ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);303-320

Comparative Analysis of Common Statistical Models Used for Value-Added Assessment of School Performance*

Okul Performansının Katma-Değerli Değerlendirilmesinde Kullanılan Yaygın İstatistik Modellerinin Karşılaştırmalı Analizi

Sedat Şen ** Seock-Ho Kim *** Allan S. Cohen ****

Abstract
The purpose of this study was to compare three popular value-added models used in measuring school effectiveness based on their distinguishing characteristics. In this study, the simple fixed effects model (SFEM) and two hierarchical models (UHLMM and AHLMM) were analyzed using value-added measures obtained from a common data set with two years standard assessment data. Value-added measures obtained from these three models were analyzed to determine the impact of the differences of each model. Correlational analyses were also conducted to see whether there were meaningful relationships among these value-added models. SFEM and UHLMM models produced very similar rank orders of school effects while SFEM and AHLMM had only a moderate correlation. Thus there was not much difference between SFEM and two HLM models in terms of the rank orders of schools.

Keywords: School effectiveness, value-added assessment, value-added models, hierarchical linear models.

Öz
Bu çalışmanın amacı, okul etkililiğini ölçmede yaygın olarak kullanılan üç katma-değerli modeli ayırt edici özelliklerine dayanarak karşılaştırmaktır. Bu çalışmada iki yıllık bir standart test verisi kullanılarak bu veriden elde edilen katma-değerli ölçümler vasıtasıyla basit sabit etki modeli (SFEM) ve iki hiyerarşik doğrusal model (UHLMM ve AHLMM) analiz edilmiştir. Bu üç modelden elde edilen katma-değer ölçümleri, her modelin farklılıklarının etkisini belirlemek için analiz edildi. Bu katma değerli modellerin sonuçları arasında anlamlı ilişki olup olmadığını görmek için korelasyon analizine başvurulmuştur. SFEM ve UHLMM modelleri okul etkilerini benzer derecede sıralarken, SFEM ve AHLMM sonuçları orta derecede bir korelasyona sahiptir. Bu nedenle, okulların sıralamasına göre SFEM ve iki HLM modelinden elde edilen sonuçlar arasında çok fazla fark bulunmamıştır.

Anahtar Kelimeler: Okul etkililiği, katma-değerli değerlendirme, katma-değerli modeller, hiyerarşik doğrusal modeller.

INTRODUCTION

Over the past few decades, there has been growing interest in the effectiveness and accountability of schools around the world. As an example, this has been the case with the U.S., especially since the adoption of the No Child Left Behind act of 2001 which requires states to measure student academic achievement and to report on progress using Adequate Yearly Progress (AYP) measures (Amrein-Beardsley, 2008). This system is based on an approach which gives rewards to schools that make contributions to students' learning and sanctions those that do not make any improvement on student test scores. Early applications of this state-wide assessment have focused on the current status of

* This study was produced from the first author's Master's thesis.
** Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa-Turkey, e-mail: sedatsen@harran.edu.tr, ORCID ID: orcid.org/0000-0001-6962-4960
*** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia, USA, e-mail: shkim@uga.edu, ORCID ID: orcid.org/0000-0002-2353-7826
**** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia, USA, e-mail: acohen@uga.edu, ORCID ID: orcid.org/0000-0002-8776-9378

DOI: 10.21031/epod.321840
Received: 15.06.2017
Accepted: 09.08.2017

students. The current-status approach compares different cohorts of students at a single point in time (Doran & Izumi, 2004). It simply uses the percentage of students who passed the state test at the end of the school year.

Educators recognize that a one-time test score is not always a useful way to estimate school effects on student performance. Differences among schools may be due to student and school variables that are not measured in tests but that influence test scores. Current-status methods don't take socioeconomic factors into account, for example, when assessing schools' effectiveness. Although these methods are located at the heart of the state accountability system, there are at least two reasons why they're invalid and inappropriate to use for the purpose of school comparisons.

First, students come to school with different backgrounds. In other words, there is no random assignment of students to schools (Doran & Izumi, 2004) yet the statistical methodology underlying this approach assumes random assignment. This results in making unfair comparisons between disadvantaged and advantaged schools in terms of socioeconomic status.

Second, current-status methods are cumulative. They reflect the impact of learning obtained from all previous schools on students' performance scores (Doran & Izumi, 2004) but they do not differentiate current effects from previous effects. Thus, we cannot hold only the latest school accountable for a student's good or poor test score if the student has changed schools in the past. As Ballou, Sanders, and Wright (2004) note, holding schools accountable based on mean achievement levels makes no sense, when students enter those schools with large mean differences in achievement.

