properties in the pilot study. More specifically, items with difficulty values higher than 0.9 or lower than 0.1 were avoided, a point-biserial correlation of minimally 0.30 was required, and item information functions were investigated to make sure that items provided reasonable amount information.

All three test forms comprised number, algebra, geometry, and probability items. The items in the successive test forms were, in general, increasingly more demanding. For example, an algebra item for the 6th grade expected students to solve a simple linear equation involving only one variable, whereas an item in the same content area for the 8th grade expected students to solve pairs of simultaneous equations involving two variables. However, the test forms also included items that were suitable for neighboring grades (i.e., grades 6 and 7, or grades 7 and 8). These items provided the anchor items required to link the test forms on a common scale.

IRT Models and Estimation

The 2-parameter IRT model describes performance on a multiple-choice item with respect to two parameters: item difficulty and item discrimination. In addition to these two parameters, the 3-parameter model also takes into account the probability of a correct response to a multiple-choice item as a result of guessing (Hambleton, Swaminathan, & Rogers, 1991).

Logistic models are mathematical functions of the expression $z_j = a_j(\theta - b_j)$, which is referred to as a logistic deviate, where a_j and b_j are discrimination and difficulty parameters of item *j*, respectively. Theta in the expression stands for the respondent's ability.

The two-parameter logistic model is defined as

$$P_{2j}(\theta) = \frac{1}{1 + e^{-z_j}} \tag{1}$$

The three-parameter model as stated in (2) is an extension of the two-parameter model considering the probability g_j of a correct response to the multiple-choice item *j* as a result of guessing.

$$P_{3j}(\theta) = g_j + (1 - g_j)P_{2j}(\theta) \tag{2}$$

The formulas (1) or (2) determine the relationships between item responses and the trait being measured. These relationships provide the basis for test scoring. But first, the parameters for each item of the test are to be estimated. BILOG-MG uses the method of marginal maximum likelihood (MML) to this purpose (Bock & Aitkin, 1981). The process of estimating item parameters and checking the fit of the models is referred to as item calibration.

Estimating respondent's ability, which is called scale score estimation, follows the item calibration in BILOG-MG. In estimating the ability parameters, estimated item parameters at the item calibration phase are treated as known constants. Three types of IRT scale score estimation methods used in this study are maximum likelihood estimation, Bayes estimation, and Bayes modal estimation.

Maximum likelihood (ML) estimate of the scale score of respondent *i* in an *n*-item test is the value of θ that maximizes the expression given in (3) (Bock, 2003a).

$$\sum_{j=1}^{n} \{x_{ij} \log_e P_j(\theta) + (1 - x_{ij}) \log_e (1 - P_j(\theta))\} \tag{3}$$

In the expression (3), x_{ij} is either 1 or 0 according as the person *i* responses correctly or incorrectly to item *j*. $P_j(\theta)$ is the IRT model used; that is one of the models given in (1) or (2).

The Bayes estimate is the mean of the conditional posterior distribution of θ which is also called expected a posteriori (EAP) estimator. The condition is the observed response pattern. And finally Bayes modal or maximum a posteriori (MAP) estimator is the value of θ that maximizes the expression (4) in which $g(\theta)$ indicates the density function of population distribution of θ.

$$\sum_{j=1}^{n} \{x_{ij} \log_e P_j(\theta) + (1 - x_{ij}) \log_e (1 - P_j(\theta))\} + \log_e g(\theta) \tag{4}$$

Vertical Scaling

As David Thissen states, "the magic of IRT arises in placing all of the test scores on the same scale…" (2003; p. 592). The reason why he calls it a magic may lie in the power of IRT models to meet, at least to some extent, the strict requirements stated by Angoff (1984) to qualify that the scale scores from different tests have been equated.

IRT vertical scaling is a general name of the procedures for placing results from grade-specific test forms on a common scale. It is an application of non-equivalent groups equating in which the students are administered a certain grade-specific test corresponding to their groups (Kim, Lee, Kim, & Kelley, 2009).

Vertical scaling may provide a significant opportunity at the school or country level to monitor students' growth on a specific subject across the grade levels. Fixed estimation and simultaneous estimation are the two methods that can be conducted for vertical scaling in BILOG-MG (du Toit, 2003). Both methods require anchor items between consecutive test forms for the grade levels.

Fixed Estimation. To carry out the fixed estimation in this study, first data from the 6th grade-test (reference test) was used for the item calibration and the scale score estimation. To deal with the scale indeterminacy problem as stated in the introduction section, the ability scale was determined such that the mean of the sample θ estimates was 0 and its variance was 1.

