ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);98-111

A Comparison of IRT Vertical Scaling Methods in Determining the Increase in Science Achievement*+

Fen Başarısındaki Artışın Belirlenmesinde Madde Tepki Kuramına Dayalı Dikey Ölçekleme Yöntemlerinin Karşılaştırılması

Aylin ALBAYRAK SARI** Hülya KELECIOGLU ***

Abstract
This study is based on a vertical scaling implemented with reference to the Item Response Theory, and involves a comparison of vertical scaling results obtained through the application of proficiency estimation methods and calibration methods. The vertical scales thus developed were assessed with reference to the criteria of grade-to-grade growth, grade-to-grade variability, and the separation of grade distributions. The data used in the study pertains to a dataset composed of a total of 1500 students from twelve primary schools in the province of Ankara, characterized by different levels of socio-economic cultural development. The comparison of the findings pertaining to the first and the second sub-problems reveals that the mean differences found through separate calibration were lower than those applicable to concurrent calibration, while the standard deviation found in the case of separate calibration were again lower than the values established through concurrent calibration. Furthermore, the scale of impact in the case of separate calibration was again lower than the values applicable to concurrent calibration. The results reached for all three criteria, using the concurrent calibration method were ranked in the order ML < MAP < EAP, with ML leading to the lowest value while EAP producing the highest one. In case of separate calibration, on the other hand, the ranking of results was found to vary with reference to the criteria applied.

Key words: Item response theory, vertical scaling, calibration methods, proficiency estimation methods.

Öz
Bu araştırmada Madde Tepki Kuramına dayalı dikey ölçekleme çalışması yürütülmüş, kalibrasyon yöntemleri ve yetenek kestirim yöntemleri kullanarak elde edilen dikey ölçekleme sonuçları karşılaştırılmıştır. Elde edilen dikey ölçekler, bir sınıf düzeyinden diğer sınıf düzeyine olan büyüme, sınıf düzeyleri arasındaki çeşitlilik ve düzey dağılımlarının ayrımı kriterlerine göre değerlendirilmiştir. Çalışmanın verileri Ankara ili farklı sosyoekonomik kültüre sahip on iki ilköğretim okulundan toplam 1500 öğrenciden toplanmıştır. Birinci ve ikinci alt probleme ait elde edilen bulgular karşılaştırıldığında, ayrı kalibrasyon ile elde edilen ortalama farkların eş zamanlı kalibrasyon ile elde edilen ortalama farklarından daha düşük olduğu, ayrı kalibrasyon ile elde edilen standart sapma değerlerinin genel olarak eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu ve ayrı kalibrasyon ile elde edilen etki büyüklüğü değerlerinin eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu görülmektedir. Eş zamanlı kalibrasyon yöntemi ile her üç kriter için de elde edilen sonuçların ML < MAP < EAP şeklinde sıralandığı; en küçük değerlerin ML, en büyük değerlerin ise EAP ile elde edildiği görülmektedir. Ayrı kalibrasyon da ise sonuçların sıralamalarının kullanılan kriterlere göre farklılaştığı görülmektedir.

Anahtar Kelimeler: Madde tepki kuramı, dikey ölçekleme, kalibrasyon yöntemleri, yetenek kestirim yöntemleri.

* This study is a part of Aylin Albayrak Sarı's doctoral dissertation titled "A Comparison of IRT Vertical Scaling Methods in Determining of the Increase in Achievement of Science Education" and conducted under the supervision of Professor Hülya Kelecioğlu.

+ This study was supported by Hacettepe University Scientific Research Projects Coordination Unit (Project Nu: 014 T03 700 001-587).

** Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: aylinalb@hacettepe.edu.tr
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: hulyaebb@hacettepe.edu.tr

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 1, İlkbahar 2017, 98-111.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 1, Spring 2017, 98-111.
DOI: 10.21031/epod.286221
Received: 17.01.2017
Accepted:03.03.2017

INTRODUCTION

Exams applied at schools serve for a wide range of objectives. When deciding on the school a student will attend, or setting the test score a candidate is expected to have for admission for a university, deciding on what to do to enhance the education system, and assessing the changes in educational practices, information derived from exams is used (Kolen, & Brennan, 2004).

In order to ascertain the level of change in academic development from one year to the next, developmental scale scores established by converting the scores pertaining to students at different levels of class into a common scale is used (Kolen, & Brennan, 2004). An awareness of the level of development through the years can provide dependable knowledge about the continuity of success, whereupon improvements at the student and class level can be effected. Large-scale assessments covering the period from K-12 grade involved numerous studies to assess the academic achievement levels of the students. It is necessary to develop a single scale score for all students' performances in all levels for reviewing and comparing academic development through the years and presenting all test scores in a single scale regardless of the year.

The fundamental problem regarding the level of academic development from one year to the next is the differences in the level of difficulty of tests, as well as their contents, even if the general topic may be the same. In order to overcome this issue, a common set of items are directed to students from consecutive years of education and the scores of students at different proficiency levels are converted into a common scale by using these items.

The process of establishing a link between the scores received in tests applied to different years is called vertical scaling (Kolen, & Brennan, 2004; McBridge, & Wise, 2001). The primary reason of applying scaling on test batteries is to provide a developmental scale score to the test developers to enable monitoring the progress in students' achievement levels (Loyd, & Hoover, 1980).