It is widely accepted that status-based accountability systems are likely to be flawed, resulting in inaccurate judgments of school quality (Doran & Izumi, 2004; Tekwe et el., 2004). As the shortcomings of this method increasingly become apparent, an alternative way of assessing school effectiveness using growth models has gained acceptance. This new method focuses on the improvement students in the school made during the year. Instead of considering how cohort groups have increased in knowledge, measuring individual student progress over time from one time point to the next is more reasonable in terms of "learning," which is meant to be "change." Growth models are designed to generate estimates from these kinds of data (Doran & Izumi, 2004).

In this regard, researchers have developed a method called value-added analysis (VAA) which enables them to use individual student achievement scores over time in order to identify effective schools. As defined by Tekwe et al. (2004) "Value-added is a term used to label methods of assessment of school/teacher from one year to the next and then use that measure as the basis for a performance assessment system" (p. 31). Pioneers of VAA claim that VAA generates fairer and more accurate estimates than those generated by state tests that measure only the achievement of a single year. The primary purpose of VAA is to determine the impact of teachers or schools on the progress of their students (Raudenbush, 2004). To do this, VAA computes gain scores by taking the differences between students' scores on state tests from one year to the next (Sanders et al., 2002).

The VAA approach evaluates schools based simply on how they increased the level of their students' knowledge. The two basic ideas underlying value-added measurement are that it is calculated for each individual nested within the schools and that it is based on changes in student performance from one year to the next (Ladd & Walsh, 2002). Another advantage cited of VAA is that, unlike the current-status method, it can control the effect of confounding variables such as student and school socioeconomic status that may influence the test scores. In this way, it is an attempt to minimize the influence of experiences, privilege, and ethnicity on student performance.

In general, value-added models (VAMs) are a class of statistical model procedures that analyze students' standardized test scores over time to identify the degree to which a student's progress is a function of their own characteristics or of the characteristics of their school (Doran & Izumi, 2004). VAMs have recently received a great deal of interest from both policy makers and researchers due to a belief that these models can adequately determine how individuals are growing over time while

appropriately attributing that portion of their gain scores to their schools (Sanders, & Horn, 1994; Sanders, & Horn, 1998; Sanders, Saxton, & Horn, 1997). It is an area of research in education that has achieved a significant role in shaping the school accountability system.

Several VAM approaches have been suggested by researchers. Current-status methods all rely on regression models and assume that school effects are fixed (Tekwe et al., 2004). They are also confounded with nonschool factors (Sanders, 2000), whereas VAMs require the use of more complex statistical models such as mixed models and hierarchical models which assume school effects to be random. Hanushek (1972) is generally credited as the first to use VAM methods in the accountability system. Sanders, who developed the Tennessee Value Added Assessment System (TVAAS), was the first to implement VAMs in a statewide testing system (Stewart, 2006).

According to a report by the RAND corporation (McCaffrey, Lockwood, Koretz, & Hamilton, 2003) early VAM applications (e.g., Hanushek, 1972; Murnane, 1975) primarily used fixed effects models. More recent applications, including the TVAAS layered model, have used random effects models exclusively.

Another important model is one developed by Raudenbush and Bryk (1986) and Aitkin and Longford (1986). This model relies on hierarchical linear models to measure student growth. Although there are several VAMs which are based on different statistical assumptions (Braun, 2004; Tekwe et al., 2004), the most popular has been the TVAAS (Olson, 2004). For any of these models to be useful in VAA analysis, however, the test scores must be vertically scaled (Ballou et al., 2004; Doran & Cohen, 2005). That is, the test scores must all be expressed on a common scale that extends over the time periods included in the analysis. In brief, longitudinal data, annual assessment, and vertically equated tests are said to be basic elements of VAMs. Typically, standardized assessment scores are used in VAM studies. Though no VAM has yet been obvious to be clearly superior over another, VAMs are considered to be fairer and more accurate than conventional methods (Doran & Izumi, 2004).

To date, several alternative models, ranging from simple gain scores to complex mixed models, have been suggested by researchers with regard to assessment of school effectiveness. However, there have been a limited number of studies which make comparisons among these different models (Ballou et al., 2004; McCaffrey et al., 2003; Tekwe et al., 2004). Selection of the most useful model for an accountability analysis requires determining which model is most accurate. Fortunately, a few important studies have been conducted to determine the most desirable model for computing school effects. The Journal of Educational and Behavioral Statistics published one volume solely concerning the VAA and popular VAMs (Wainer, 2004). The papers in that volume concluded that there are numerous acceptable models as opposed to only a single acceptable model.

