Assessment of Achievement and Growth by Vertical Scaling: Comparison of Vertical Scaling Methods[1]

Aylin ALBAYRAK SARI[2] & Hülya KELECİOĞLU[3]

ABSTRACT

In this study, item response theory-based vertical scaling was conducted, and the vertical scaling results obtained by using calibration methods and proficiency estimation were compared. The obtained vertical scales were evaluated according to the criteria of grade-to-grade growth, grade-to-grade variability, and the separation of grade distributions. For this study, the data was simulated by R program. According to the results, the mean differences in both the methods of concurrent calibration and separate calibration methods increased compared to the levels observed in 8th grade, and that the largest value was obtained through the EAP method. The mean differences obtained through separate calibration were lower than those obtained through concurrent calibration. When standard deviation values were compared, it was observed that the largest values were obtained through EAP in both calibration methods, and that the standard deviation values obtained in both methods were generally close to each other. When effect size values were examined, in both calibration methods, the effect size values increased toward the 8th grade. The effect size values obtained through separate calibrations were lower than those obtained through concurrent calibration. The results generated by all three proficiency estimation methods were similar to each other.

Key Words: Item response theory, Vertical scaling, Calibration methods, Proficiency estimation methods, Artificial data

DOI Number: http://dx.doi.org/10.12973/jesr.2016.62.2

[1] This study is based on Aylin Albayrak Sarı's doctoral dissertation titled "A Comparison of IRT Vertical Scaling Methods in Determining of the Increase in Achievement of Science Education". This study was supported by Hacettepe University Scientific Research Projects Coordination Unit. Project Number: 014 T03 700 001-587.
[2] Corresponding Author: Dr., Hacettepe University, Faculty of Education - aylinalb@hacettepe.edu.tr
[3] Prof. Dr. - Hacettepe University, Faculty of Education - hulyaebb@hacettepe.edu.tr

INTRODUCTION

Information obtained from tests are used to determine which school/university a student should attend, establishing the test scores students should have to be accepted to a university, discussing what should be done to improve the education system, and evaluating changes in educational practices (Kolen & Brennan, 2004). Implementers of the test also want to be able to compare test scores received from different forms and different groups, and to replace one test with another, should the need arise. However, to be able make comparisons, the raw scores need to be converted into standard scores (AERA/American Educational Research Association, APA/American Psychological Association & NCME/National Council on Measurement in Education, 1999). By converting the scores obtained from different tests prepared with the aim of measuring similar characteristics into a common scale, these scores can be compared and such scores are called comparable scores (Angoff, 1971).

One of the fields in which comparable scores are most frequently used is developmental scale scores. In order to identify the year-to-year progress of academic development, developmental scale scores are used; obtained by converting the scores of students from different grades into a common scale (Kolen & Brennan, 2004). The main reason to conduct scaling processes in test batteries is to provide the developmental score scale to the test developers so that they can monitor the progress in the achievements of students (Loyd & Hoover, 1980). The basic problem in identifying the year-to-year progress in academic development is the fact that the groups that receive the test and the difficulty levels and contents of tests are different. In order to overcome this problem, common questions are asked of the students in consecutive grades and the scores of students at different proficiency levels are converted into a common scale. In this case, a linking procedure called vertical scaling is applied (Kolen & Brennan, 2004).

During the process of vertical scaling, different data collection designs, scaling methods, calibration methods, proficiency estimation methods, and evaluation criteria can be used. The researchers are required to make certain decisions about the designs and methods to be used in the scaling process. It was seen that such decisions affected vertical scaling, and accordingly, the patterns showing the progress in the achievement of students (Tong & Kolen, 2007). In this study, the data collection design used was the non-equivalent groups anchor test design and the scaling design used was the Item Response Theory-based logistic model with two parameters. While the scale conversion calibration method used to link the grades to a common scale was separate and concurrent calibrations, the estimation methods used for estimating the item parameters were Maximum Likelihood Estimation (ML), Expected a Posteriori (EAP), and Maximum a Posteriori (MAP). In the last phase of scaling study, the results obtained were *compared by using the evaluation criteria* of grade-to-grade growth, grade-to-grade variability, and the separation of grade distributions.

Although there is no common opinion in the literature regarding which method is best and most accurately reflects the increase in the achievement of students, vertical scaling is used by many test developers. However, each test developer designates by himself the development processes of vertical scaling for the scale he develops (Tong & Kolen, 2007). The subject of vertical scaling, which enables the presentation of student achievement progress in line with their increasing grades, has gained incremental importance and more studies have begun in this field. This study may serve as a model for monitoring the progress in the achievement of students.