Different data collection designs, scaling methods, calibration methods, proficiency estimation methods or evaluation criteria can be applied in vertical scaling processes. The researchers would be required to make certain decisions about the designs and methods to be used in the scaling process. Such decisions were observed to have an impact on vertical scaling, and therefore the patterns indicating the change in the achievement levels of students (Tong, & Kolen, 2007). There is a brief discussion of the designs and methods chosen for this study.

Data Collection Designs

In equating, the data collection design is often called the "scaling design" (von Davier, & Wilson, 2008). Non-equivalent groups anchor test design, scaling design, and equal-to-group design are the most common used designs in vertical scaling. As the non-equivalent groups anchor test design is used in the present study, the following section will provide a brief description of the method.

The non-equivalent groups anchor test design enables the comparison of the performance of groups with reference to anchor items by building on the overlapping structure of test batteries in elementary education. For each grade, a test compatible with the level of the grade would be developed, and each such test would be applied only to the relevant grade. The test-takers' level of success with the anchor items are then used to establish the level of growth from one year to the next (Kolen & Brennan, 2004). As the design is applied on two non-equivalent groups, it is called non-equivalent groups anchor test (or anchor item) design (NEAT) (von Davier, Holland, & Thayer, 2004). Where anchor items are chosen correctly, this design helps reduce the equating error in the scaling (Hambleton, Swaminathan, & Rogers, 1991; Holland, & Dorans, 2006).

Scaling Design

Each equating method is based on a distinct theory and assumption. The equating methods are categorized as methods based on the Classical Test Theory (CTT) or on the Item Response Theory (IRT), with reference to the underlying theoretical framework.

Equating based on IRT involves the development of a mathematical relationship between the scores in two distinct forms of a test (Dongyang, 2009). Equating methods based on IRT are developed on the basis of the assumption of the existence of a mathematical function defining the relationship between the respondents' proficiency level (θ) and the probability to provide a correct response (Kolen & Brennan, 2004). Understanding, implementing, and explaining IRT methods are harder compared to CTT methods; yet IRT methods are more flexible (Harris, 2003).

One-parameter logistic model, two-parameter logistic model, and three-parameter logistic models may be applied with reference to the scale, in case of items scored on a binary scale (1-0). The present study applies a two-parameter logistic model (2-PLM).

Calibration Methods

When NEAT design is used in vertical scaling, the anchor items enable the establishment of a shared scale linking the test levels of different grades. With NEAT design, IRT parameters are either estimated for each test level by running the program separately, or estimated concurrently as the program is ran only once (Kolen, & Brenan, 2004). These calibration methods are called concurrent and separate calibration methods (Meng, 2007).

Concurrent calibration: Data pertaining to all grades is calibrated at once, to produce a vertical scale in concurrent calibration. The item parameters of the forms are estimated on the basis of the assumption that anchor items present the same item parameters for consecutive grades (Meng, 2007). In this context, the first thing to do is to set a reference grade, followed by the development of a scale with a mean of 0 and standard deviation of 1, pertaining to the scaled proficiency estimations for consecutive grades (Çetin, 2009). The item parameters for the anchor items included in the target test are estimated once again after adjustment to the values of the reference test. The item parameters pertaining to anchor items are known, while IRT calibrations are used to place non-anchor items of the target test with reference to the reference test scale (Meng, 2007).

Separate calibration: In separate calibration, the item parameters are calculated separately for each grade. As the item and proficiency parameters established separately for two different test forms have different scales, they are not readily comparable. With a view to enabling comparisons, a grade is chosen as the reference level, and θ scale is set as the starting scale for a grade. Item and proficiency parameters' estimation are used to place on the starting scale by using a series of linear conversions, with reference to the anchor items in the NEAT design (Kolen, & Brennan, 2004). Numerous linking procedures were developed in order to place the results obtained through the separate calibration on a single shared scale. The studies comparing various equating methods proposed in the literature recommend the use of Haebara and Stocking Lord (SL) methods utilizing item and test characteristics curves, instead of moment methods applying item parameters (Hanson, & Béguin, 2002; Kim, & Kolen, 2006; Kolen, & Brennan, 2004). Furthermore studies note that SL method generates less error compared to alternative methods (Hanson, & Béguin, 2002; Karkee, & Wright, 2004; Kim, 2007). Therefore, the present study applied Stocking Lord method as a characteristic curve equating method.

Furthermore, the present study compares the results obtained through scaling via both concurrent and separate calibration.

Proficiency Estimation Methods

Once the item parameters are converted into a common scale using an appropriate calibration method, the methods for estimating proficiency level should be decided. Total score or pattern scoring can be used when applying θ proficiency level estimation with reference to item response theory. The total score method, which offers a more practical and simpler approach, is used more frequently compared to the pattern scoring method. However, its error rate is larger compared to pattern scoring, while the amount of information it provides is smaller (Tong, & Kolen, 2010). For proficiency estimation regarding the binary items coded as 1-0 in IRT, often three distinct proficiency estimation methods are used. These are Maximum Likelihood (ML), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP) estimation methods. The present study provides a comparison of the results achieved through all three proficiency estimation methods.

Evaluation Criteria

The final stage of the scaling study involves the comparison of the results obtained. The normative characteristics of developmental scale scores constitute the subject matter of numerous studies. The characteristics of the scale scores are compared in order to be able to compare the results of the vertical scaling analysis. These characteristics refer to grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. Grade-to-grade growth is assessed with reference to mean difference between consecutive grades, grade-to-grade variability is assessed with reference to standard deviation between consecutive grades, and separation of grade distributions are interpreted with reference to the effect size index proposed by Yen (1984) (Kim, 2007; Kolen, & Brennan, 2004). The present study provides a comparison of the results through all three evaluation criteria.