Tekwe et al. (2004), Ballou et al. (2004), and McCaffrey et al. (2003) describe differences among VAMs. As these studies have noted, compared to other methods, VAMs are less biased and produce more precise estimates. Although there is a lack of comparative studies showing which VAM is better than the others, the LMEM model has been used frequently for accountability purposes. Ballou et al. (2004) conducted a simulation study to evaluate the TVAAS model which is based on the LMEM. Results indicated that the TVAAS uses a highly parsimonious model that omits controls for contextual factors such as SES and demographics that influence achievement.

Unlike the LMEM model, HLM models include school and student variables and attempt to control such factors by statistical adjustment (Bryk & Raudenbush, 1992). Sanders et al. (2002) noted that inclusion of these factors in HLM affects the school estimates resulting in biased measures of schools towards zero. Sanders' LMEM model does not account for these variables. That model attempts to eliminate controls for these variables by use of multiple measures on each student (Ballou et al., 2004). Sanders found that the inclusion of these factors to the model did not result in a significant difference between the two models (Ballou et al., 2004). Results of a simulation study comparing the general model, which is similar to the AHLMM, with those of a layered model which

is similar to the LMEM, however, suggested that the AHLMM fit the data better than the layered model (McCaffrey et al., 2003).

Tekwe et al. (2004) found little or no benefit from use of more complex models. The simpler SFEM model provided results that were more accurate compared to estimates from the other models. Results also indicated that the AHLM model would be preferred when there is a need for controlling the effects of student and school variables estimates and that selection of one of the two models should be based on non-empirical considerations.

Although VAMs have been shown as an important tool for accountability system, a number of researchers criticized the VAMs application for determining school or teacher effectiveness. An important criticism of VAMs is that they do not yet solve the problem of randomization completely (Wiley, 2006). Another criticism of VAMs is about the precision of the value-added estimates obtained from longitutional data sets. Schochet and Chiang (2010) examined the likely system error rates for measuring teacher and school performance in the upper elementary grades using ordinary least squares (OLS) and Empirical Bayes (EB) methods applied to student test score gain data.

Similarly, Guarino, Reckase, and Wooldridge (2015) investigated the accuracy of the value-added estimates of teachers obtained from commonly used value-added models. They found that no one method accurately captures true teacher effects and classifies teachers in realistic conditions. In addition, VAM approach has been shown to be invalid when there is endogeneity which may be due to correlation between the random effect in the hierarchical model and some of its covariates (Manzi, San Martín, & Van Bellegem, 2014). Another criticism of VAMs is about the data requirements of these models. As mentioned above vertically equated test results from multiple years are basic elements of VAMs. This makes VAM useful for a single developmental scale. However, most of the VAMs cannot be used for multiple test instruments (on different scales) administered within a school year. A few researchers have discussed how to use VAMs to analyze longitudinal student achievement data obtained from multiple instruments (Green, 2010; Rivkin, Hanushek, & Kain, 2005).

There have been numerous studies that show the strengths of the VAMs over the conventional methods. However, the concern remains that simpler models are as efficient as more complex models (Doran & Fleischman, 2005). Several models introduced in VAA calculate the value-added measures based on different assumptions. SFEM and UHLMM do not account for school/non-school variables, while AHLMM attempts to control these factors by statistical adjustments. In this study, the impact of school and non-school factors are compared on school-level value-added scores using an empirical data with an eye to better understanding problems associated with model complexity. Three popular VAMs (i.e., SFEM, UHLMM, and AHLMM) were examined in this study. The models selected for the present study show similarities to a previous study conducted by Tekwe et al. (2004). Tekwe et al. (2004) have also examined the LMEM in their study in addition to the models compared in this study. LMEM was excluded from our study due to data requirements of this model.

METHOD

Instrumentation

Data for this study were taken from 2002 and 2003 statewide mathematics and reading test results of the Florida Comprehensive Assessment Test (FCAT) for Grades 6 to 8. Separate analyses were done for each grade. The FCAT is a criterion-referenced test that aims to assess student achievement in high-order cognitive skills represented in the Sunshine State Standards (Florida Department of Education, 2003) in reading, mathematics, writing, and science. The FCAT includes three types of questions: multiple choice items, graded response items, and open-ended items. FCAT scaled scores used in this study were vertically scaled, thus making them appropriate for VAA.