In the second step, data from the 7th grade-test (target test) was used for the item calibration and the scale score estimation. However, the parameters of the common items between the 6th and the 7th grade tests were fixed to the values estimated in the first step where data from the 6th grade test was calibrated, and were not re-estimated in the second step. The parameters of the remaining non-common items in the target test were estimated in the item calibration phase. This procedure sets the parameter estimates of the items in the target test to the scale of the reference test. Thus, the scale of the target test is not indeterminate anymore as it is already aligned to that of the reference test. In other words, the estimates are not to be rescaled anymore to have a mean of 0 and a variance of 1. To this purpose in BILOG-MG analyses, the FIX keyword in the >TEST command and NOADJUST option in the >CALIB command were used. This second step ended with the estimation of the scale scores of the 7th grade students.

The fixed estimation procedure in this study was completed by the third step in which the 8th grade-test data was analyzed. This step was conducted in the same way as the second step. However, in this step the 7th grade-test was considered as the reference test and the 8th grade-test was considered as the target test.

Simultaneous Estimation. In simultaneous estimation, as compared to the fixed estimation, the item and ability parameter estimates for all grades may be put on a common scale in a single run. In other words, there is no need for separate linking and scaling steps.

The underlying principle for simultaneous estimation is as follows: Let n_6, n_7 and n_8 be the number of students who responded the tests in the 6th, 7th and 8th grades, respectively. In addition, suppose that m_6, m_7 and m_8 are the number of unique items only in the tests for the 6th, 7th and 8th grades, and that $m_{6\text{-}7}$ and $m_{7\text{-}8}$ are the number of anchor items between the 6th and 7th grades, and between the 7th and 8th grades, respectively. Then, item responses of students in all 3 grade levels are entered into a large data matrix of dimensions $(n_6 + n_7 + n_8)$ x $(m_6 + m_7 + m_8 + m_{6\text{-}7} + m_{7\text{-}8})$. In these blocks of item responses, items not administered to the corresponding students are treated as missing values. The multi-group design of BILOG-MG allows the group of students from different levels to differ in the mean and standard deviation of their ability distribution. Then within this design, parameters of all the items in the tests and scale scores of all the students in the groups can be estimated concurrently (Fischer & Molenaar, 1995).

In this study, simultaneous estimation was conducted as defined above. To deal with the indeterminacy problem the mean and standard deviation of ability estimates of the 6th grade students was determined to be 0 and 1, respectively. The mean and standard deviation of ability distributions for 7th and 8th grade were estimated freely, but adjusted by the BILOG-MG with respect to the values determined for the 6th grade distribution.

In this current study, both the fixed estimation and the simultaneous estimation procedures were replicated (2 x 3) = 6 times considering the 2 IRT models (i.e., 2-parameter and 3-parameter models), and the 3 scale score estimation methods (i.e., ML, EAP and MAP).

Evaluation Criteria

In empirical studies on vertical scaling that compare different models, it is often difficult, if not impossible, to find an objective basis for deciding which of the models better represents the students' true growth in the construct being measured. This is due to the fact that the true scores of the students are not known but only estimated through a measurement model. Moreover, it is a well-known fact that different IRT models could fit the response data equally well and yet exhibit different relationships across the grade level scores (Bock, 2003b).

The general tendency to deal with this obscurity is building a somehow reasonable basis for an evaluation (Kim et al., 2009). In this study, results were evaluated with respect to two criteria. First criterion depends on the assumption that, on a vertically equated scale, the average scale score of students at a higher grade should also be somewhat higher than the average scale score of students at a lower grade. Second criterion depends on the fact that the relationship between IRT ability scales derived from two different calibrations is always linear. In other words ability scales derived from two different calibrations differ only in origin and unit of measurement (Cook & Eignor, 1991). A corollary of this fact is that individually estimated scale scores of a certain grade level (i.e., scores estimated using only the test data of a certain grade level) should coincide with the vertically equated scale score estimates of the same grade level after a linear transformation that equalizes the mean and the standard deviation values of these two estimates.

Results

Checking the Model Assumptions

The IRT models used in this study require the tests to be unidimensional. For a test to be unidimensional, the variance in the responses of examinees to the test items should be accounted for by a single latent trait (Hambleton et al., 1991). To check this assumption, scree plots of the test data in all three grade levels were investigated separately. The plots obtained through principal component analysis are given in Figure 1. The plots clearly indicate that the test data in all three grades can be considered as unidimensional.