The objective of this study is to compare different vertical scaling results obtained through calibration methods (separate and concurrent calibration) and proficiency estimation methods (maximum likelihood, expected a posteriori, maximum a posteriori) in terms of mean, standard deviation, and effect size values by carrying out item response theory *(IRT)*-based vertical scaling. In order to compare the vertical scales obtained, the criteria of grade-to-grade growth, grade-to-grade variability, and the separation of grade distributions were used. While means and mean differences were examined in order to evaluate the grade-to-grade growth, the standard deviation values for each grade were examined to evaluate the grade-to-grade variability; and effect size values were examined to evaluate the separation of grade distributions. Thus, it is thought that this study will contribute to the literature.

This research aims to answer to the question: "How does the evaluation criteria obtained using different calibration methods and different estimation methods in item response theory-based vertical scaling differ by using artificial dataset derived through simulation?" The sub-problems examined in line with this problem statement are:

1. How do the

    a. grade-to-grade growth

    b. grade-to-grade variability

    c. the separation of grade distributions

    of maximum likelihood, expected a posteriori, and maximum a posteriori proficiency estimation obtained through concurrent calibration method differ?

2. How do the

    a. grade-to-grade growth

    b. grade-to-grade variability

    c. the separation of grade distributions

    of maximum likelihood, expected a posteriori, and maximum a posteriori proficiency estimation obtained through separate calibration method differ?

METHOD

Type of Research

Because the existing methods and techniques in the research were tested through artificial data and since the aim was to contribute to theoretical studies by designating the methods with minimum error, the research is classed as basic research (Creswell, 2013).

Research Design

In this research, non-equivalent groups anchor test design was used. While this design is one of the most widely used designs in implementation, it is also one of the most flexible and most complicated designs (Sinharay & Holland, 2007. This method, which is a preferred method in terms of practicality, is also less restrictive compared to other designs (Zhu, 1998).

Research Data

The working group of the research consists of artificial data derived in line with the parameters designated for items and skills. The answers from a total of 1,500 students were simulated, with 500 students in each grade.

For each of these tests, ten items were designated as common items to enable chain-linking among consecutive classes. While Hambleton, Swaminathan and Rogers (1991) stated that the reasonable number of common items corresponds to 20% of the total items in the test, many studies state that the increase in the number of common items results in a decrease in the standard error of measurement in the test (Boughton, Lorie, & Yao, 2005; Kim, Lee, Kim, & Kelley, 2009). Thus, within this study, common items were used, the number of which correspond to 25% of the total number of items.

In line with the test of science achievement applied, a test pattern consisting of 40 items, ten of which were common with the consecutive class, was developed and the answers of 1,500 students were simulated by using the R program. What kind of results would be achieved with the item parameters was tested on the proficiency distributions designated.

Data Simulation

Prior to the analysis, the item and proficiency parameter intervals designated by the researchers were also used to derive artificial datasets in the R program. The parameter intervals of the artificial data derived are given in Table 1.

Table 1. *Parameter intervals of artificial data*

Proficiency Parameters	6th grade	7th grade	8th grade
Parameter θ	[-1<θ<0]		
		[0<θ<1]	
			[1<θ<2]
Item Parameters	6th grade	7th grade	8th grade
Parameter a	[0.5<a<1.5]		
	[0.5<a<1.5]	[1<a<2]	
		[1<a<2]	
		[1<a<2]	[1.5<a<2]
			[1.5<a<2]
Parameter b	[-3<b<3]		
	[-1<b<0]	[-1<b<+1]	
		[-2<b<+3]	
		[0<b<1]	[-1<b<2]
			[1<b<2]

The proficiency parameters for artificial datasets were set by increasing the values by one unit for each grade. Artificial data for the 6th, 7th, and 8th grades were generated with proficiency parameter intervals of [-1<θ<0], [0<θ<1] and [1<θ<2], respectively.

In this study, 2 Parameter Logistic Model (PLM) was applied in the analyses. Hence, item difficulty (parameter a) and item discrimination (parameter b) were calculated. The level of item discrimination power was set to be narrowed as the grades increased. Artificial data for 6th, 7th, and 8th grades were generated in intervals of [0.5<a<1.5], [1<a<2], and [1.5<a<2], respectively. On the other hand, item difficulty levels for the dataset were

designated separately for the common and non-common items. Parameter b for the first 30 non-common items in the 6th grade was derived in the interval of [-3<b<3]; parameter b for the last ten items common with the 7th grade was derived in the interval of [-1<b<0]; parameter b for the first ten items that were common in the 6th and 7th grades was derived in the interval of [-1<b<+1]; parameter b for the next 20 items of the 7th grade was derived in the interval of [-2<b<+3]; parameter b for the last ten items that are common in the 7th and 8th grades was derived in the interval of [0<b<1]; parameter b for the first ten items that were common in the 8th and 7th grades was derived in the interval of [-1<b<2]; and parameter b for the next 30 items of the 8th grade was derived in the interval of [1<b<2]. Parameter b for common items was increased in accordance with the grade and the artificial data was generated.

RESULTS AND INTERPRETATION

The study results obtained based on grades, calibration methods, and proficiency estimation methods were examined in regard to the criteria of mean, standard deviation and effect size.