Purpose of the Study

The literature has not yet to come up with a common view about which method reveal the best and most accurate depiction of the increase in the level of the students' achievement. Nevertheless, vertical scaling is used by numerous test developers, and every test developer determine its own vertical scaling processes (Tong, & Kolen, 2007).

Vertical scaling as a means of revealing the development of students' achievement from one grade to the next, has subsequently became an important field, and there is an increase in the number of the vertical scaling studies. The present study can provide a model about monitoring of the development in terms of students' achievement levels.

A glance at the literature reveals the rarity of studies based on real data, while studies based on simulated data are more common. The present study, on the other hand, is based on the results of science achievement tests applied with 1500 students enrolled in six different schools. In this vein, the study is expected to contribute to the literature as a model based on real data.

The purpose of the study is to implement a vertical scaling analysis based on the item response theory, and to come up with a comparison of the developmental scale scores established through the application of calibration methods (separate and concurrent calibration) and estimation methods (maximum likelihood, maximum a posteriori, and expected a posteriori estimation), with reference to the mean, standard deviation and effect size. That is why the study discusses the grade-to-grade growth, grade-to-grade variability, and separation of grade distribution characteristics pertaining to developmental scale scores. Mean and mean differences were employed to assess grade-to-grade growth, standard deviation figures for each grade were used to assess the grade-to-grade variability, and effect size were analyzed to assess the separation of grade distribution.

Research Questions

This study maintains vertical scales over three forms and investigated the question "How does the evaluation criteria vary by using various calibration methods and proficiency estimation methods in terms of vertical scaling on the basis of item response theory?". Specifically, the research questions to be investigated in line with this problem statement are as below:

1. How do;
    a. grade-to-grade growth,
    b. grade-to-grade variability, and
    c. separation of grade distribution

vary with respect to maximum likelihood, maximum a posteriori, and expected a posteriori estimations using concurrent calibration?

2. How do;
    a. grade-to-grade growth,
    b. grade-to-grade variability, and
    c. separation of grade distribution

vary with respect to maximum likelihood, maximum a posteriori, and expected a posteriori estimations using separate calibration?

METHOD

Type of Study

Because the existing methods and techniques in the research were tested through real data, and since the aim was to contribute to theoretical studies by designating the methods with minimum error, the research is a fundamental study (Creswell, 2013).

Participants

The participants of the study consist of 6^{th}, 7^{th}, and 8^{th} grades. The data used in the study were gathered from a total of 1500 students from 12 distinct schools; two from each of the Altindag, Cankaya, Golbasi, Kecioren, Sincan, and Mamak districts of Ankara province.

The science achievement test applied was developed using items selected out of Placement Exam (SBS), High School Entrance Examination (OKS), and Free Boarding and Scholarship Examination (PYBS) applied between the years 2008-2012 by checking the item discrimination and item difficulty indices, whereupon the items were compiled to achievement tests of 40 items for each of the three grades. Ten items were identified as anchor items to enable chain scaling between consecutive grades. While Hambleton, Swaminathan and Rogers (1991) note that 20% of the overall test would be a sufficient guideline to establish the number of anchor items, many studies note that increase in the number of anchor items would help reduce the standard deviation regarding the assessment sought through the test (Boughton, Lorie, & Yao, 2005; Kim, Lee, Kim, & Kelley, 2009). Therefore, the present study employed an anchor item ratio of 25% of the total number of items.

Research Design

In this research, the non-equivalent groups anchor item design was used. Even though the design is one of the most frequently employed ones, it is also one of the most flexible and most complex

designs (Sinharay, & Holland, 2007). Even though it is a design preferred on practical grounds, it is also less restrictive compared to other designs (Zhu, 1998).

Data Analysis

Before running the analyses, data was subjected to preprocessing to remove incomplete or missing data from the dataset. Furthermore, the scores received from the science achievement test were checked for unidimensionality, local independence, and model-data fit compliance among major IRT assumptions.

When unidimensional Item Response Theory (IRT) is used for equating, it is necessary to test the unidimensionality assumption for the tests (Hambleton, & Swaminathan, 1985). In order to test the unidimensionality assumption of the item response theory, confirmatory factor analysis (CFA) was applied to all three grade levels of the science tests given to students, leading to the testing of the model for a significance level of 0.05. Numerous goodness of fit indices are used in order to evaluate the model-data fit. Among these, the most frequently used indices, namely Chi-Squared Test ($\chi 2$ / sd), Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Comparative Fit Index (CFI), and Normed Fit Index (NFI) were checked. The obtained results are presented in Table 1.

Table 1. Good Fit Indices Calculated Through Confirmatory Factor Analyses for Science Test

Level of Fit	Perfect Fit Value	Allowable Fit Value	Model Value		
			6th Grade	7th Grade	8th Grade
χ^2 / sd	$0 < \chi^2 / sd \leq 2$	$2 < \chi^2 / sd \leq 5$	1.76	2.35	1.98
RMSEA	$0 < RMSEA < 0.05$	$0.05 < RMSEA < 0.10$	0.05	0.08	0.05
GFI	$0.95 \leq GFI \leq 1$	$0.90 \leq GFI \leq 0.95$	0.93	0.93	0.94
AGFI	$0.90 \leq AGFI \leq 1$	$0.85 \leq AGFI \leq 0.90$	0.92	0.90	0.95
CFI	$0.97 \leq CFI \leq 1$	$0.95 \leq CFI \leq 0.97$	0.97	0.97	0.98
NFI	$0.95 \leq NFI \leq 1$	$0.90 \leq NFI \leq 0.95$	0.97	0.94	0.95

(Ref.: Schermelleh-Engel, Moosbrugger & Müller, 2003)

A review of the goodness of fit indices obtained through CFA analysis and presented in Table 1 reveals that the model presents a high level of fit for all three grades, and the model meets the requirements of the unidimensionality assumption. Based on the CFA analysis, it can be said that data meets the unidimensionality assumption; hence the science achievement test assesses a single feature in all grades involved.