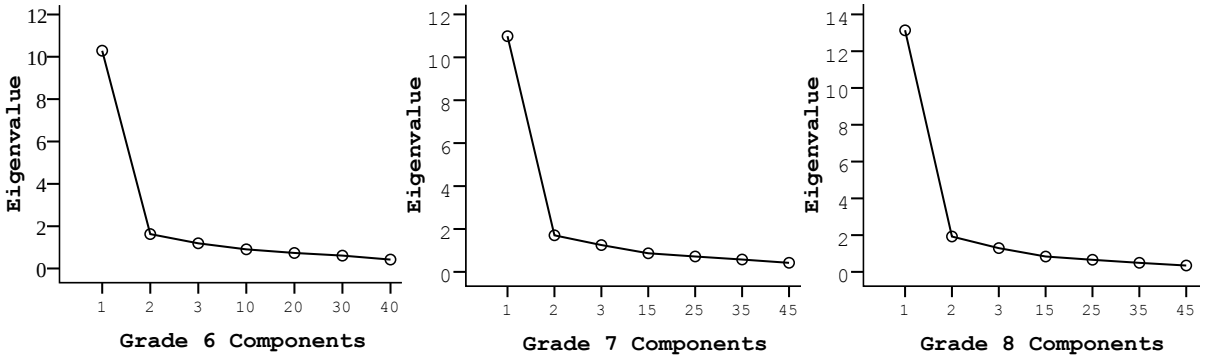


Figure 1. Scree plots for the 6th, 7th and 8th grades test data.

The second important requirement for the vertical scaling analyses is the psychometric quality of the anchor items. Anchor items should have relatively high discrimination value and middle-range difficulty, and they should not function differentially between the grade levels. (Kim et al., 2009).

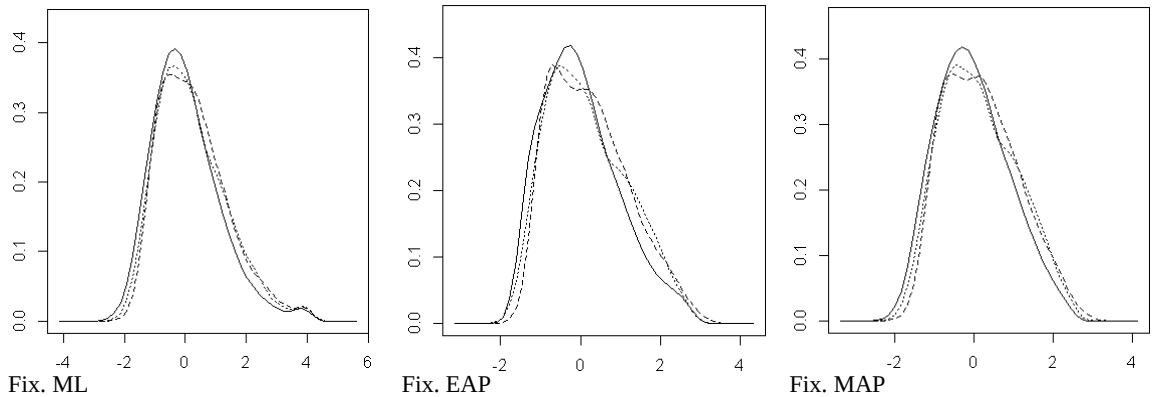
Biserial correlations of the anchor items in this study ranged between 0.48 and 0.57, and the proportion correct values ranged between 0.39 and 0.61. These statistics indicated that anchor items had appropriate discrimination and difficulty values. However, there was a problem on the DIF side. Differential item functioning (DIF) analyses in this study were conducted with the built-in DIF model in the program BILOG-MG.

Unexpectedly, 2 of the 6 anchor items between the 6th and 7th grades, and 4 of the 6 anchor items between the 6th and 7th grades were detected as functioning differentially between the corresponding grade levels. Thus, the detected items were not used as anchor items, leaving only 4 anchor items between the 6th and 7th grade levels and 2 anchor items between the 7th and 8th grade levels. However, these detected items were not excluded from the single grade level analyses in which scale scores of students of a certain grade level were estimated using only the test data of that grade level.

Fixed Estimation versus Simultaneous Estimation

Figure 2 presents the distribution of estimated scale scores for the fixed and the simultaneous estimation procedures using the 2-parameter IRT model. For each of the procedures, scale scores were estimated three times using the three scale score estimation methods (i.e., ML, EAP and MAP).

Graphs in Figure 2 are the smoothed density estimates of the histograms of the scale scores obtained in the analyses (Bowman & Azzalini, 1997). The graphs were obtained using the software system R (Crawley, 2007).