Sample

Separated analyses were performed for each of the grade cohorts for Grades 6, 7 and 8 in a large Florida school district with 44 secondary schools for 2002 and 2003. Only standard curriculum students were used in the analyses. Special education students with any exceptionality and students in the limited English proficiency (LEP) program for two or fewer years were excluded due to following reasons. Generally, it is impossible to collect two years of score from students with severe cognitive disabilities that are required for most of the VAMs. In addition, students with limited English cannot show real performance on state test and this may have a negative effect on the value-added measures of schools. Students whose reported ages were outside the acceptable age range for a given grade were excluded from the analyses. Listwise deletion was applied to exclude these students' information.

A total of 60,718 students were available for analyses after the exclusions: 19,611 for Grade 6, 20,433 for Grade 7, and 20,674 for Grade 8. Non-school variables for socioeconomic status and minority status were included in the data set. Socioeconomic status information was provided in the form of student's eligibility for the free-or-reduced lunch program. Minority status is a school-level variable is based on the proportion of African-American or non-African-American students in the school. Descriptive statistics based on grade and subject combination are presented in Table 1.

Table 1. Sample Size, Mean FCAT and Standard Deviation by Subject, Grade and Year, and Percent Minority and Percent Poverty in 2003 by Grade

		Reading			Math			Demographics in 2003	
		2002 score	2003 score	Change score	2002 score	2003 score	Change score	Poverty	Minority
6th	*N*	19,611	19,611	19,611	19,611	19,611	19,611	19,611	19,611
	M	1421.32	1527.89	106.57	1566.02	1581.17	15.15	73.7%	28.6%
	SD	368.52	371.85	235.62	294.80	297.80	189.48		
7th	*N*	20,433	20,433	20,433	20,433	20,433	20,433	20,433	20,433
	M	1493.98	1623.32	129.33	1554.14	1692.70	138.56	72.2%	28.4%
	SD	385.43	348.92	244.52	293.74	255.18	191.43		
8th	*N*	20,674	20,674	20,674	20,674	20,674	20,674	20,674	20,674
	M	1606.93	1782.10	175.16	1675.76	1804.40	128.64	70.3%	28.6%
	SD	345.79	276.42	223.87	274.60	216.95	169.142		

Value-Added Models Used in This Study

As noted above, VAMS have the capability of controlling the effects of non-school variables as well as prior performance. In this study, results for three commonly used VAMs were compared: a simple fixed effects model and two hierarchical linear models. It should be noted that layered mixed effects model (LMEM) is another popular VAM that is useful for data sets collected from students attending multiple schools. This model was not examined in this study as the data set in this study does not have students attending multiple schools within a school-year. This makes present study different from Tekwe et al. (2004).

Simple fixed effects model (SFEM)

Fixed effects models (FEM) used for VAA assume school effects to be fixed rather than random. These have the advantage of being the simplest VAM, requiring less computation than the others. As a result, estimates from FEM are more easily understood by policymakers and educators with little statistics experience (Wiley, 2006). The simple fixed effects model (SFEM) is an extension of the FEM. One concern with this model is that it does not incorporate student-level covariates and does

not apportion variance for students who have attended multiple schools. Thus it does not produce any shrunken estimates. As SFEM uses only two years of data in a single subject, however, its application is very straightforward.

Model parameterization:

$$d_{ijs} = \beta_{0s} + \sum_{k=1}^{44} \beta_{1ks} S_{kij2} + \varepsilon_{ijs}, \quad (1)$$

where

$d_{ijs} = \gamma_{ijs2} - \gamma_{ijs1}$,

$d_{ijs} =$ is a simple change score obtained from difference between two examinations of a student i in school j on the same subject area s,

γ_{ijst} =is the test score on the subject area s $(s = 1, 2)$ at time t $(t = 1, 2)$ for the student j $(j = 1, \cdots, n_j)$ in school i $(i = 1, \cdots, n_i)$,

S_{kij2} = is effect coding at time (t = 2) for school k $(k = 1, \cdots, 44)$ with coding numbers m $(m = 1, \cdots, 43)$,

S_{kij2} = 1 for $k = m$ and $k \neq 44$; 0 for $k \neq m$ and $k \neq 44$; -1 for $k = 44$,

and ε_{ijs} is the random error for student *j* in school *i* for subject area *s*.

It is assumed that $\varepsilon_{ijs} \sim N\left(0, \sigma_{\varepsilon s}^2\right)$.