Local independence means that a response given to each item is independent from others, and the possibility of giving a positive answer to an item is not affected by other items. When the proficiency level is fixed, the correlation between items is expected to approach to zero. With a view to meeting the requirements of the local independence assumption, where just a single proficiency is required for responding all items, these items are considered unidimensional (Nandakumar, 1994). The compliance with the unidimensionality assumption can provide evidence regarding the local independence assumption (Hambleton, Swaminathan, & Rogers, 1991; Lord, & Novick, 1968). Given the fact that the present study meets the requirements of the unidimensionality assumption, it is also deemed to have met the requirements of the local independence assumption.

Once the assumptions were tested in accordance with the Item Response Theory, model-data fit was checked in order to identify the model offering the highest level of fit with the data set. The fit statistics calculated through separate calibrations for each grade revealed a state of affairs wherein

the 1 Parameter Logistics Model (PLM) and 2 PLM had model-data fit, while no model-data fit was observed for 3 PLM. Therefore, the analyses were applied in line with 2 PLM model.

FINDINGS and INTERPRETATION

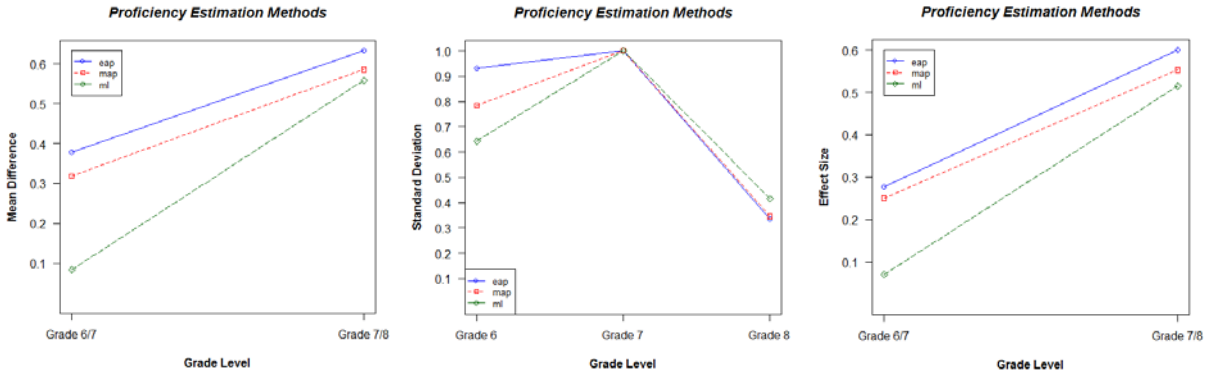
The findings of the study and the results obtained with reference to grade levels, calibration methods, and proficiency estimation methods employed were reviewed in light of mean, standard deviation, and effect size criteria.

In order to come up with an answer to first sub-problem, data pertaining to all grade levels were compiled in a single file, and all data were calibrated concurrently, using the software BILOG-MG 3. Concurrent calibration method was applied to estimate the item and proficiency parameters for each grade. The θ proficiency level means, mean differences, standard deviations and effect size values were established on the basis of ML, EAP and MAP proficiency estimation methods. The values thus calculated are presented below, in Table 2.

Table 2. Results of ML, EAP, and MAP Proficiency Estimation Obtained for Science Test through Concurrent Calibration Method

	Grade	ML	EAP	MAP
Mean	6	-0.084	-0.379	-0.318
	7	0.000	0.000	0.000
	8	0.558	0.633	0.585
Mean difference	7-6	0.084	0.379	0.318
	8-7	0.558	0.633	0.585
Standard deviation	6	0.643	0.930	0.785
	7	1.000	1.000	1.000
	8	0.415	0.336	0.346
Effect size	7-6	0.0709	0.2777	0.2505
	8-7	0.5154	0.6000	0.5530

Table 2 presents the evaluation criteria values for each grade. The graphs pertaining to these values are shown below, in Graph 1.



(a) (b) (c)

Graph 1. Graphs of Values Obtained Through the Concurrent Calibration Method: (a) Mean Differences, (b) Standard Deviations, (c) Effect Size.

As shown in both Table 2 and Graph 1 reveals, the means calculated through concurrent calibration on the basis of the data from the science test suggest that the proficiency level of the students increase as they progress from grade 6th to 8th. The review of mean differences with a view to ascertaining the criteria of development between individual grades suggests that the highest mean difference figures were observed with EAP, while the lowest ones were achieved with ML method.

The review of standard deviations, to assess the variability criteria between individual grades, on the other hand, reveals that the standard deviation fell as one moved from 6th grade to 8th, and the highest standard deviation was observed with EAP, while ML produced the lowest ones. As 7th grade was chosen as the reference year, all estimation methods stipulated a standard deviation of one (1) for that grade.