Fix. ML

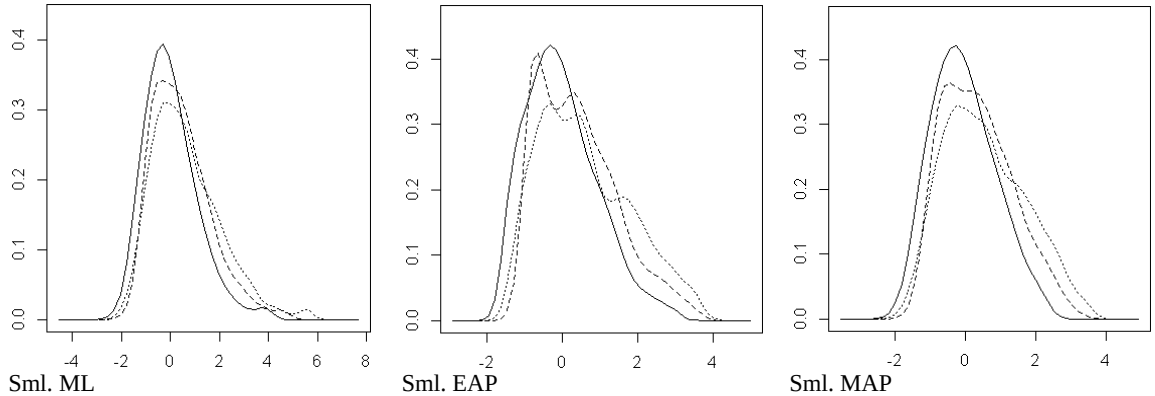
	6th	7th	8th
M	0.002	0.243	0.179
S	1.062	1.088	1.096
R	0.925	0.930	0.941

Fix. EAP

	6th	7th	8th
M	0.000	0.233	0.174
S	0.956	0.982	0.998
R	0.906	0.914	0.921

Fix. MAP

	6th	7th	8th
M	-0.031	0.190	0.132
S	0.908	0.934	0.950
R	0.915	0.922	0.933



Sml. ML

	6th	7th	8th
M	-0.002	0.421	0.600
S	1.051	1.174	1.334
R	0.925	0.936	0.944

Sml. EAP

	6th	7th	8th
M	0.000	0.407	0.593
S	0.954	1.054	1.214
R	0.904	0.915	0.923

Sml. MAP

	6th	7th	8th
M	-0.034	0.370	0.551
S	0.901	0.999	1.159
R	0.915	0.925	0.936

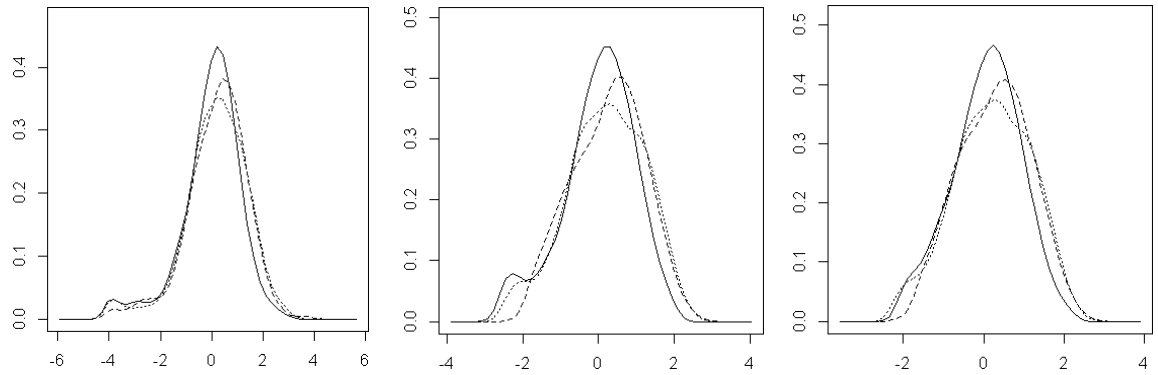
Fix.: Fixed estimation; Sml.: Simultaneous estimation; ML: Maximum likelihood; EAP: Expected a posteriori; MAP: Maximum a posteriori; M: Mean of scale scores; S: Standard deviation of scale scores; R: Reliability of scale scores;

————: 6th grades, - - - - - - - : 7th grades, ... : 8th grades.

Figure 2. Distributions of scale score estimates for the fixed and simultaneous estimation methods using the 2-parameter IRT model.

Both in the fixed and the simultaneous estimations the mean and the standard deviation of the 6th grade population were determined to be 0 and 1, respectively. However, one should notice that these are the population parameters; thus the estimated mean and standard deviation values are not exactly 0 and 1 due to the estimation error, yet very close to these values.

Figure 3 presents the same graphs as Figure 2 but which were obtained using the 3-parameter IRT model.