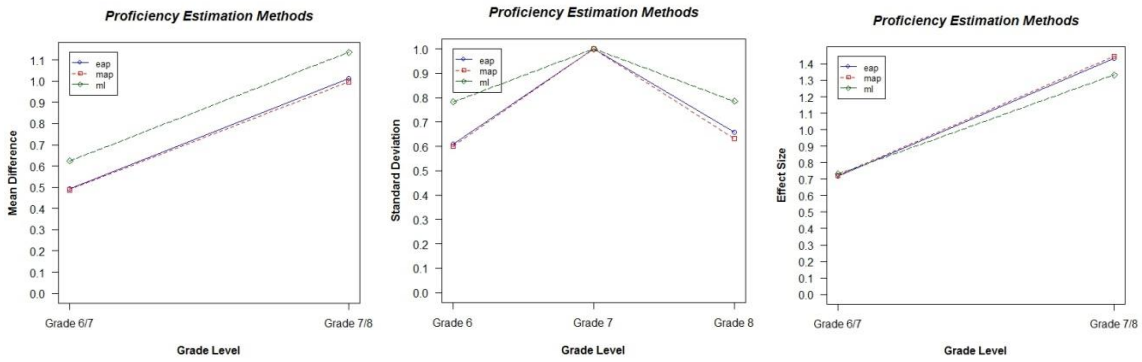
Results and Interpretations Related to the First Sub-Problem

During the data analysis phase for the first sub-problem, artificial data was derived using the parameter intervals given in Table 1. Each analysis was iterated 100 times under the conditions mentioned and the values obtained then averaged. In order to be able to make a forecast through the concurrent calibration method, codes were written in the R program, and the item and proficiency parameters for each grade were forecasted through a single analysis using the BILOG-MG 3 program. Means, inter-mean differences, standard deviations, and size effect values of the θ proficiency level obtained through proficiency estimation methods, namely ML, EAP, and MAP, were calculated. The mentioned values are given in Table 2.

Table 2. *Results of ML, EAP, and MAP proficiency estimations obtained through concurrent calibration method in the context of the artificial dataset*

	Grade	ML	EAP	MAP
Mean	6	-0.624	-0.492	-0.488
	7	0.000	0.000	0.000
	8	1.136	1.013	0.997
Mean difference	6-7	0.624	0.492	0.488
	7-8	1.136	1.013	0.997
Standard deviation	6	0.783	0.610	0.601
	7	1.000	1.000	1.000
	8	0.785	0.659	0.632
Effect size	6-7	0.732	0.719	0.722
	7-8	1.332	1.433	1.445

Table 2 shows the mean, mean difference, standard deviation, and size effect values obtained as a result of the ML, EAP, and MAP proficiency estimation for each grade. When estimating the IRT parameters through concurrent calibration, it is seen that the mean of the θ proficiency level of the 7th grade, which was selected as the reference grade, is zero and its standard deviation is 1 in all three proficiency estimations. For a better interpretation of these values, see Graph 1 for related graphics.



Graph 1. *Mean differences, standard deviations, and size effect values obtained through concurrent calibration method in the context of the artificial dataset*

As seen in both Table 2 and Graph 1, when the mean differences calculated as a result of the concurrent calibration method are examined, it is seen that the mean differences increase from the 6th grade to the 8th grade in all three proficiency estimation methods. Graph 1 shows that while the highest mean differences were obtained through the ML method, values that are close to each other were obtained through the MAP and EAP methods.

When the standard deviation values are examined in order to evaluate the criteria of grade-to-grade variability, it is seen that the standard deviation values of the 6th and 8th grades were close to each other. While the lowest standard deviation value was obtained through the MAP method, the highest was obtained through the ML method.

When the criteria of calculated size effect values are examined to evaluate the separation criteria of grade distributions, it is seen that size effect values that are close to each other are obtained in all three methods. When the values given in Table 2 are examined, the size effect between the 6th and 7th grades can be interpreted as a moderate effect, while the effect value between the 7th and 8th grades can be interpreted as a strong effect.

When the literature is examined, it is seen that these findings are similar to those obtained in the study of Tong and Kolen (2007). On the other hand, Meng, Kolen, and Lohman (2006), Meng (2007) and Tong (2005) indicated in their studies that the smallest forecasts are obtained through the ML method.

Results and Interpretations Related to the Second Sub-Problem

Artificial data derived through simulation were also used for the second sub-problem and each analysis was iterated 100 times under the conditions mentioned. In order to be able to make a forecast through the separate calibration method, codes were written in the R program, and the item and proficiency parameters for each grade were forecasted separately using the BILOG-MG 3 program. Quadrature points required for making conversions through the Stocking Lord method were calculated through the icl_win program and A and B constants were forecasted through ST program by using the quadrature points calculated. The slope and intersection values obtained are given in Table 3.