The analysis of effect sizes, with a view to evaluate the differentiation criteria between level distributions, reveals that effect size grew from 6th grade to 8th, with the largest effect sizes were observed with EAP, while the lowest ones were obtained with ML method. An analysis of the figures in Table 2 reveals that the effect size changes between the 6th and 7th grade can be considered small, while the one between the 7th and 8th is medium.

These findings run in parallel to the studies by Tong and Kolen (2010) and Kim (2007), using concurrent calibration method. Furthermore, the studies by Meng, Kolen and Lohman (2006) and Tong (2005) also found, in a similar vein, that the smallest effect size value was obtained through ML estimation.

In order to come up with an answer to second sub-problem, data for each grade level were calibrated separately using 2PLM. Item and proficiency parameters were calculated using BILOG-MG 3 software. In order to present the parameter estimations for each grade on the scale for the 7th grade, which is accepted as the reference level, the ST (Hanson, Zeng, & Chien, 2004) software, which is calculating IRT scaling constants and written in C programming language, was used. And also, Stocking Lord method was used to estimate the gradient and intersection values as a characteristics curve method.

Quadrature points are used for conversions applying Stocking Lord method. The analyses required for the calculation of Quadrature points were affected using the icl_win software. The quadrature points established thus were added to codes, to come up with SL conversion.

SL method was applied using the test-characteristic curves. The slope and intersection values produced are presented below in Table 3.

Table 3. Constants A and B calculated for Stocking Lord Conversion

Grade	A (Slope)	B (Intercept)
6-7	1.121	0.767
7-8	1.574	-0.962

The conversions are effected using the constants A and B obtained through the SL conversion presented in Table 3. Since 7th grade is set as the reference level, when converting the 6th grade to the 7th, proficiency estimations are effected through the equation "$\theta_{new}=\theta_{old}$ x 1.121 + (0.767)". On the other hand, conversion of the 8th grade to the 7th is done through the equation $\theta_{new}=\theta_{old}$ x 1.574 + (-0.962). A two-step conversion is required for transition from the 8th grade to the 6th. The equation $\theta_{new}=(\theta_{old}$ x 1.121 + (0.767)) x 1.574 + (-0.962) was used for the conversion of the 8th grade. The intersection values between the 6th and the 7th grades are positive, while those between the 7th and the 8th are negative.

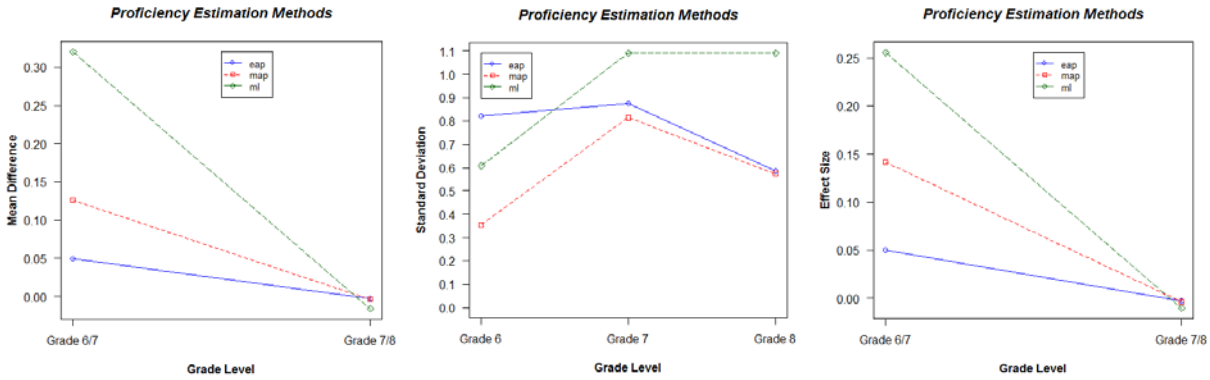
Estimation was effected using separate calibration method with the BILOG-MG 3 software using the calculated estimation values as well. The θ proficiency level means, mean differences, standard

deviations and effect size values were established on the basis of ML, EAP and MAP proficiency estimation methods. The results are presented below in Table 4.

Table 4. The Results of ML, EAP, and MAP Proficiency Estimations Obtained for Science Test through Separate Calibration Method

	Grade	ML	EAP	MAP
Means	6	-0.317	-0.058	-0.147
	7	0.007	0.002	-0.021
	8	-0.009	-0.005	-0.025
Mean difference	6-7	0.320	0.060	0.126
	7-8	-0.016	-0.007	-0.004
Standard deviation	6	0.608	0.822	0.354
	7	1.091	0.874	0.814
	8	1.091	0.586	0.575
Effect size	6-7	0.2558	0.0498	0.1420
	7-8	-0.0104	-0.0030	-0.0040

Table 4 presents the evaluation criteria values for each grade. To present a clearer picture of these figures, the graphs pertaining to these values are shown below, in Graph 2.



(a) (b) (c)

Graph 2. Graphs of Values Obtained Through the Separate Calibration Method: (a) Mean Differences, (b) Standard Deviations, (c) Effect Sizes.