Fix. ML

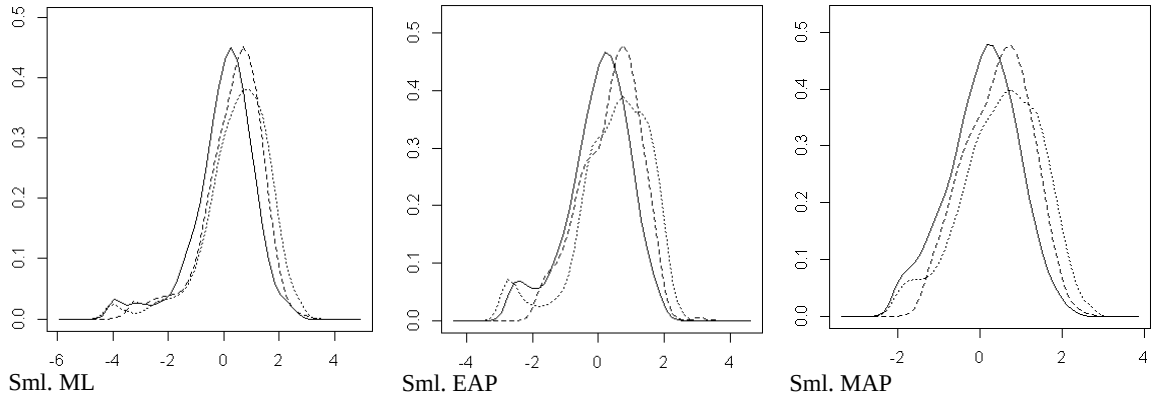
	6th	7th	8th
M	0.026	0.233	0.210
S	1.073	1.129	1.154
R	0.939	0.941	0.946

Fix. EAP

	6th	7th	8th
M	-0.002	0.245	0.192
S	0.954	0.954	1.033
R	0.898	0.905	0.925

Fix. MAP

	6th	7th	8th
M	0.059	0.273	0.210
S	0.863	0.901	0.975
R	0.921	0.924	0.936



Sml. ML

	6th	7th	8th
M	0.026	0.373	0.510
S	1.057	0.994	1.130
R	0.941	0.946	0.953

Sml. EAP

	6th	7th	8th
M	-0.002	0.365	0.462
S	0.952	0.875	1.104
R	0.895	0.891	0.919

Sml. MAP

	6th	7th	8th
M	0.066	0.430	0.543
S	0.848	0.789	0.962
R	0.923	0.928	0.943

Fix: Fixed estimation; Sml.: Simultaneous estimation; ML: Maximum likelihood; EAP: Expected a posteriori; MAP: Maximum a posteriori; M: Mean of scale scores; S: Standard deviation of scale scores; R: Reliability of scale scores;

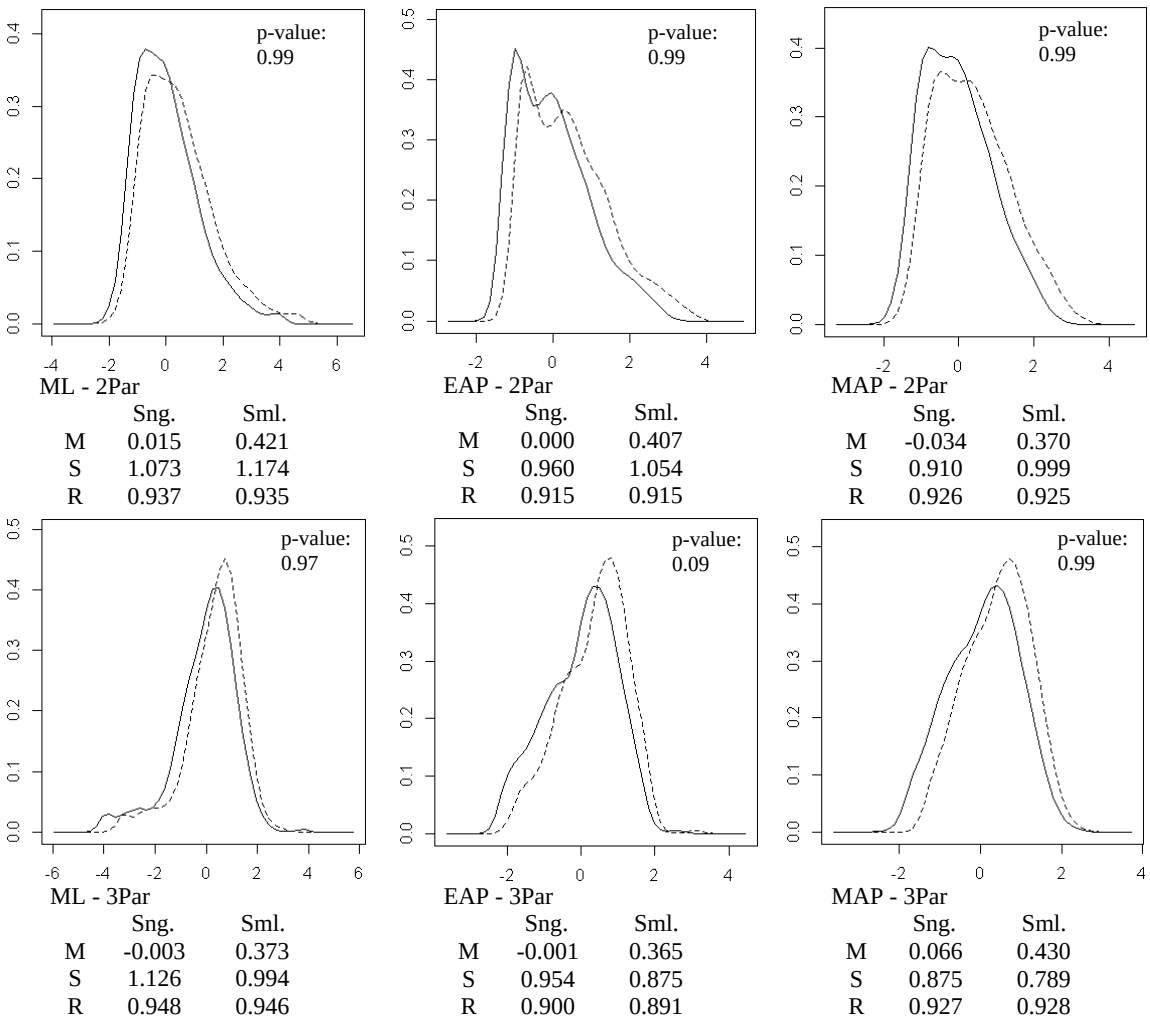
———— : 6th grades, - - - - - - - : 7th grades, ... : 8th grades.