Table 3. *A and B constants obtained for the stocking-lord conversion*

Grade	A (Slope)	B (Intercept)
6-7	1.417	0.357
7-8	1.250	0.812

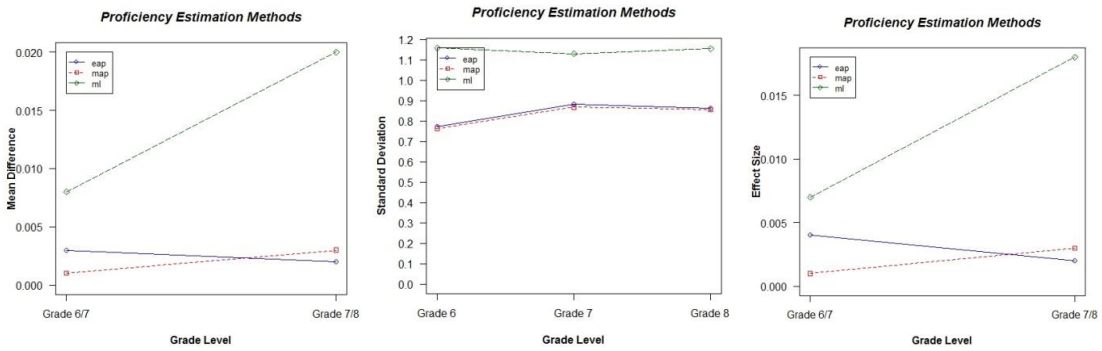
A conversion was made using the slope and intersection values given in Table 3. Since the 7th grade is the reference grade, when converting the 6th grade into the 7th grade scale, proficiency estimations are made with the equation $\theta_{new}=\theta_{prev}x1.417+(0.357)$. And when converting the 8th grade scale into the 7th grade, proficiency estimation is calculated with the equation $\theta_{new}=\theta_{prev}x1.250+0.812$). On the other hand, two conversions are required in order to convert from 8th grade to the 6th grade. In order to convert the 8th grade, the equation $\theta_{new}=(\theta_{prev}x1.417+(0.357))x1.250+(-0.812)$ was used. Intersection values are positive for all grades.

Estimation was made through a separate calibration method, using the forecast values calculated, as well as the BILOG-MG 3 program. Means, inter-mean differences, standard deviations, and size effect values of the θ proficiency level obtained through proficiency estimation methods, namely ML, EAP, and MAP, were calculated. The mentioned values are given in Table 4.

Table 4. *Results of ML, EAP, and MAP proficiency estimations obtained through the separate calibration method in the context of the artificial dataset*

	Grade	ML	EAP	MAP
Mean	6	-0.015	-0.002	0.004
	7	-0.007	0.001	0.005
	8	0.013	0.003	0.008
Mean differences	6-7	0.008	0.003	0.001
	7-8	0.020	0.002	0.003
Standard deviation	6	1.159	0.773	0.764
	7	1.130	0.884	0.868
	8	1.155	0.863	0.856
Effect size	6-7	0.007	0.004	0.001
	7-8	0.018	0.002	0.003

Table 4 shows the mean, mean difference, standard deviation, and size effect values obtained as a result of the ML, EAP, and MAP proficiency estimation for each grade. For a better interpretation of these values, the related graphics are given in Graph 2.



Graph 2. *Mean differences, standard deviations, and size effect values obtained through the separate calibration method in the context of the artificial dataset*

As seen in both Table 4 and Graph 2, when the mean differences calculated as a result of concurrent calibration method are examined, it is seen that the mean differences were very small and close to each other in all three estimation methods.

When the standard deviation values are examined in order to evaluate the criteria of grade-to-grade variability, it is seen that the standard deviation values of all three grades were close to each other and that while the results obtained through the MAP method and EAP method were very similar, the highest standard deviation value was obtained through the ML method.

When the criteria of size effect values calculated are examined to evaluate the separation criteria of grade distributions, it is seen that the size effect values are close to each other in all three methods. When the values given in Table 4 are examined, the size effect between the 6th and 7th grades and the size effect between the 7th and 8th grades can be interpreted as a weak effect.

DISCUSSION AND CONCLUSION

The objective of this current study is to compare different vertical scaling results obtained through different calibration methods (separate and concurrent calibration) and different proficiency estimation methods (maximum likelihood, expected a posteriori, maximum a posteriori) in terms of mean, standard deviation, and size effect values by carrying out item response theory-based vertical scaling.

When the findings obtained for the first and second sub-problems are examined, it is observed that the mean differences in the cases of both concurrent and separate calibration methods increased when compared to the 8th grade, and that the highest values are obtained through the EAP method. It is seen that the mean differences obtained through separate calibration were lower than those obtained through concurrent calibration. When standard deviation values were compared, it was observed that the largest values were obtained through EAP in both calibration methods, and that the standard deviation values obtained in both methods were generally close to each other. When the size effect, which is another evaluation criteria, is examined, it is seen that the size effect values increased toward the 8th grade in both calibration methods. The size effect values obtained through separate calibration were lower than those obtained through concurrent calibration, and that all three estimation methods generated results close to each other.