β_{1ks} in Equation 1 is the value-added component in subject area *s* for school *k*.

Hierarchical linear models.

Hierarchical linear models (HLM) require using hierarchically ordered nested data. The hierarchical nature of the structure is that students are considered nested within classes and classes as nested within schools. Due to the nature of the data used in education, HLM has been used extensively for analysis of school effects (Raudenbush & Bryk, 2002). HLM is a special type of the general mixed models family and can be used to obtain value-added measures. These models demand more computation than SFEM, but unlike SFEM, HLM-based models produce shrunken effects.

The HLM analysis consists of four parts as follows (Raudenbush & Bryk, 1988-1989):

i. Apportioning variation between and within units of analysis
ii. Assessing the homogeneity of regression assumption
iii. Testing for compositional effects
iv. Assessing the effect of the method

Traditional regression methods assume that individuals are independent of each other although students in the same school might have similar results when compared to students from different schools. HLM can handle this violation of the independence assumption unlike linear models.

In this study, two different types of HLM were examined, unadjusted HLM (UHLMM) with random intercept and adjusted HLM (AHLMM). The AHLMM consists of two equations called student-level and school-level models. The two-level HLM provides an analytical framework for examining the effects of schools on student outcomes. An extension of two-level model (i.e., three-level HLM) can

be used to obtain value-added estimates of schools and teachers using a data set structure which has students nested within teachers and teachers nested within schools.

Unadjusted hierarchical linear model (UHLMM)

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

Student-level model:

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

where d_{ijs} is the change score defined as in Equation 1, β_{0is} is a random intercept associated with the school *i*, and ε_{ijs} is a random error.

School-level model:

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

where γ_{0s} is the mean of the random intercepts, β_{0is}, and ξ_{is} are the random effect and random error of school *i* on the random intercept for subject area *s*. β_{0is} and ξ_{is} are assumed to be independent. ε_{ijs} and ξ_{is} are assumed to have normal distribution.

Single equation form:

$$d_{ijs} = \beta_{0s} + \xi_{is} + \varepsilon_{ijs}. \quad (2)$$

Adjusted hierarchical linear model (AHLMM)

The AHLMM model is adjusted for student-level and school-level covariates.

Student-level model:

$$d_{ijs} = \beta_{0is} + \beta_{1s}\gamma_{ijs1} + \beta_{2s}Min_{ij} + \beta_{3s}Pov_{ij} + \varepsilon_{ijs},$$

where $d_{ijs} = \gamma_{ijs2} - \gamma_{ijs}$, β_{0is} is a random intercept associated with the school *i* and subject area *s*, Min_{ij} = an indicator of minority status (Yes or No) for student *j* in school *i*, Pov_{ij} = an indicator of poverty in which the status of a student eligible for a free-and-reduced lunch is considered to be poverty (Yes or No) for student *j* in school *i*, β_{1s}, β_{2s}, and β_{3s}, are the fixed effects of previous year's test score, minority status, and poverty on learning gain in subject area *s*, and ε_{ijs} is a random error.

School-level model:

$$\beta_{0is} = \gamma_{0s} + \gamma_{1s}Z_{1i} + \gamma_{2s}Z_{2i} + \xi_{is},$$

where Z_{1i} is the mean input score for the school *i*, Z_{2i} is the percentage of students in poverty in the school *i*, ξ_{is} is is the random error associated with the value of the random intercept for the subject area test (*s*) and the school *i* in the student level model, and the γ's are fixed effects coefficient parameters. The within and between school error terms, ε_{ijs} and ξ_{is}, are assumed to be independent.

Single equation form:

$$d_{ijs} = \gamma_{0s} + \gamma_{1s}Z_{1i} + \gamma_{2s}Z_{2i} + \beta_{1s}\gamma_{ijs1} + \beta_{2s}Min_{ij} + \beta_{3s}Pov_{ij} + \xi_{is} + \varepsilon_{ijs}, \quad (3)$$

RESULTS

Assumptions and characteristics of each of the VAMs used in this study are shown in Table 2. Thus, differences in characteristics of the models can be seen in Table 2. Interpretations of results for each model are based on distinguishing characteristics of the model. Correlations between VAM measures of schools generated from each model are given in Table 3. Schools were ranked based on their VAM estimates from different models. Correlational results provide information about the rank order of school effects generated from each model. Tables with these rankings are also presented in Appendices.