Figure 3. Distributions of scale score estimates for the fixed and simultaneous estimation methods using the 3-parameter IRT model.

The results given in Figure 2 and Figure 3 showed that, both for the 2-parameter and the 3-parameter IRT models, fixed estimation procedure did not reflect a possible increase in the 8th grade average scale score as compared to that of the 7th grade. This is possibly due to the fact that only 2 anchor items were used between the 7th and 8th grade test forms. In addition, the average score gain between the 6th and 7th grades in all the analyses obtained through the fixed estimation procedure was considerably lower than the gain obtained through the simultaneous estimation procedure. Thus, one may conclude that the fixed estimation procedure is an insufficient way in vertical equating with a small number of anchor items. To this reason, investigating the results with respect to the second evaluation criterion was conducted only for the simultaneous estimation procedure.

As stated before, second evaluation criterion based on the comparison of the pairs of scale score estimates of the same group; one obtained from the single group, and the other from the simultaneous analyses. Single group analyses were those that used only the test data of the group under investigation. In single group analyses the mean and the standard deviation of the score estimates of the group under investigation was determined to be 0 and 1, respectively. Simultaneous analyses were exactly the same as given above in Figure 2 and Figure 3.

Figure 4 presents 6 pairs of scale score distributions for the 7th grades obtained through the 2-parameter and the 3-parameter IRT models for each of the 3 scale score estimation methods.



ML: Maximum likelihood; EAP: Expected a posteriori; MAP: Maximum a posteriori; M: Mean of scale scores; S: Standart deviation of scale scores; R: Reliability of scale scores; 2Par: 2-parameter model; 3Par: 3-parameter model; Sng.: Single group estimation; Sml.: Simultaneous estimation.

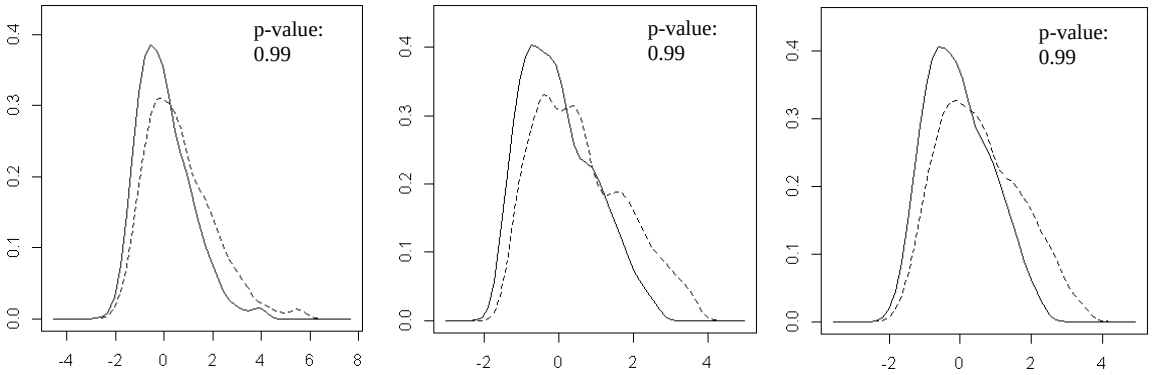
———— : Scores from single group estimation, - - - - - - -: Scores from simultaneous estimation

Figure 4. Distributions of scale score estimates for the 7th grade level.