RECOMMENDATIONS

According to the findings of the research, the vertical scaling process is a complicated process and is not the only correct method. Since there is no single correct method on which there is common agreement, the complexity of the methods applied shall be taken into account in accordance with the results of the analysis, and the most appropriate method shall again be determined by the researcher depending on the nature of the research. Since the conditions that are taken into consideration through this process may affect the result of vertical scaling and thus the improvement of student achievement, it may be recommended that different methods are used and compared to each other when deciding on the students' achievements. Hanson and Béguin (2002) expressed that no single method can be designated and that in order to designate the correct method under different conditions, equating methods should be used together and their results should be compared in order to achieve an effective outcome.

This current study is one part of the whole vertical scaling process. Test developers and implementers are advised to examine studies about the equating process for observed-real data that is a part of the final phase of the vertical scaling process, and to examine the factors affecting the scores observed. In this current study, data was derived through simulation in the defined parameter intervals. For cases in which it is difficult to access real data, analyses can be carried out by deriving data through simulation.

In the current study, it was found that student achievement increased as the consecutive grades increased. However, further studies should be conducted in order to evaluate whether or not this increase is at the desired level. Vertical scaling implementations are quite important for defining the changes in the year-to-year achievement of students. It is therefore recommended that vertical scaling studies be initiated and implemented in order to monitor the achievement of students at the K-12 level.

In this current study, factors such as the length of test (40 items each), the number of common items (ten items each), the number of sampling (500 students each), and the model applied (2-PLM) were not defined as conditions and were kept constant. Future research could use these conditions as variables in order to explore their effects on the results of vertical scaling.

Researchers should also carry out a longitudinal study in order to examine the achievement of the same students over multiple years, and conduct their analyses based on data to be obtained through the longitudinal study. Since there is no single and precise criteria to be used for evaluating the accuracy of the methods used in vertical scaling, researchers are recommended to use more than one evaluation criteria (mean, mean differences, standard deviation, effect size, horizontal distance, Root Mean Square Error of Approximation [RMSEA], and bias values) when evaluating the results of scaling.

REFERENCES

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508-600). Washington D.C.: American Council on Education.

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

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

Felan, G. D. (2002, February). *Test equating: mean, linear, equipercentile and item response theory.* Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, Texas.

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.

Kim, J., Lee, W. C., Kim, D., & Kelley, K. (2009, April). *Investigation of vertical scaling using the rasch model.* Paper presented at 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 edition). New York: Springer Verlag.

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

McBride, J., & Wise, L. (2001) *Developing the vertical scale for the Florida comprehensive assessment test (FCAT).* San Antonio, Texas: Harcourt Educational Measurement.

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

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

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 Dissertation). 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.

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

Başarının ve Başarıdaki Artışın Dikey Ölçekleme İle Değerlendirilmesi: Dikey Ölçekleme Yöntemlerinin Karşılaştırılması[1]

Aylin ALBAYRAK SARI[2] & Hülya KELECİOĞLU[3]

Giriş

Bir öğrencinin hangi okula devam edeceğine karar verilirken, bir üniversite alacağı öğrencilerin sahip olması gereken test puanını belirlerken, eğitim sistemini geliştirmek için neler yapılması gerektiği hakkında kararlar alınırken ve eğitim uygulamalarındaki değişiklikleri değerlendirirken uygulanan sınavlardan elde edilen bilgiler kullanılmaktadır (Kolen & Brennan, 2004). Test uygulayıcıları da farklı formlardan ve farklı gruplardan elde edilen test puanlarını karşılaştırabilmek ve ihtiyaç duyulduğunda testleri birbirinin yerine kullanabilmek isterler. Fakat karşılaştırmanın yapılabilmesi için ham puanların standart puanlara dönüştürülmesi gerekmektedir (AERA, APA & NCME, 1999). Benzer özellikleri ölçmesi amacı ile hazırlanan farklı testlerden elde edilen puanların ortak bir ölçeğe dönüştürülmesi ile bu puanlar karşılaştırılabilmekte ve bu tür puanlara karşılaştırılabilir puanlar denilmektedir (Angoff, 1971).

Karşılaştırılabilir puanların en çok kullanıldığı alanlardan biri de, gelişimsel ölçek puanlarıdır. Akademik gelişimin yıldan yıla ne kadar olduğunun belirlenmesi için, farklı sınıf seviyelerindeki öğrencilerden elde edilen puanların ortak bir ölçeğe dönüştürülmesi ile elde edilen gelişimsel ölçek puanları kullanılır (Kolen & Brennan, 2004). Test bataryalarında ölçekleme işlemlerinin yapılmasının temel nedeni, test geliştiricilere öğrenci başarısındaki ilerlemeyi izleyebilecekleri gelişimsel ölçek puanı sağlamaktır (Loyd & Hoover, 1980). Akademik gelişimin yıldan yıla ne kadar olduğunun belirlenmesinde en temel sorun, uygulanan grubun, testlerin güçlüklerinin ve test içeriklerinin farklı olmasıdır. Bu sorunu aşabilmek için, ardışık sınıf seviyelerindeki öğrencilere ortak sorular sorularak iki farklı yetenek düzeyindeki öğrencilerin puanları ortak bir ölçeğe dönüştürülür. Bu durumda dikey ölçekleme olarak adlandırılan bağlama prosedürlerine başvurulur (Kolen & Brennan, 2004).

