TURJE
www.turje.org

Volume:7 Issue:4

DOI:10.19128/turje.456656

Turkish Journal of Education

Research Article

Received	*01.09.2018*
Accepted	*06.10.2018*

Does equating matter in value-added models?

Sedat Şen
Harran University, Şanlıurfa, Turkey, sedatsen@harran.edu.tr
orcid.org/0000-0001-6962-4960
Ragıp Terzi
Harran University, Şanlıurfa, Turkey, terziragip@harran.edu.tr
orcid.org/0000-0003-3976-5054
İbrahim Yıldırım
Harran University, Şanlıurfa, Turkey, iyildirim84@gmail.com
orcid.org/0000-0002-4137-2025
Allan S. Cohen
University of Georgia, Athens, GA, USA, acohen@uga.edu
orcid.org/0000-0002-8776-9378

ABSTRACT The purpose of this study was to examine the effect of equated and non-equated data on value-added assessment analyses. Several models have been proposed in the literature to apply the value-added assessment approach. This study compared two different value-added models: the unadjusted hierarchical linear model and the generalized persistence model. The former model assumes equated tests while the latter one relaxes this assumption. Two different data sets (equated and non-equated) were analyzed with both models. Value-added estimates for both models based on a statewide examination (equated) and a countrywide examination (non-equated) data were generally consistent. School rankings showed differences between the two models. The practical implication of this study is that although there were small differences in school rankings, a model requiring an equating assumption can be applied to a non-equated data set in a case when equating between test forms is not possible.

Keywords *Value-added assessment, Hierarchical linear model, Generalized persistence model, Test equating,*

Katma-değerli değerlendirme modellerinde eşitleme önemli mi?

ÖZ Bu çalışmanın amacı, eşitlenmiş ve eşitlenmemiş verilerin katma-değerli değerlendirme analizlerine etkisini incelemektir. Katma-değerli değerlendirme yaklaşımını uygulayabilmek için literatürde birçok model önerilmiştir. Bu çalışma, iki farklı katma değerli değerlendirme modeli karşılaştırmıştır: düzeltilmemiş hiyerarşik doğrusal model (UHLMM) ve genelleştirilmiş süreklilik (GP) modeli. Birinci model eşitlenmiş testler için kullanılırken, ikincisi bu varsayımı esnetir. Her iki modelde iki farklı veri seti (eşitlenmiş ve eşitlenmemiş) analiz edildi. Her iki model için eyalet çapında yapılan bir sınav (eşitlenmiş) ve ülke çapında yapılan bir sınav (eşitlenmemiş) verilerine dayanan katma-değer kestirimleri genellikle tutarlı bulundu. Okul sıralamalarında iki model arasında bazı farklılıklar gözlendi. Bu çalışmanın pratik çıkarımı, okul sıralamasında küçük farklılıklar olmasına rağmen, test formları arasında eşitlemenin mümkün olmadığı durumlarda eşitlenmemiş bir veri seti gerektiren bir modelin uygulanabileceğidir.

Anahtar Kelimeler *Katma-değerli değerlendirme, Hiyerarşik doğrusal model, Genelleştirilmiş süreklilik modeli, Test eşitleme,*

Cite This Article: Şen, S., Terzi, R., Yıldırım, İ., & Cohen, A. S. (2018). Does equating matter in value-added models? *Turkish Journal of Education, 7*(4), 186-195. DOI:10.19128/turje.456656

INTRODUCTION

The effectiveness of a school or a teacher has been debated for decades. How to identify a qualified teacher or an effective school constitutes a major problem in education? The focus of our study which emerged from this problem on value-added analysis (VAA) is a recent approach to determining school/teacher effectiveness. In this process, various approaches have been developed for the evaluation of the school effectiveness such as using school's overall achievement scores and data envelopment analysis (Bessent & Bessent, 1980) that take into account teachers' proficiency levels in teaching. Since the problem of "effectiveness" contains many variables (Marzano, 2003), only measurable benefits can be expressed in an unbiased manner. In this context, "student achievement" can be considered the most important indicator of school effectiveness (Balcı, 1988).