With respect to the second evaluation criterion, to claim that the simultaneous estimation procedure for vertical scaling produces comparable test scores, the pairs of scale score distributions given in Figure 4 should be similar up to a linear transformation. That is, the distributions should coincide when one the distributions is linearly transferred to have the same mean and standard deviation values as the other one.

In line with this manner, the p-values given on the graphs can be interpreted as the degree of overlap between the two distributions after the first one is linearly transferred to have the same mean and variance as the second distribution. The values were obtained through the software R (Bowman & Azzalini, 1997; Crawley, 2007).

Similarly, Figure 5 presents the results at the 8th grade level. With respect to the results one should notice that the EAP estimates of the scale scores obtained through the 3-parameter IRT model produced relatively small amount of overlap between the corresponding pairs of distributions both at the 7th and at the 8th grades.



ML - 2Par

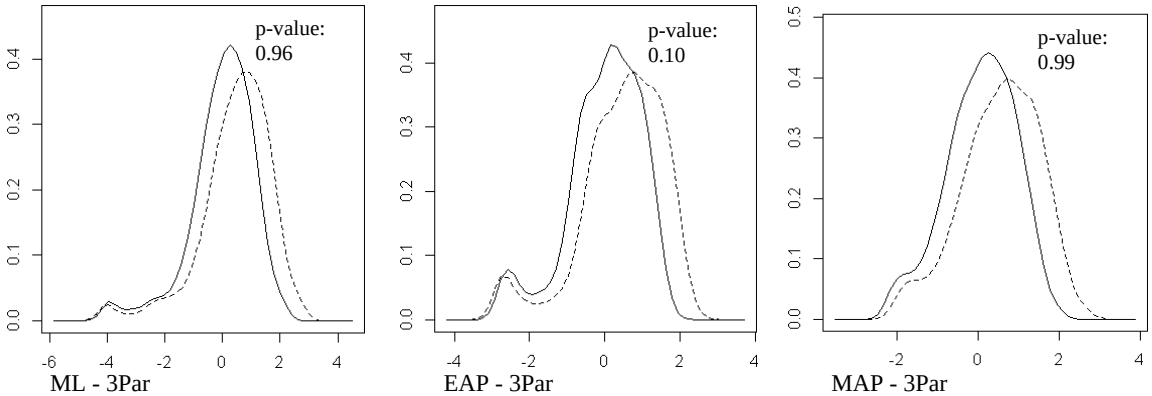
	Sng.	Sml.
M	0.005	0.600
S	1.056	1.334
R	0.943	0.944

EAP - 2Par

	Sng.	Sml.
M	0.000	0.593
S	0.963	1.214
R	0.921	0.923

MAP - 2Par

	Sng.	Sml.
M	-0.033	0.551
S	0.918	1.159
R	0.935	0.936



ML - 3Par

	Sng.	Sml.
M	0.040	0.510
S	1.008	1.130
R	0.950	0.953

EAP - 3Par

	Sng.	Sml.
M	-0.001	0.462
S	0.966	1.105
R	0.921	0.919

MAP - 3Par

	Sng.	Sml.
M	0.062	0.543
S	0.863	0.962
R	0.941	0.943

ML: Maximum likelihood; EAP: Expected a posteriori; MAP: Maximum a posteriori; M: Mean of scale scores; S: Standart deviation of scale scores; R: Reliability of scale scores; 2Par: 2-parameter model; 3Par: 3-parameter model; Sng.: Single group estimation; Sml.: Simultaneous estimation.

——— : Scores from single group estimation, - - - - - - -: Scores from simultaneous estimation

Figure 5. Distributions of scale score estimates for the 8th grade level.

Discussion and Conclusion

Vertical scaling presented in this current study belongs to a general family of the equating methods known as Nonequivalent Groups Anchor Test (NEAT) design. There is a set of criteria to be met for a NEAT design equating to be successful (Angoff, 1984; Lord, 1980).