Dikey ölçeklemede, aynı bilgi veya becerileri ölçen fakat güçlükleri farklı olan iki testten elde edilen puanlar eşitlenir. Testi alan bireylerin yetenek düzeyleri ve testlerin güçlük düzeyleri farklıdır (Felan, 2002). Dikey ölçekleme, daha çok ilköğretimde uygulanan başarı testlerinin eşitlenmesinde kullanılmaktadır. Okulöncesi eğitimden on ikinci sınıfa kadar olan büyük çaplı değerlendirmelerde, öğrencilerin akademik gelişimlerini belirlemek için birçok çalışma yapılmaktadır. Belirlenen yıllar arasında karşılaştırma yapabilmek veya sınıf düzeyine bakılmaksızın tüm test puanlarını aynı ölçekte gösterebilmek için verilen sınıf düzeyleri arasında, geniş bir aralıktaki tüm öğrenci performansları için tek bir ölçek puanı elde etmek gerekmektedir. Böyle ölçeklere dikey ölçek, böyle bir ölçek geliştirme sürecine ve bütün sınıf düzeylerindeki değerlendirme puanlarını böyle bir ölçeğe yerleştirmeye dikey ölçekleme denmektedir (McBride & Wise, 2001). Öğrencilere, sınıf düzeylerine uygun olarak

[1] Bu çalışma Aylin Albayrak Sarı'nın "Fen Başarısındaki Artışın Belirlenmesinde Madde Tepki Kuramına Dayalı Dikey Ölçekleme Yöntemlerinin Karşılaştırılması" başlıklı tezinden üretilmiştir. Bu çalışma Hacettepe Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından desteklenmiştir. Proje Numarası: 014 T03 700 001-587.
[2] Dr. - Hacettepe Üniversitesi, Eğitim Fakültesi - aylinalb@hacettepe.edu.tr
[3] Prof. Dr. - Hacettepe Üniversitesi, Eğitim Fakültesi - hulyaebb@hacettepe.edu.tr

hazırlanan testler uygulanarak farklı sınıf düzeylerinden elde edilen puanlar aynı puan ölçeğinde ifade edilmektedir. Bu işlemler sayesinde, farklı sınıf düzeylerinden elde edilen puanlar karşılaştırılabilmekte ve bireylerin gelişimleri hakkında bilgi edinilmektedir (Kolen, 1988).

Türkiye'de de uygulanan ve öğrencilerin başarısını hem ulusal hem uluslararası platformda karşılaştırabilmeye olanak sağlayan PISA, PIRLS, TIMSS gibi uluslararası sınavların sayısı arttıkça öğrencilerin akademik başarısının belirlenebilmesi için yapılan çalışmaların sayısında ve bu başarının yıldan yıla nasıl değiştiğinin belirlenebilmesi için yapılan çalışmaların sayısında son yıllarda artış görülmektedir. Amerika Birleşik Devletleri'nde 2001 yılında alınan No Child Left Behind - Hiçbir Çocuk Geride Kalmasın- (NCLB, 2002; Public Law 107-110) yasaları gereği ülkedeki tüm çocukların akademik başarı gelişimleri izlenmeye bu konuda çalışmalar yapılmaya başlanmıştır. NCLB uygulamaları ile çoğu ülke öğrencilerinin başarı gelişimlerini izleyebilmek için yapılan çalışmalara ağırlık vermektedir. Her ne kadar alanyazında bir sınıftan daha üst sınıfa başarı gelişimlerini ölçmek için üzerinde hem fikir olunan belli bir yöntem olmasa da dikey ölçekleme çalışmaları ile öğrencilerin gelişimleri belirlenmektedir.

Alanyazında hangi yöntemlerin öğrencilerin başarılarındaki artışı en iyi ve doğru ortaya koyduğu konusunda ortak bir görüş yoktur. Buna rağmen dikey ölçekleme birçok test geliştiricisi tarafından kullanılmaktadır. Ancak her test geliştirici geliştirdiği ölçek için dikey ölçek geliştirme süreçlerini kendisi belirlemektedir (Tong & Kolen, 2007).