Table 2. Summary of Distinguishing Characteristics of Models

Model identifier	Dependent variable	School effects	Student-level variable	School-level variables
SFEM	Change score	Fixed	No	No
UHLMM	Change score	Random	No	No
AHLMM	Change score	Random	Yes	Yes

Note. Adapted from Tekwe et al. (2004, p.23). SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

Table 3. Table of Correlations Between Value-added Measures of the Models

	6th grade		7th grade		8th grade	
	Math	Reading	Math	Reading	Math	Reading
SFEM vs. UHLMM	.99	.99	.99	.99	.99	.99
SFEM vs. AHLMM	.75	.85	.80	.55	.73	.74
UHLMM vs. AHLMM	.75	.85	.80	.54	.73	.74

Note. SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

With respect to the assumption of school effects as random, the SFEM is the only one that accounts for school effects as fixed effects. Therefore, it is appropriate to compare the SFEM to the UHLMM. The UHLMM differs only in that it considers the school effect to be random. The most important finding that is evident in Table 3 is the very high correlation between SFEM and UHLMM value-added estimates (r = .99) in all cohorts. This suggests that the two models provide the same rank ordering of schools. Thus, it is possible to conclude that there was no difference between taking school effects as random or fixed in terms of rank order of school effects.

A second concern in measuring school effectiveness is to include school and non-school covariates in the models. Among the models in this study, only the AHLMM can take both student-level and school-level effects into account. Apart from this characteristic, the AHLMM and UHLMM are identical. As a result, we can make inferences based on the comparison of these two models. As can be seen in Table 3, there were moderate correlations ranging from .54 to .85 between AHLMM and UHLMM for the different cohorts. This indicates that the effects of including school and non-school variables in the AHLMM had a clear impact on the VAA estimates.

Another comparison with the AHLMM can be made with SFEM. This comparison will help to see the effects of employing shrinkage or including school and non-school variables in the AHLMM model. Correlations between these two models showed moderate values ranging from .55 to .85. These results suggest there is a noticeable difference between SFEM and AHLMM. Although the AHLMM is appropriate when seeking to adjust for confounding variables, the only thing we can

really conclude is that there was a difference between the rank orders of schools based on these two models.

Strong correlations were observed between results generated by the SFEM and UHLMM, but much more modest correlations were observed between the AHLMM and all other models. We conclude on the basis of these results that there was not much difference between the SFEM and hierarchical models in terms of the rank order of school estimates.

Once a model is chosen, value-added measures for students can be converted to standardized grades to determine the relative performance of the teachers within each school (or attributed to each school). To obtain standardized grades, standardized value-added measures were divided by their standard errors and assigned grade point average (GPA) values using the following criteria from Tekwe et al. (2004):

If $z > 2$, then assign a grade of A and 4 growth points;
If $1 < z \leq 2$, then assign a grade of B and 3 growth points;
If $-1 < z \leq 1$, then assign a grade of C and 2 growth points;
If $-2 < z \leq -1$, then assign a grade of D and 1 growth points;
If $z \leq -2$, then assign a grade of F and 0 growth points.

Results of the standardized grade conversions are presented in Table 4.

Since grades from the SFEM and UHLMM models were found to be similar, we present only results for the SFEM and AHLMM in Table 4. Results in Table 4 suggest that large schools with higher value-added estimates tended to have lower GPA values than smaller schools with lower value-added estimates, although it was also possible that large schools with lower value-added estimates could have higher GPA values.

Individual school estimates and their rankings were obtained for each grade from three different VAMs. Only estimates for Grade 6 are presented (see Tables 5 and 6 in Appendices A and B). (Estimates for Grades 7 and 8 are available on request from the first author.). For the SFEM, estimates can be interpreted as the difference between the school specific sample average change and the average changes overall. Estimates from the UHLMM are shrunken estimates of school effects from the SFEM. These can be calculated as estimates of the best linear unbiased predictors of the random effects for each school and each grade. Value-added estimates of the AHLMM were also calculated as estimates of best linear unbiased predictors.

The ranks of the school estimates from the SFEM were similar to those of the school estimates from the UHLMM. It is interesting to note that estimates from both models were very similar. This result also suggests that there was little difference in estimating school effects as either random or fixed. Results from the AHLMM had moderate agreement with results from SFEM. Results from each of the models suggested that VAM rankings of schools differed across different grades. Results compared for each grade, however, were very consistent with the results of correlational analyses.