The assessment of student learning is a major policy issue in the field of education (Ercikan, 2006). As is well known, there are many factors that affect student achievement. Among a number of factors, teacher and school effects in students' test score gains can be detected by a recent statistical approach to assessment of student learning (Sanders et al., 2002). Hanushek (1972) first used VAA in an accountability system. Sanders and his colleagues implemented VAA in the Tennessee Value Added Accountability System (TVAAS), a statewide testing system (Sanders & Horn, 1994; Sanders & Rivers, 1996; Sanders et al., 1997; Wright, Horn, & Sanders, 1997).

Sanders et al. (1997) defined a teacher value-added score as the differences between the predicted level of achievement and current achievement in a classroom taught by the teacher. In this definition, the magnitude of differences in the predicted and observed test scores is assumed to reflect teacher and school effectiveness. If the measured score is higher than the predicted score, it is interpreted to mean that the teacher and school "add" to student achievement, otherwise, they detract from student achievement.

A number of Value-Added Models (VAMs) have been developed to track individual students' academic growth over years and in different subjects so that teachers' contributions to that growth can be estimated (Braun, 2005). In this way, these models are intended to control for student-level socio-demographic variables (e.g., age, gender, and ethnicity) that may have effects on student achievement. The purpose of VAMs is to obtain accurate and reliable comparisons of student achievement across schools regardless of large demographic or ability differences in student populations. Some VAMs rely on fixed school effects while others rely on random school effects. For instance, early VAM applications (e.g., Hanushek, 1972; Murnane, 1975) assume fixed effects models, whereas, more recent applications (e.g., the TVAAS layer model) assume random effects models (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). The former methods are based on regression models (Tekwe et al., 2004), while the latter use more complex statistical models such as mixed models or hierarchical models (Aitkin & Longford, 1986; Raudenbush & Bryk, 1986) to assess school and teacher effects (Şen, Kim, & Cohen, 2017).

Estimating teacher effectiveness based on student achievement using VAMs requires longitudinal data in order to track the impacts of prior educational inputs on future achievement (Mariano et al., 2010). In order for test scores to be used to estimate growth, however, they need to be vertically equated and scaled to a common metric (Ballou et al., 2004; Briggs & Domingue, 2013; Doran & Cohen, 2005). Therefore, one challenging issue in modeling longitudinal data is the need for equating test scores as most standardized tests consist of different items varying in difficulty. This is a crucial point because inferences based on VAMs should be made with respect to the validity standards (Hill, 2009). Concerns have been raised about the limitations of vertical scaling and equating in part because latent constructs are subject to changes at each grade level. Furthermore, equating constructs that shift across grades can result in biased and distorted value-added teacher effectiveness (Braun & Wainer, 2007;

Martineau, 2006). Another issue that may affect the estimates of VAMs is the choice of vertical scaling methodology as this can affect the subsequent results from a given VAM (Briggs & Domingue, 2013; Briggs & Weeks, 2009).

The majority of VAMs require equating between consecutive grades, while some VAMs relax it by simply requiring scores across grades be linearly related (Shaw, 2012). In the absence of vertically scaled data sets, some possible alternative methods have been proposed for use with VAA. Mariano et al. (2010) and McCaffrey et al. (2003; 2004) proposed the variable persistence model for data with non-constant variances and covariances obtained from different developmental scales. It is also the case, however, that inferences about teacher effectiveness may be biased due to measurement error in previous test scores (Shaw, 2012). Reckase (2004) noted that comparing results across years does not provide unbiased estimates if different skills and academic domains are included in a VAA. A number of models have been introduced to deal with aforementioned issues (Broatch & Lohr, 2012; Lockwood et al., 2007; Mariano et al., 2010). Some of these VAMs ignore construct shift entirely and directly carry out analyses with vertically scaled test scores, while some VAMs model construct shift with vertically scaled test scores, and some other VAMs ignore vertically scaled test scores completely. There is no research reported, however, comparing the sensitivity of VAMs based on equated and non-equated data sets.