As seen in both Table 4 and Graph 2 reveals, the means calculated through separate calibration on the basis of the data from the science test suggest that the proficiency level of the students increase as they progress from grade 6th to 7th, and fall from grade 7th to 8th. Mean differences, which reflect the level of improvement from one grade to another allows a better understanding of this criterion. While the mean differences are positive between grades 6th and 7th, they are negative between grades 7th and 8th, and tend to fall from grade 6th to 8th. This finding can be interpreted as the fact that the 7th grade students are more successful than the 8th grade students and that the desired and expected growth from one class level to the other class level cannot be achieved. The reason for the 8th grade students being less successful than the 7th grade may be the TEOG (Basic Education to Secondary Transition) exam. The increase in students' anxiety levels may have adversely affected their success. In addition, the fact that eighth grade students have entered adolescence may have affected their psychology and achievements negatively. In the study of Briggs, Weeks and Wiley (2009), parallel to this finding, it was stated that the growth patterns did not show an increase from one year to the

other year as a linear. It seems that there are studies supporting this finding in the literature (Tong, & Kolen, 2008; Cetin, 2009; Wysel, & Reckase, 2011; Altun, 2013). In Tong and Kolen (2010)'s study, it was found that the mean difference was higher in the lower class levels and the mean difference decreased as the class level increased. Similar to the results of Tong and Kolen's (2010) study, Ito, Skykes and Yao (2008)'s and Tong and Kolen (2007)'s studies, compared vertical scaling methods, have stated that he increase in the scores of the students in the lower grade level is higher than the increase in the scores of the students in the higher grade level. As a result of the IRT analyzes the scores of the students increase and decrease according to the grade levels. In other words, the success levels of unsuccessful students are increasing in 6th grade to 7th grade, compared to the transition from 7th grade to 8th grade. And, when the estimation methods are compared, it is seen that the highest mean differences were obtained with ML, while EAP produced the lowest ones.

A glance at standard deviation figures shows that overall standard deviation between grades 6th and 8th tend to fall. While the lowest standard deviation is established with ML method, EAP produced the highest level of standard deviation.

The analysis of effect sizes indicates that in all three methods, effect sizes tend to fall towards grade 8th, with the largest effect sizes being observed when ML is applied, in contrast to the smallest ones are obtained through EAP. An analysis of the figures in Table 4 reveals that the effect size changes between the 6th and 7th grades as well as between the 7th and 8th grades can be interpreted as a weak effect. The review of the literature reveals that these findings run in parallel to those of Tong and Kolen (2007).

DISCUSSION and CONCLUSIONS

The objective of this study is to apply vertical scaling based on item response theory, leading to a comparison of calibration methods and proficiency estimation methods, and the developmental vertical scale scores calculated with reference to the mean, standard deviation, and effect size values.

The means calculated through concurrent calibration on the basis of the data from the science test showed that the proficiency level of the students increase as they progress from grade 6th to 8th. The mean differences for all three grades present a picture where largest differences are produced with EAP method. A glance at standard deviation figures shows that standard deviation between grades 6th and 8th tends to fall, and the lowest standard deviation value is established with ML method. Effect size picture suggests an increase from grade 6th to 8th, with the largest effect size values being produced with EAP method.

When the separate calibration method is applied as another calibration, the developmental scale scores present an increase in the means from grade 6th to 8th, while mean differences fall approaching from 6th to grade 8th. The highest mean difference was observed with EAP method. The mean differences generated through separate calibration were also notably lower than those generated through concurrent calibration. Standard deviation picture presents falling rates as one move from grade 6th towards 8th. The lowest standard deviation was observed with ML method. The standard deviation values calculated in separate calibration were generally lower than those produced through concurrent calibration. On the effect size front, it is observed that the effect sizes values decreasing from 6th grade to 8th grade. The highest effect size was observed with ML method. The effect size values calculated in separate calibration were lower than those produced through concurrent calibration.

The comparison of the findings pertaining to the first and the second sub-problems reveals that the mean differences found through separate calibration were lower than those applicable to concurrent calibration, while the standard deviation found in the case of separate calibration were again lower than the values established through concurrent calibration. Furthermore, the scale of impact in the case of separate calibration was again lower than the values applicable to concurrent calibration. The results reached for all three criteria, using the concurrent calibration method were ranked in the order ML < MAP < EAP, with ML leading to the lowest value while EAP producing the highest one. In

case of separate calibration, on the other hand, the ranking of results was found to vary with reference to the criteria applied.

The conclusions reached through the study reveal that vertical scaling is a complex process, and that there is no single all-applicable method. Since there is no single method supported by a wide-ranging consensus, taking into account the complexities of the methods applied and the results of the analyses, it is recommended that the researcher should decide on the method to apply, within the context of her specific study. The interactions between the issues discussed in this process can have an impact on the results of vertical scaling, and hence on the interpretations about the ongoing development of the students' achievements, one can recommend effective comparisons employing a range of methods, to lead to decisions regarding the achievements of students. Hanson and Béguin (2002) also emphasized that no single all-applicable method can be designated, and that comparing results through a combination of various equating methods under different conditions is the way to go.

Such an analysis should actually be considered an inherent part of the overall vertical scaling process. Test developers and users can be recommended to work on the process of equating the observed and actual scores in the final stage of the vertical scaling process, with the review of factors affecting observed scores.

Achievement levels of the students were observed to increase as one move from earlier grades to subsequent ones. However, further studies may be needed to assess whether such increases are at required levels or not. In order to ascertain the level of change students experience from one grade to another, vertical scaling practices are crucial. Vertical scaling assessments can be recommended to review the students' achievements at the K-12 level.

In the present study, test length (40 items), number of anchor items (10), sample size (1500), and applied model (2PLM) were fixed, and not subjected to analysis as determining factors or independent variables. Other studies can use these as variables in their own right, and investigate their impact on vertical scaling results as well. It is also possible to carry out a longitudinal study to review the achievement levels of individual students through extended years, followed up by an analysis on the basis of data from such longitudinal study. Since there is no single and exact criteria to assess the applicability of the methods employed in vertical scaling, the researchers are recommended to use more than one evaluation criteria (mean, mean differences, standard deviation, effect sizes, vertical distance, root-mean square error of approximation (RMSEA) and bias values) when comparing scaling results.