Table 4. Growth Point Averages for Each School Based on Value-Added Measures from SFEM and AHLMM

	SFEM						AHLMM					
School	M	R	G6	G7	G8	T	M	R	G6	G7	G8	T
1	0.00	0.66	0.50	0.50	0.00	0.33	0.33	1.33	0.50	1.50	0.50	0.83
2	3.00	3.66	2.00	4.00	4.00	3.33	2.33	2.33	1.50	2.50	3.00	2.33
3	1.00	2.33	0.50	1.00	3.50	1.66	2.33	2.66	2.00	2.00	3.50	2.5
4	2.00	1.00	2.00	0.50	2.00	1.50	2.00	1.66	2.50	1.00	2.00	1.83
5	3.33	3.00	2.00	4.00	3.50	3.16	2.66	1.66	2.00	2.50	2.00	2.16
6	2.66	1.66	2.50	2.50	1.50	2.16	2.66	2.00	3.00	2.00	2.00	2.33
7	1.00	2.33	1.50	1.00	2.50	1.66	1.66	2.00	2.00	1.50	2.00	1.83
8	3.00	2.66	4.00	2.00	2.50	2.83	1.66	2.00	2.00	1.50	2.00	1.83
9	2.00	1.00	0.00	1.00	3.50	1.50	2.33	2.00	2.00	2.00	2.50	2.16
10	3.66	3.00	4.00	3.00	3.00	3.33	3.00	2.33	2.50	2.50	3.00	2.66
11	2.33	1.66	3.50	2.50	0.00	2.00	2.33	1.33	2.00	2.50	1.00	1.83
12	3.00	2.33	3.00	2.50	2.50	2.66	1.66	1.66	2.00	1.50	1.50	1.66
13	2.66	2.00	4.00	2.00	1.00	2.33	2.33	2.33	3.00	2.00	2.00	2.33
14	1.66	2.00	4.00	1.50	0.00	1.83	2.00	2.00	2.00	2.00	2.00	2.00
15	1.00	2.33	2.00	1.00	2.00	1.66	1.00	2.00	2.00	1.00	1.50	1.50
16	0.00	1.00	1.00	0.00	0.50	0.50	1.33	2.00	1.00	2.00	2.00	1.66
17	0.33	1.00	1.50	0.00	0.50	0.66	0.66	1.33	2.00	0.50	0.50	1.00
18	0.00	1.33	0.00	0.00	2.00	0.66	2.00	2.66	2.00	2.00	3.00	2.33
19	3.33	3.00	3.50	4.00	2.00	3.16	2.66	1.66	3.00	2.50	1.00	2.16
20	1.66	1.66	1.00	2.00	2.00	1.66	2.00	2.00	2.00	2.00	2.00	2.00
21	1.00	1.00	0.00	1.50	1.50	1.00	0.66	1.00	1.00	1.50	0.00	0.83
22	2.00	2.33	2.00	0.50	4.00	2.16	2.00	2.66	2.50	1.50	3.00	2.33
23	3.00	2.00	1.50	3.00	3.00	2.50	3.66	2.33	3.00	3.00	3.00	3.00
24	1.33	2.66	1.50	3.00	1.50	2.00	0.66	2.00	2.00	1.00	1.00	1.33
25	3.33	2.66	4.00	2.50	2.50	3.00	3.33	2.33	4.00	2.00	2.50	2.83
26	1.00	1.66	2.50	1.50	0.00	1.33	0.66	1.33	1.50	1.50	0.00	1.00
27	2.00	2.33	1.50	3.50	1.00	2.16	2.33	2.66	2.50	3.50	1.50	2.50
28	1.66	2.66	3.50	1.00	2.00	2.16	1.66	2.00	2.00	1.50	2.00	1.83
29	3.66	2.66	3.00	3.00	3.50	3.16	2.00	2.00	2.00	2.00	2.00	2.00
30	0.66	2.33	0.50	1.00	3.00	1.50	2.66	3.00	2.00	2.50	4.00	2.83
31	1.66	2.66	2.00	1.50	3.00	2.16	2.66	2.66	2.50	2.00	3.50	2.66
32	2.33	3.00	1.50	2.50	4.00	2.66	3.33	3.33	3.00	3.00	4.00	3.33
33	2.00	2.66	0.00	4.00	3.00	2.33	1.33	1.66	0.00	2.50	2.00	1.50
34	1.33	1.33	4.00	0.00	0.00	1.33	2.00	2.00	2.00	2.00	2.00	2.00
35	0.00	0.66	0.00	0.00	1.00	0.33	1.33	2.00	1.00	1.50	2.50	1.66
36	2.33	1.66	1.50	2.50	2.00	2.00	2.00	1.33	1.00	2.00	2.00	1.66
37	3.33	1.33	3.00	2.50	1.50	2.33	2.66	1.33	2.50	2.50	1.00	2.00
38	2.66	0.66	1.00	2.00	2.00	1.66	2.66	1.66	1.50	2.50	2.50	2.16
39	2.66	3.33	4.00	2.50	2.50	3.00	2.00	2.33	2.00	2.00	2.50	2.16
40	2.66	2.66	1.50	3.50	3.00	2.66	1.33	1.33	1.50	1.50	1.00	1.33
41	1.33	0.66	3.00	0.00	0.00	1.00	2.66	1.66	2.50	1.50	2.50	2.16
42	2.00	1.66	2.50	2.50	0.50	1.83	2.00	2.00	1.50	2.50	2.00	2.00
43	2.00	0.66	0.00	3.00	1.00	1.33	1.00	1.00	0.50	1.50	1.00	1.00
44	-	-	-	-	-	-	2.33	2.33	2.00	2.50	2.50	2.33