Two approaches to VAA were investigated in this study, the unadjusted hierarchical linear model (UHLMM; Tekwe, et al., 2004) and the generalized persistence model (GP model; Mariano et al., 2010). Results from these two models were compared to determine whether they were consistent with one another and the two models differed with respect to school rankings?

METHODOLOGY

In this study, two different data sets (equated and non-equated) were analyzed. Equated data for this study were taken from a vertically scaled statewide mathematics test administered to 8th graders from 2002 and 2003 in a large Southeastern state in USA. As these data were vertically scaled, they were consistent with models requiring equating. This test is a part of a criterion-referenced test that aims to assess student achievement in the high-order cognitive skills represented in the state standards in reading, mathematics, writing, and science. Three types of questions – multiple choice items, graded response items, and performance tasks – were used in this test. Non-equated data for this study were taken from the 2015 November and 2016 April administrations of a countrywide exam in Turkey, namely, the Exam for the Transition from Basic Education to Secondary Education (also known as TEOG; Ministry of Education). The exam scores from the Grade 8 mathematics section of this test were used for analyses under a non-equating condition. The exams consist of 20 multiple choice questions and data were collected from schools in a province located in southeastern Turkey. Twenty schools were randomly selected from each of the two data sets. The samples consisted of 9,811 students for the vertically equated statewide examination data set and 941 students for the countrywide examination data set.

VAMs Used in This Study

Unadjusted hierarchical linear model and generalized persistence model were used to examine the effect of vertical equating on value-added estimates. A brief explanation about these models is presented below.

Unadjusted hierarchical linear model

The UHLMM uses unadjusted change scores with a random intercept. This model consists of a two-level HLM described by the following equations;

Student-level model,

$$d_{ijs} = \beta_{0is} + \varepsilon_{ijs} \tag{1}$$

where d_{ijs} is the change score, β_{0is} is a random intercept associated with the school i, and ε_{ijs} represents random error.

School-level model,

$$\beta_{0is} = \gamma_{0s} + \xi_{is} \tag{2}$$

where γ_{0s} is the mean of the random intercepts, β_{0is} and ξ_{is} are the random effect and random error for school i on the random intercept for subject area s, respectively. β_{0is} and ξ_{is} are assumed to be independent**.** The single equation form can be written as

$$d_{ijs} = \beta_{0s} + \xi_{is} + \varepsilon_{ijs}. \tag{3}$$

Generalized persistence model

The GP model is a general multivariate model for estimating teacher or school effects based on a longitudinal data set, which was developed by Mariano et al. (2010), it was intended to accommodate both school effect decay and scale changes. The GP model estimates are computed from a Bayesian framework for non-equated longitudinal data set. A student's year *t* score depends on an overall year *t* mean for all students, plus a cumulative sum of the current year and past year schools' effects, plus a random residual error term for the student in the current testing year. y_{it} is the achievement score of student *i* in year *t* and the GP model for this score is

$$y_{it} = \mu_t + \left(\sum_{g=1}^{t} \sum_{j=1}^{J_g} \phi_{igj} \theta_{g[jt]} \right) + \varepsilon_{it}, \tag{4}$$

where μ_t is the overall mean for the year, ϕ_{igj} equals 1 if student was taught by school *j* in year *g*, and 0 otherwise. Therefore, the products of $\phi_{igj}\theta_{g[jt]}$ provide the school effects for the current and prior grades, and ε_{it} is the residual error term.

Results of the UHLMM and GP models based on the equated and non-equated data sets were compared in this study. As mentioned, the UHLMM requires equated test scores while the GP model relaxes this assumption. UHLMM analyses were conducted with SAS software using the code provided by Tekwe et al. (2004); GP model estimations were conducted using GPvam R package (Karl et al., 2012).