REFERENCES

Altun, A. (2013). *Dikey ölçeklemede madde tepki kuramına dayalı farklı kalibrasyon ve yetenek kestirim yöntemlerinin karşılaştırılması* (Unpublished Doctoral Thesis). Ankara: Hacettepe University.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008, April). *Vertical scaling in value-added models for student learning*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Boughton, K. A., Lorie, W., & Yao, L. (2005). *A multidimensional multi-group IRT models for vertical scales with complex test structure: An empirical evaluation of student growth using real data*. Paper presented at the annual meeting of the National Council on Measurement in Education, Monreal, Canada.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative and mixed methods approaches* (4th edition). University of Nebraska, Lincoln: Sage.

Cetin, E. (2009). *Dikey ölçeklemede klasik test ve madde tepki kuramına dayalı yöntemlerin karşılaştırılması* (Unpublished Doctoral Thesis). Ankara: Hacettepe University.

Dongyang, L. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design* (Unpublished Doctoral Thesis). University of Maryland.

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

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

Hanson, B. A., & Béguin, 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*, 3-24.

Hanson, B. A., Zeng, L., & Chien, Y. (2004). *ST: A computer program for IRT scale transformation* [Computer software]. Retrieved January 24, 2005, from http://www.education.uiowa.edu/casma.

Harris, D. J. (2003). Equating the multistate bar examination. *The Bar Examiner, 72*(3), 12-18.

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

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurment in Education, 21,* 187-206.

Karkee, T. B. & Wright, K. R. (2004). *Evaluation of linking methods for placing three-parameter logistic item parameter estimates onto a one-parameter scale*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Kim, J. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales* (Unpublished Doctoral Thesis). University of Iowa.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kim, J., Lee, W. C., Kim, D., & Kelley, K. (2009). *Investigation of vertical scaling using the Rasch model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

McBridge, J., & Wise, L. (2001). *Developing the vertical scale for the Florida comprehensive assessment test (FCAT).* Paper presented at the annual meeting of the Harcourt Educational Measurement, San Antonio, Texas.

Meng, H (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.* (Unpublished Doctoral Thesis). University of Iowa, Iowa.

Meng, H., Kolen, M. J., & Lohman, D. (2006). *An empirical investigation of IRT scaling methods: How different IRT models, parameter estimation procedures, proficiency estimation methods, and estimation programs affect the results of vertical scaling for the cognitive abilities test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Tong, T. (2005). *Comparison of methodologies and results in vertical scaling for educational achievements tests* (Unpublished Doctoral Thesis). University of Iowa, Iowa.

Tong, Y., & Kolen, M. (2007). Comparison of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.

Tong, Y., & Kolen, M. (2008, March). *Maintenance of vertical scales.* Paper presented at the National Council on Measurement in Education, New York City.

Tong, Y., & Kolen, M. (2010). Scaling: An ITEMS module. *Educational Measurement: Issues and Practice, 29*(4), 39-48

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating.* New York: Springer.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of Item Response Theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11-26.

Wysel, A. E., & Reckase, M. D. (2011). A graphical approach to evaluating equating using test characteristic curves. *Applied Psychological Measurement, 35*(3) 217–234.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93-111.

Zhu, W. (1998). Test equating: What, why, who? *Research Quarterly for Exercise and Sport, 69*(1), 11–23.

UZUN ÖZET

Giriş

Dikey ölçekleme sürecinde farklı veri toplama desenleri, ölçekleme yöntemleri, kalibrasyon yöntemleri, yetenek kestirimi yöntemleri ve değerlendirme ölçütleri kullanılabilir. Araştırmacıların ölçekleme sürecinde kullanılacak desen ve yöntemlere ilişkin çeşitli kararlar vermesi gerekmektedir. Bu kararların dikey ölçeklemeyi dolayısıyla da öğrenci başarısındaki gelişimi gösteren örüntüleri etkilediği görülmüştür (Tong & Kolen, 2007). Bu çalışmada veri toplama deseni olarak denk olmayan gruplarda ortak madde deseni, ölçekleme deseni olarak Madde Tepki kuramına dayalı 2 Parametreli lojistik model kullanılmıştır. Sınıf seviyelerinin ortak bir ölçeğe bağlanması için kullanılan ölçek dönüştürme kalibrasyon yöntemlerinden ayrı ve eş zamanlı kalibrasyon; madde parametrelerini kestirebilmek için kullanılan kestirim yöntemlerinden ise, Maximum Likelihood Estimation (ML) (Maksimum Olabilirlik), Expected A Posteriori (EAP) (Beklenen Önsel Dağılım) ve Maximum A Posteriori (MAP) (Maksimum Önsel Dağılım) kestirim yöntemleri kullanılmıştır. Ölçekleme çalışmasının son aşamasında ise elde edilen sonuçlar bir sınıf düzeyinden diğer sınıf düzeyine olan büyüme, sınıf düzeyleri arasındaki çeşitlilik ve düzey dağılımlarının ayrımı değerlendirme ölçütleri kullanılarak karşılaştırılmıştır.