With respect to these criteria the tests to be equated should measure the same construct (validity), the same measurement model should fit the data in all groups (invariance), equation that maps test score X on test A to test score Y on test B should also map score Y on test B to score X on test A (symmetry), and it should not matter to a student which test to take (equity). Arguments in the psychometrical literature indicate that latent variable models are better in meeting these criteria (van der Linden, 2000; Maris, Schmittmann, & Borsboom, 2010). Because of this, the vertical scaling in this study was conducted by the 2-parameter and the 3-parameter IRT models. In addition, the software BILOG-MG used in this study provides an opportunity to define separate parameters (i.e., mean and standard deviation) for each group so that the existence of group differences can be accounted for. Otherwise, the simultaneous estimation method used in this study would not be realized.

The results of this study revealed that the number of anchor items might be an important factor in vertical scaling through the fixed estimation method. The increase on the average group scale scores obtained in the fixed estimation between the grades 6 and 7 was almost as half as the increase obtained in the simultaneous estimation. There were 4 anchor items between the tests at these grade levels. Besides, when the number of anchor items happened to drop to 2 between the grades 7 and 8, no increase between these grade levels was observed in the fixed estimation as opposed to the simultaneous estimation. However, the simultaneous estimation method managed to reflect a possible increase in the group means with as few as 2 items which is in line with the finding of Vale (1986). Thus, this might be interpreted as, especially with the modest number of anchor items, the simultaneous estimation procedure should be preferred to the fixed estimation method.

However, in some cases it may not be possible to use the simultaneous estimation due to some practical constraints of the test conditions. For example, test data from all the group levels may not be ready at the same time, and consequently, the analyses may be conducted only with the available data. Then, if one wants to analyze the rest of the data at the same metric scale of the first calibration phase, the fixed estimation method is to be used. Therefore, renewing this current study with an increased number of anchor items would be an important supplement to this study, which may reveal some information on the number of anchor items required for a reasonable scaling in the fixed estimation process.

When the results of the simultaneous estimation are further investigated, it is clear that the vertically scaled EAP ability estimates with the 3-parameter IRT model should be used with caution. As mentioned before, EAP estimates use the observed response pattern as a condition to the conditional posterior distribution of ability. In addition the 3-parameter IRT model takes into consideration the guessing effect (Hambleton et al., 1991). The results of this study showed that this combination had somehow deteriorating effect on the vertical scaling results. Some further theoretical research on this issue may clarify the causes for this effect.

In sum, the NEAT design is a very challenging issue in the psychometric literature that there are researchers who suggest avoiding it whenever possible (eg., Maris et al., 2010). However, in the large scale tests which aim to monitor the development across the grade levels using a NEAT design seems to be inevitable. The results of this current study showed that in such test occasions, the simultaneous estimation procedure with the 2-parameter and the 3-parameter IRT models would be a reasonable choice, even with a few anchor items.

References

Angoff, W.H. (1984). *Scales, norms, and equivalent scores*. Princeton NJ: Educational Testing Service.

Bock, R.D. (2003a). Estimation in BILOG-MG. In M. du Toit (Ed.), *IRT from SSI,* (pp. 599-611). Illinois: Scientific Software International.

Bock, R.D. (2003b). Uses of Item Response Theory. In M. du Toit (Ed.), *IRT from SSI,* (pp. 618-633). Illinois: Scientific Software International.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443-445.

Bowman, A.W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis.* Oxford: Clarendon Press.

Cook, L.L. & Eignor, D.R. (1991). An NCMF instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, 10, 37-45.

Crawley, M.J. (2007). *The R book.* West Sussex: John Wiley & Sons. du Toit, M. (Ed.) (2003). *IRT from SSI.* Illinois: Scientific Software International.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Hambleton R.K., Swaminathan, H., Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* New York: Springer-Verlag.

Kim, J., & Hanson, B.A. (2002). Test equating under the multiple choice model. *Applied Psychological Measurement*, 26(3), 255-270.

Kim, J., Lee, W., Kim, D., & Kelley, K. (2009). *Investigation of vertical scaling using the Rasch model.* Paper presented at the annual meeting of NCME, April 2009, San Diego.

Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Maris, G., Schmittmann, V.D. & Borsboom, D. (2010). Who needs linear equating under the NEAT design. *Measurement*, 8, 11-15.

MEB. (2009). *İlköğretim matematik dersi 6-8. sınıflar öğretim programı ve kılavuzu.* Ankara: T.C. M.E.B. Talim ve Terbiye Kurulu Başkanlığı.

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement*, 7(2), 201-210.

Thissen, D. (2003). Estimation. In M. du Toit (Ed.), *IRT from SSI,* (pp. 592-599). Illinois: Scientific Software International.

Vale, C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement,* 10(4), 333-344.

van der Linden, W.J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika,* 65, 437-456.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International, Inc.