RESULTS

Value-added estimates from each VAM used in this study are shown in Table 1. The UHLMM provides value-added estimates as best linear unbiased predictors (BLUP), while the GP model provides empirical best linear unbiased predictors (EBLUP). As shown in Table 1, value-added estimates for the statewide test from both models appeared to be consistent in terms of sign for most of the schools except for Schools 3, 4, 6, 9, 13, and 20 (presented in bold). Similarly, value-added estimates for the countrywide test from both models appeared to be consistent in terms of sign for most of the schools except for Schools 8, 13, 15, 17, and 20 (presented in bold). However, the magnitude of the school estimates varied.

Table 1
Estimates of the school effects obtained from two VAMs based on grade 8 statewide and countrywide math test results

	Statewide Test					Countrywide Test			
School ID	UHLMM		GP		School ID	UHLMM		GP	
	BLUP	SE	EBLUP	SE		BLUP	SE	EBLUP	SE
5	45.58	8.44	20.71	8.72	12	2.47	1.79	1.12	1.23
2	23.65	7.86	17.57	8.47	4	1.79	1.86	0.67	1.26
10	21.38	9.32	26.53	9.10	3	0.82	1.78	0.59	1.23
9	**19.24**	**8.85**	**-2.79**	**8.89**	6	1.39	1.97	0.52	1.29
3	**10.66**	**8.82**	**-4.46**	**8.88**	1	0.98	1.81	0.43	1.24
12	8.84	7.98	9.59	8.52	7	1.07	1.79	0.29	1.23
8	6.99	7.57	7.96	8.36	**17**	**-0.26**	**1.81**	**0.25**	**1.24**
4	**2.11**	**9.00**	**-6.89**	**8.96**	**15**	**-0.78**	**1.19**	**0.14**	**1.08**
20	**0.46**	**8.34**	**-5.50**	**8.68**	**20**	**-0.13**	**1.83**	**0.09**	**1.25**
6	**-0.09**	**8.88**	**4.34**	**8.91**	5	0.39	1.65	0.08	1.19
19	-0.85	8.37	-2.55	8.69	2	0.21	1.67	0.00	1.20
7	-4.83	8.53	-0.73	8.75	**8**	**0.25**	**1.93**	**-0.02**	**1.28**
13	**-5.79**	7.82	**1.65**	8.46	**13**	**0.35**	**1.79**	**-0.09**	**1.23**
15	-8.79	9.09	-8.54	9.00	11	-1.43	1.85	-0.33	1.25
18	-13.06	7.68	-10.34	8.39	18	-0.59	1.87	-0.38	1.26
11	-14.58	7.62	-2.93	8.37	10	-0.66	1.87	-0.47	1.26
14	-14.76	7.42	-2.27	8.29	16	-1.12	1.85	-0.58	1.25
17	-16.93	9.06	-17.09	8.98	19	-1.16	1.64	-0.66	1.19
16	-19.55	7.48	-6.04	8.32	9	-1.76	1.08	-0.75	1.06
1	-39.67	7.66	-20.49	8.39	14	-1.83	1.79	-0.93	1.23

School IDs are sorted by UHLMM; inconsistent results (in terms of signs) between two models are presented in bold.

Value-added scores are typically used for school ranking. The schools were ranked based on the value-added scores obtained from two different models. The school ranks based on the UHLMM and GP model are presented in Table 2. As shown in Table 2, school rankings based on statewide test data showed differences between UHLMM and GP models. Only the least successful school (i.e., School 1) was found to be the same in both models. The three most successful schools were the same but they were in different orders for the two models. School rankings for the countrywide test data also showed differences between UHLMM and GP models (see Table 2). As shown in Table 2, twenty-five percent of the schools (Schools 12, 4, 1, 9, and 14) were ranked the same by both models (presented in bold). Although other schools' ranks did differ, the school rankings appeared to be close to each other from both models. Spearman rank correlations were calculated between the ranks obtained from both models. The correlations were .699 and .878 for the statewide (equated) data and the countrywide (non-equated) data, respectively.