Alan yazın incelendiğinde gerçek veri ile yapılan çalışmaların oldukça az olduğu, daha çok simülasyon verileri ile yapılan çalışmalara ağırlık verildiği görülmektedir. Bu çalışmada araştırmacılar tarafından geliştirilen fen bilgisi başarı testi 1500 öğrenciye uygulanarak toplanan gerçek veriler üzerinde analizler yürütülmüştür, böylece bu çalışmanın alan yazına katkı sağlayacağı düşünülmektedir.

Yöntem

Araştırmada var olan yöntem ve teknikler gerçek veri ve yapay veri üzerinden sınandığı ve en az hatalı yöntemler belirlenerek kuramsal çalışmalara katkı sağlaması amacı taşıdığı için araştırma temel araştırma niteliğindedir (Creswell, 2013). Araştırmada çalışma grubu 6ncı, 7nci ve 8inci sınıf öğrencilerinden oluşmaktadır. Çalışma grubu, Ankara ili Altındağ, Çankaya, Gölbaşı, Keçiören ve Mamak ilçelerinden ikişer okul olmak üzere 12 farklı okuldan toplam 1500 öğrenciden oluşmaktadır. Uygulanan fen bilgisi başarı testi için 2008-2012 yılları arasında uygulanan SBS (Seviye Belirleme Sınavı), OKS (Ortaöğretim Kurumları Seçme ve Yerleştirme Sınavı) ve PYBS (Parasız Yatılılık ve Bursluluk Sınavı) testlerinden ayırt edicilik düzeyleri ve madde güçlük indeksleri kontrol edilerek maddeler seçilmiş ve üç sınıf düzeyine uygun 40ar maddelik birer başarı testi geliştirilmiştir. Bu testlerde ardışık sınıflar arası zincirleme ölçeklemeyi sağlayacak 10'ar madde ortak madde olarak belirlenmiştir. Hambleton, Swaminathan ve Rogers (1991), ortak maddelerin sayısının testin tamamının %20'si kadar olmasının uygun olduğunu belirtirken, birçok araştırmada ortak madde sayısındaki artışın testteki ölçmenin standart hatasını azalttığını belirtilmektedir (Boughton, Lorie & Yao, 2005; Kim, Lee, Kim & Kelley, 2009). Bu nedenle bu çalışmada toplam madde sayısının %25'i kadar ortak madde kullanılmıştır. Bu araştırmada denk olmayan gruplarda ortak madde deseni kullanılmıştır. Bu desen uygulamada yaygın olarak kullanılan desenlerden biri olmakla birlikte, en esnek ve en karmaşık desenlerden biridir (Sinharay & Holland, 2007). Pratiklik açısından tercih edilen bir yöntem olmakla birlikte, diğer desenlere göre de daha az sınırlayıcıdır (Zhu, 1998).

Sonuçlar ve Tartışma

Analizler yapılmadan önce veri temizleme yapılarak eksik ve kayıp veriler veri setinden çıkarılmış ve fen bilgisi başarı testinden elde edilen puanların MTK varsayımlarından tek boyutluluk, yerel bağımsızlık ve model veri uyumu kontrol edilmiştir. Birinci ve ikinci alt probleme ait bulgular incelendiğinde; dikey ölçekleme analizinde farklı kalibrasyon yöntemlerinden elde edilen sonuçlar

karşılaştırıldığında; ayrı kalibrasyon ile elde edilen ortalama farkların eş zamanlı kalibrasyon ile elde edilen ortalama farklarından daha düşük olduğu, ayrı kalibrasyon ile elde edilen standart sapma değerlerinin genel olarak eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu ve ayrı kalibrasyon ile elde edilen etki büyüklüğü değerlerinin eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu görülmektedir. Eş zamanlı kalibrasyon yöntemi ile her üç kriter için de elde edilen sonuçların ML < MAP < EAP şeklinde sıralandığı; en küçük değerlerin ML, en büyük değerlerin ise EAP ile elde edildiği görülmektedir. Ayrı kalibrasyon da ise sonuçların sıralamalarının kriterlere göre değiştiği görülmektedir. Araştırma bulgularına göre, dikey ölçekleme sürecinin karmaşık bir süreç olduğu ve tek bir doğru yöntem olmadığı görülmektedir. Üzerinde hemfikir olunan doğru bir yöntem olmadığı için, uygulanan yöntemlerin karmaşıklığı analizlerin sonuçları göz önünde bulundurularak en uygun yöntemi yine araştırmacı araştırmasına uygun olarak belirleyebilir. Bu süreçte ele alınan koşulların birbiriyle etkileşimi dikey ölçekleme sonucunu dolayısıyla öğrenci başarısının gelişimine yönelik yapılacak yorumları etkileyebileceği için öğrenci başarıları hakkında karar verirken farklı yöntemlerin de kullanılarak karşılaştırma yapılması önerilebilir. Hanson ve Béguin (2002) de tek bir doğru yöntem belirtilemeyeceği, farklı koşullarda doğru yöntemi belirleyebilmek için eşitleme yöntemlerini bir arada kullanarak, sonuçlarını karşılaştırmanın etkili olacağını vurgulamışlardır. Öğrenci başarılarının genel olarak ardışık sınıf seviyesi arttıkça arttığı görülmüştür, fakat bu artışın istendik düzeyde olup olmadığını değerlendirebilmek için çalışmalar yapılabilir. Öğrencilerin yıldan yıla başarılarındaki değişimin belirlenebilmesi için dikey ölçekleme uygulamaları oldukça önemlidir. Öğrencilerin K-12 seviyesinde başarılarının takibi için dikey ölçekleme çalışmalarının başlatılması ve yürütülmesi önerilebilir.