Bu çalışmanın amacı, madde tepki kuramına dayalı olarak dikey ölçekleme çalışması yürüterek, farklı kalibrasyon yöntemleri (ayrı ve eş zamanlı kalibrasyon) ve farklı yeterlik tahmini yöntemleri (maksimum olabilirlik, beklenen önsel dağılım, maksimum önsel dağılım) kullanılarak oluşturulan farklı dikey ölçekleme sonuçlarının ölçeklerin ortalama, standart sapma ve etki büyüklüğü değerlerine göre karşılaştırılmasıdır. Bu çalışmada üretilen dikey ölçeklerin, bir sınıf düzeyinden diğer sınıf düzeyine olan büyüme, sınıf düzeyleri arasındaki çeşitlilik ve düzey dağılımlarının ayrımı özellikleri üzerinde durulmuştur. Bir sınıf düzeyinden diğer sınıf düzeyine olan büyümeyi değerlendirebilmek için ortalamalar ve ortalama farkları, sınıf düzeyleri arasındaki çeşitliliği değerlendirebilmek için her sınıf düzeyi için standart sapma değerleri ve düzey dağılımlarının ayrımını değerlendirebilmek için de etki büyüklüğü değerleri incelenmiştir.

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).

Çalışma grubu için yapay veri seti oluşturulurken alanyazın incelenmiş ve alanyazındaki çalışmalara göre belirlenen madde ve yetenek parametreleri kullanılmıştır. 10'ar maddesi ardışık sınıf ile ortak madde olmak üzere 40'ar maddelik bir test örüntüsü oluşturulmuş ve 1500 öğrenci yanıtı R programı kullanılarak simüle edilmiştir. Yapay veri seti için yetenek parametreleri, sınıf seviyelerine göre birer birim artırılarak belirlenmiştir. Yapay veriler 6., 7. ve 8. sınıflar için yetenek parametreleri sırasıyla $[-1<\theta<0]$; $[0<\theta<1]$ ve $[1<\theta<2]$ aralıklarında üretilmiştir. Veri setindeki maddelerin ayırıcılık gücü düzeyleri sınıf seviyesi arttıkça daralacak şekilde belirlenmiştir. Yapay veriler, 6., 7. ve 8. sınıflar için sırasıyla $[0.5<a<1.5]$, $[1<a<2]$ ve $[1.5<a<2]$ aralıklarında üretilmiştir. Veri setinin madde güçlük

düzeyleri ise ortak ve ortak olamayan maddeler için ayrı ayrı belirlenmiştir. 6. sınıf düzeyindeki ortak olmayan ilk 30 maddenin b parametresi [-3<b<3] aralığında; madde 7. sınıf ile ortak son on maddenin b parametresi [-1<b<0]; 7. sınıfın 6. sınıf ile ortak olan ilk on maddesinin b parametresi [-1<b<+1]; 7. sınıfın sonraki 20 maddesinin b parametresi[-2<b<+3]; 7. sınıf ile 8. sınıfın ortak son on maddesinin b parametresi [0<b<1]; 8. sınıfın 7. sınıf ile ortak ilk on maddesinin b parametresi [-1<b<2]; 8. sınıfın sonraki 30 maddesinin b parametresi [1<b<2] aralığında türetilmiştir. Ortak maddelerin b parametreleri sınıf seviyesine göre artırılarak yapay veriler üretilmiştir.

Bulgular

Birinci alt problemde yer alan eş zamanlı kalibrasyon ile elde edilen bulgular belirlenen üç kritere göre karşılaştırılmıştır. Sınıf düzeyleri arasındaki çeşitlilik kriterini değerlendirmek için hesaplanan standart sapma değerleri incelendiğinde, 6. sınıf ile 8. sınıf standart sapmalarının birbirine yakın değerler olduğu, en düşük standart sapma değerinin MAP yöntemi kullanılarak, en yüksek standart sapma değerlerinin ML yöntemi kullanılarak elde edildiği görülmektedir. Düzey dağılımları arasındaki ayrım kriterini değerlendirmek için hesaplanan etki büyüklüğü kriterleri incelendiğinde, etki büyüklüklerinin her üç yöntemde de birbirine yakın değerler elde edildiği görülmektedir. 6. sınıf ile 7. sınıf arasındaki etki değeri orta etki, 7. sınıf ile 8. sınıf arasındaki etki değeri ise güçlü etki olarak yorumlanabilir. Alanyazın incelendiğinde elde edilen bu bulguların Tong ve Kolen (2007)'un çalışmasındaki bulguları ile paralel olduğu görülmektedir. Meng (2007) ve Tong (2005) ise çalışmalarında en küçük kestirimlerin ML yöntemi ile elde edildiği bulgusuna ulaşmışlardır.