Table 2
School ranks obtained from the two VAMs based on grade 8 statewide and countrywide math test results

Ranking	Statewide Test School ID		Countrywide Test School ID	
	UHLMM	GP	UHLMM	GP
1	5	10	**12**	**12**
2	2	5	**4**	**4**
3	10	2	6	3
4	9	12	7	6
5	3	8	**1**	**1**
6	12	6	3	7
7	8	13	5	17
8	4	7	13	15
9	20	14	8	20
10	6	19	2	5
11	19	9	20	2
12	7	11	17	8
13	13	3	18	13
14	15	20	10	11
15	18	16	15	18
16	11	4	16	10
17	14	15	19	16
18	17	18	11	19
19	16	17	**9**	**9**
20	**1**	**1**	**14**	**14**

Consistent rankings between two models are presented in bold.

DISCUSSION and CONCLUSION

In this study, equated statewide test data and non-equated countrywide test data were analyzed with both UHLMM and GP models. In general, the estimated effects of most of the schools are compatible for both data sets. On the other hand, school rankings showed differences between the UHLMM and GP model for both data sets. The school rankings based on the two VAMs were closer for the non-equated data set than for the equated data set. The correlation of school effects across models also appeared to be stronger in the non-equated data (i.e., the countrywide test data) than in the non-equated data (i.e., the statewide examination test data). Thus, the differences between results of the UHLMM and GP model appeared to be smaller for the non-equated data set. One possible explanation for the differences between two data sets may be due to the equating effect. As a result, it can be concluded that practitioners should be careful about the model choice. When test scores are equated, then the data should be analyzed with the UHLMM. When data are not equated, then tests can be estimated by either the UHLMM or GP model. As the test score equating is a tedious process and is not always possible in the real testing applications (e.g., the lack of anchor items), practitioners may prefer either the UHLMM or GP model. These results are consistent with Yıldırım and Şen's (2018) study. Yıldırım and Şen (2018) have compared the GP model to the UHLMM under non-equated data set and they found that tests can be estimated by either the UHLMM or GP model for non-equated data set.

Test equating is an important process if one wants to compare results from different forms of the same test. This is likewise important when the test scores from multiple years are to be compared. However, this issue has not been studied sufficiently for comparison of value-added assessments. A relatively small number of studies have been reported examining scaling effects on value-added estimates (e.g., Briggs & Domingue, 2013; Briggs & Weeks, 2009; Briggs, Weeks, & Wiley, 2008). Briggs and Weeks (2009), for example, examined the effect of different scaling methods on school level estimates, and they found that scaling did have an effect on the estimates. Similarly, in the present study showed, there were differences observed in terms of school-level value-added estimates and in school rankings between equated and non-equated data sets. Briggs and Domingue (2013) note that choices in vertical scaling may also have an effect on teacher and school effects. Although only one data set with vertical scaling was examined in this study, results provide evidence that test equating may have an effect on model selection and school estimates. Although vertical scaling appears to be important for growth models, vertical equating using IRT does not guarantee an equal interval scale in value-added assessment applications (Ballou, 2009).

Several VAMs have been developed for determining teacher and school effectiveness. Each model has some strengths and weaknesses. Persistence models are different from gain score models in that they incorporate persistence of school effects. Another possible explanation for the difference between the two data sets in this study may be due to school effect estimates that are sensitive to different modeling specifications, such as the persistence of school effects. Although VAMs appear to provide an objective tool for use in educational accountability systems, these models should be used cautiously along with other tools to determine effective and ineffective schools (Beardsley, 2008).

Thinking of a different perspective, some countries such as Turkey and the US give a key role to private schools and tutoring centers for high-stakes tests. It could be particularly helpful for parents to decide which school or tutoring center they should send their children. In this regard, ranking of these schools and centers based on its effectiveness could be considered as an alternative procedure because tracking students' academic growth can be explained by teachers' contributions using VAMs. Furthermore, considering the performance salaries of teachers in these private institutions (Boran, Atalmis, & Sagir, 2015), the importance of using VAMs cannot be ignored.

In this study, models were compared using empirical data sets for two consecutive years for a single subject (i.e., mathematics). VAM applications are also potentially biased if school- and student-related covariates are excluded, although some VAMs can statistically control school- and student-related variables. Research on the effects of equating on value-added scores might benefit by inclusion of covariates for school- and student-related variables.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modeling in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1–43.

Balcı, A. (1988). Etkili okul. *Eğitim ve Bilim, 12*(70), 21-30.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4,* 351–383.

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*, 37–66.

Beardsley, A. A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Bessent, A. M., & Bessent, E. W. (1980). Determining the comparative efficiency of schools through data envelopment analysis. *Educational Administration Quarterly, 16*(2), 57–75.

Boran, A., Atalmis, E. H., Sagir, E. (2015). Özel öğretim kurs merkezi öğretmenleri ve çalışma koşulları [Private tutoring centers and their working conditions]. *Turkish Journal of Education 4*(4), 17–29.

Braun, H. (2005). Value-added modeling: What does due diligence require? In R.W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 19–40). Maple Grove, MN: JAM Press.

Braun, H., & Wainer, H. (2007). Value added modeling. In C.R. Rao & S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26.* Amsterdam: Elsevier.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*(6), 551–576.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy, 4*, 384–414. doi:10.1162/edfp.2009.4.4.384

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

Broatch, J., & Lohr, S. (2012). Multidimensional assessment of value added by teachers to real-world outcomes. *Journal of Educational and Behavioral Statistics, 37*, 256–277.

Doran, H. C., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 80–104). Maple Grove, MN: JAM Press.

Ercikan, K. (2006). Development in assessment of student learning. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 929–952). Mahwah, NJ: Erlbaum.

Hanushek, E. A. (1972). *Education and race: An analysis of the educational production process.* Lexington, MA: Lexington Books.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*, 700–709. doi:10.1002/pam.20463

Karl, A. T., Yang, Y., & Lohr, S. (2012). *GPvam: maximum likelihood estimation of multiple membership mixed models used in value-added modeling*. R Package Version 2.0-0.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics, 35*(3), 253–279.

Martineau, J. (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for value-added accountability. *Journal of Educational and Behavioral Statistics, 31*, 35–62.

Marzano, R. J. (2003). *What works in schools: Translating research into action*? Alexandria, VA: ASCD.

McCaffrey, D., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. Washington, DC: RAND.

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 67–101.

Murnane, R. J. (1975). *The impact of school resources on the learning of children*. Cambridge, MA: Ballinger Publishing.

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

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*, 117–120.

Sanders, W. L. & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville: University of Tennessee, Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA:Corwin.

Sanders, W. L., Saxton, A., Schneider, J., Dearden, B., Wright, S. P., & Horn, S. (2002). *Effects of building change on indicators of student achievement growth: Tennessee Value-Added Assessment System.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Shaw, L. H. (2012). *Incorporating latent variable outcomes in value-added assessment: An evaluation of univariate and multivariate measurement model structures.* (Unpublished doctoral dissertation). University of Nebraska, Digital Commons at the University of Nebraska-Lincoln.

Şen, S., Kim, S.-H., & Cohen, A. S. (2017). Comparative analysis of common statistical models used for value-added assessment of school performance. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 303–320.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., … Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*, 11–36.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 1*(1), 57–67.

Yıldırım, İ., & Şen, S. (2018). Katma-değerli değerlendirme modellerinde test eşitleme durumunun incelenmesi. In S. Dinçer, (Ed.), *Değişen dünyada eğitim* (pp. 125–134). Ankara: Pegem Akademi.