İkinci alt problemde yer alan ayrı kalibrasyon ile elde edilen bulgular da belirlenen kriterlere göre incelenmiş ve sonuçları karşılaştırılmıştır. Sınıf düzeyleri arasındaki çeşitlilik kriterini değerlendirmek için hesaplanan standart sapma değerleri incelendiğinde, her üç sınıf düzeyine ait standart sapmalarının birbirine yakın değerler aldığı, MAP yöntemi ile EAP yönteminin sonuçlarının çok benzer olduğu, en yüksek standart sapma değerlerinin ML yöntemi ile elde edildiği görülmektedir. Düzey dağılımları arasındaki ayrım kriterini değerlendirmek için hesaplanan etki büyüklüğü kriterleri incelendiğinde, etki büyüklüklerinin her üç yöntem ile de birbirine yakın değerler elde edildiği görülmektedir. Analizler sonucu elde edilen etki büyüklüğü değerleri incelendiğinde, 6. sınıf ile 7. sınıf arasındaki ve 7. sınıf ile 8. sınıf arasındaki etki değeri zayıf etki olarak yorumlanabilir.

Elde edilen iki alt problemin bulguları karşılaştırıldığında, eş zamanlı ve ayrı kalibrasyon yönteminin her ikisinde de ortalama farkların 8. sınıf düzeyine doğru arttığı ve her iki yöntemde de en yüksek değerlerin EAP yöntemi ile elde edildiği görülmektedir. Ayrı kalibrasyon ile elde edilen ortalama farkları eş zamanlı kalibrasyon ile elde edilen ortalama farklarına göre daha düşüktür. Standart sapma değerleri karşılaştırıldığında, her iki kalibrasyon yönteminde de en yüksek değerlerin EAP ile elde edildiği ve her iki yöntemle de genel olarak birbirine yakın standart sapma değerleri elde edildiği görülmektedir. Etki büyüklüğü değerleri incelendiğinde her iki kalibrasyon yönteminde de 8. sınıfa doğru etki büyüklüğü değerlerinin arttığı görülmektedir. Ayrı kalibrasyon ile elde edilen etki büyüklüğü değerleri eş zamanlı kalibrasyona göre daha düşüktür. Her üç yetenek kestirim yöntemi birbirine yakın sonuçlar üretmiştir.

Tartışma ve Öneriler

Dikey ölçekleme sonucunda elde edilen bulgular incelendiğinde, dikey ölçekleme sürecinin karmaşık bir süreç olduğu ve tek bir doğru yöntemin olmadığı görülmektedir. Üzerinde hemfikir olunan doğru bir yöntem olmadığı için, uygulanan yöntemlerin karmaşıklığı analizlerin sonuçları göz önünde bulundurularak en uygun yöntemi yine araştırmacı belirlemesinin uygun olduğu söylenebilir.

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.

Bu araştırma tüm dikey ölçekleme sürecinin bir parçasıdır. Test geliştirme uzmanları ve uygulayıcılarına dikey ölçekleme sürecinin son aşamasında olan gözlenen-gerçek puan eşitleme süreci ile ilgili çalışmaları, gözlenen puanları etkileyen faktörleri incelemeleri önerilebilir. Bu çalışmada belirlenen parametreler aralığında simülasyon ile veri türetilmiştir. Gerçek verilere ulaşımın zor olduğu durumlarda simülasyon ile veri türetilerek analizler yapılabilir.

Öğrenci başarılarının ardışık sınıf seviyesi arttıkça arttığı görülmüştür, fakat bu artışın istendik düzeyde olup olmadığını değerlendirebilmek için çalışmalar yapılması önerilebilir. 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.

Bu çalışmada test uzunluğu (40ar madde), ortak madde sayısı (10ar madde), örneklem sayısı (500er kişi), uygulanan model (2PLM) gibi faktörler koşul olarak belirlenmemiş ve sabit tutulmuştur. Araştırmacılar bu koşulları değişken olarak kullanabilir ve dikey ölçekleme sonuçlarına etkisini araştırabilirler. Araştırmacılar, boylamsal bir araştırma yürütüp uzun yıllar aynı öğrencilerin başarısını inceleyebilir ve analizlerini boylamsal çalışmadan elde ettikleri veriler üzerinden yapabilirler.

Dikey ölçeklemede kullanılan bu yöntemlerin doğruluğunu değerlendirmek için tek ve kesin bir ölçüt olmadığı için araştırmacılara ölçekleme sonuçlarını karşılaştırırken birden fazla değerlendirme ölçütü (ortalama, ortalama farkları, standart sapma, etki büyüklüğü, yatay uzaklık, eşitleme hatası (RMSE) ve yanlılık (bias) değerleri) kullanmaları önerilebilir.

Anahtar Sözcükler: Madde tepki kuramı, Dikey ölçekleme, Kalibrasyon yöntemleri, Yetenek kestirim yöntemleri, Simülasyon, Yapay veri

Atıf için / Please cite as:
Albayrak Sarı, A., & Kelecioğlu, H. (2016). Assessment of achievement and growth by vertical scaling: comparison of vertical scaling methods [Başarının ve başarıdaki artışın dikey ölçekleme ile değerlendirilmesi: dikey ölçekleme yöntemlerinin karşılaştırılması]. *Eğitim Bilimleri Araştırmaları Dergisi - Journal of Educational Sciences Research, 6*(2), 25-38. http://ebad-jesr.com/