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Bir okulun veya öğretmenin etkililiğinin değerlendirilmesi yıllardır tartışılmaktadır. Bu bağlamda, “öğrenci başarısı”, okul etkililiğinin en önemli göstergesi olarak düşünülebilir (Balcı, 1988). Öğrenci öğrenmesinin değerlendirilmesi, eğitim alanında önemli bir politika konusudur (Ercikan, 2006). Bilindiği gibi, öğrenci başarısını etkileyen birçok faktör vardır. Bu faktörler arasında, öğrencilerin sınav puanlarındaki öğretmen ve okul etkileri, öğrenci öğrenmesinin değerlendirilmesine yönelik yeni bir istatistiksel yaklaşımla tespit edilebilmektedir (Sanders ve ark., 2002). Sanders, Saxton ve Horn (1997) öğretmenin katma değer puanını, öğretmenin derslerini yürüttüğü bir sınıftaki beklenen başarı düzeyi ile mevcut başarı arasındaki fark olarak tanımlamıştır. Bu tanımda, tahmin edilen ve gözlemlenen test puanlarındaki farklılıkların büyüklüğünün, öğretmen ve okul etkililiğini yansıttığı varsayılmaktadır. Ölçülen puan tahmin edilen puandan yüksekse, öğretmen ve okulun öğrenci başarısına “eklediği” anlamına geldiği, aksi halde öğrenci başarısını azalttığı söylenebilmektedir.

Her bir öğrencinin akademik gelişimini yıllar boyunca ve farklı konularda izlemek için bir dizi Katma-Değerli Model (KDM) geliştirilmiştir, böylece öğretmenlerin bu büyümeye katkısı tahmin edilebilmektedir (Braun, 2005). KDM’lerin amacı, öğrenci popülasyonundaki büyük demografik veya yetenek farklılıklarını göz önünde bulundurarak, okullardaki öğrenci başarısının doğru ve güvenilir karşılaştırmasını elde etmektir. KDM’leri kullanarak öğrenci başarısına dayalı olarak öğretmen etkililiğini tahmin etmek, eğitim çıktılarının gelecekteki başarı sonuçları üzerindeki etkilerini izlemek için boylamsal veriler gerektirebilir (Mariano, McCaffrey & Lockwood, 2010). Ancak, test puanlarının büyümeyi tahmin etmede kullanılması için, bu puanların ortak bir metriğe eşitlenmesi (ör. Dikey [vertical]) ve ölçeklendirilmesi gerekmektedir (Ballou, Sanders, & Wright, 2004; Briggs & Domingue, 2003; Doran & Cohen, 2005). Bu nedenle, boylamsal verileri modellemede zorlayıcı bir konu farklı madde güçlük değerlerine sahip standart testlerin eşitlenmiş olmasıdır. KDM’lerin çoğunluğu ardışık sınıflar arasında eşitlik gerektirirken, bazı KDM’ler basitçe notların doğrusal olarak ilişkili olmasını gerektirerek bu varsayımı esnetmektedir (Shaw, 2012). Dikey olarak ölçeklenmiş veri kümelerinin yokluğunda, Katma-Değerli Değerlendirme (KDD) yaklaşımında kullanılabilecek bazı alternatif yöntemler önerilmiştir. Mariano, McCaffrey ve Lockwood (2010), McCaffrey, Lockwood, Koretz ve Hamilton (2003) ve McCaffrey, Lockwood, Koretz, Louis ve Hamilton (2004), farklı gelişimsel ölçeklerden elde edilen sabit olmayan varyanslar ve kovaryanslara sahip veriler için süreklilik modellerini (persistence models) önermişlerdir.

Yukarıda belirtilen konuların ele alınması için bir dizi model tanıtılmıştır (Broatch & Lohr, 2012; Lockwood ve ark., 2007; Mariano ve ark., 2010). Bununla birlikte, KDM’lerde test eşitleme durumunu karşılaştırmaya yönelik kapsamlı bir araştırmaya rastlanmamıştır. Bu çalışmada KDD çerçevesinde iki yaklaşım araştırılmıştır: düzeltilmemiş hiyerarşik doğrusal model (UHLMM; Tekwe ve ark., 2004) ve genelleştirilmiş süreklilik modeli (GP modeli; Mariano ve ark., 2010). Bu iki modelin sonuçları, birbirleriyle tutarlı olup olmadıklarını ve iki modelin okul sıralamasına göre farklılık gösterip göstermediğini belirlemek için karşılaştırılmıştır. Tekwe ve ark. (2004) tarafından sağlanan kodu kullanarak SAS yazılımı ile UHLMM analizleri yapılmıştır. GP modelinin tahminleri GPvam R paketi kullanılarak gerçekleştirilmiştir (Karl, Yang & Lohr, 2012). UHLMM, en iyi doğrusal yansız tahmin ediciler (BLUP) olarak katma değerli tahminler sağlarken, GP modeli ampirik en iyi doğrusal yansız tahmin edicileri (EBLUP) sağlar.

Bu çalışmada, eşitlenmiş eyalet geneli test verileri ve ülke çapında uygulanan eşitlenmemiş test verileri, hem UHLMM hem de GP modelleri ile analiz edilmiştir. Genel olarak, okul etkilerinin tahminlerinin işaretleri her iki veri seti için uyumlu bulunurken okul etki büyüklüklerinin tahmini değerleri farklılık göstermiştir. Öte yandan, okul sıralamaları UHLMM ve GP modeli arasında her iki veri kümesi için farklılıklar göstermiştir. İki KDM’ye dayanan okul sıralamaları, eşitlenmemiş veri kümesinde eşitlenmiş veri kümesine göre daha yakın çıkmıştır. Modeller arasındaki okul etkilerinin korelasyonu eşitlenmemiş verilerde (yani, eşitlenmemiş TEOG test verileri) eşitlenmiş verilerden (yani, eyalet çapında sınav test verileri) daha güçlü olduğu görülmüştür. Bu nedenle, eşitlenmemiş veri seti için UHLMM ve GP modelinin sonuçları arasındaki fark daha küçük görünmektedir. İki veri seti arasındaki farklar için olası bir açıklama, eşitleme etkisine bağlı olabilir. Sonuç olarak, uygulayıcıların model seçimi konusunda dikkatli olmaları gerektiği sonucuna varılabilir. Test puanları eşitlendiğinde, veriler UHLMM ile analiz edilmelidir. Test verileri eşit olmadığı zaman, testler UHLMM veya GP modelleriyle tahmin edilebilir. Test puanının eşitlenmesi yorucu bir süreç olduğundan ve gerçek test uygulamalarında yapılması her zaman mümkün olmadığı için (örneğin, ortak maddelerin eksikliği), uygulayıcılar bu durumlarda UHLMM veya GP modelini tercih edebilirler.

Test eşitleme, aynı testin farklı formlarının sonuçları karşılaştırmak isteniyorsa önemli bir süreçtir. Bu, birden çok yıldaki test puanlarının karşılaştırılması gerektiğinde de aynı şekilde önemlidir. Bununla birlikte, bu konu katma-değerli değerlendirmelerinin karşılaştırılması için yeterince incelenmemiştir. Katma değerli tahminler üzerinde ölçekleme etkilerini inceleyen nispeten az sayıda çalışma rapor edilmiştir (örneğin, Briggs & Domingue, 2013; Briggs & Weeks, 2009; Briggs, Weeks, & Wiley, 2008). Briggs ve Weeks (2009) farklı eşitleme yöntemlerinin okul etkisi tahminleri üzerindeki etkisini incelemiştir. Briggs ve Weeks (2009) çalışması, eşitlemenin tahminler üzerinde bir etkisi olduğunu bulmuştur. Benzer şekilde, bu çalışmada, okul düzeyinde katma-değerli tahminler ile eşitlenmiş ve eşitlenmemiş veri setleri arasındaki okul sıralamasında farklılıklar olduğu görülmüştür. Bu çalışmada, test eşitlemenin model seçimi ve okul tahminleri üzerinde bir etkiye sahip olabileceğine dair kanıt sunulmaktadır.