Notes. M = Math GPA; R = Reading GPA; T = Total GPA; 6G = 6th Grade GPA; 7G = 7th Grade GPA.

DISCUSSION and CONCLUSION

The purpose of the present study was to determine whether there were similarities or differences among three models commonly used for value-added assessment of schools. The simplest model was

the SFEM. This model treats school effects as fixed. Two hierarchical linear models were also included. Each model has distinguishing characteristics and different assumptions. Value-added estimates of individual schools obtained from these models were analyzed to compare results from the different models on the estimates.

The primary question was to investigate whether results from simpler models, such as the SFEM, differed as effective as the more complex models such as AHLMM in terms of school rankings. Previous research has found that little difference between the results of simple and complex value-added models in that correlations between estimates from SFEM and AHLMM models ranged from .55 to .85 (Tekwe et al., 2004). Results from this study were somewhat consistent with previous research in that the simple model produced similar rank orders of school effects with the more complex AHLMM. Based on these results, it may be concluded that simple models were as effective as more complex models at estimating value added effects of schooling. Further, simpler models generally could be used in place of more complex models such as AHLMM. There is typically a desire for using simpler statistical models among policy makers as well as the general public. Results of the present study tend to support the use of simpler models such as the SFEM in value-added accountability systems.

Another concern in value-added studies is to determine the impact of the inclusion of school and student background variables into models on model estimates. Among the models in this study, only the AHLMM includes statistical adjustments for these potentially confounding variables. Tekwe et al. (2004) suggested that both inclusion and exclusion of these variables during the analysis result in biased estimates of schools. In this study, the estimates from the AHLMM model were compared to estimates from other models to determine the effects of these covariates. No major differences were observed between results of the AHLMM, the UHLMM and the SFEM. Correlations between estimates from the AHLMM and SFEM ranged from .55 to .85. Correlations between results from the AHLMM and the UHLMM also ranged from .54 to .85. These correlations were mostly consistent with results from previous research. Consistent with previous research, inclusion of these covariates did have an effect on value-added estimates. The omission of covariates from the model appeared to bias parameter estimates when students were stratified by those covariates (McCaffrey et al., 2003).

The present study also reported on standardized GPA grading and rankings of each school based on value-added estimates from each model. These results were consistent with the correlational analysis. VAM-based rankings of schools showed differences over grades. It should be noted that the conclusions drawn from this study cannot be generalized to teachers or to other test conditions.

Although, value-added models are believed to be useful in school accountability system, the credibility of these methods have been questioned by a number of researchers (AERA, 2015, Amrein-Beardsley, 2014; Ballou & Springer, 2015; Guzman, 2016; The American Statistical Association (ASA), 2014). Amrein-Beardsley (2014), emphasized that VAMs have several problems with reliability, validity, and bias, affecting their fairness and transparency. In addition to these serious problems, theoretical and methodological assumptions of VAMs have also been questioned in the literature. Thus, school (or teacher) performances should not be based on only value-added measures obtained from any of the VAMs described in this study. As Amrein-Beardsley (2014) suggested multiple measures and more holistic evaluation systems should be used for school evaluations rather than relying only on VAMs.

REFERENCES

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

American Education Research Association [AERA] Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448–452.

American Statistical Association [ASA]. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

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

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education. Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.

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.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77–86.

Braun, H. I. (2004). *Value-added modeling: What does due diligence require?* Princeton, NJ: Educational Testing Service.

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

Doran, H., & Izumi, L. T. (2004). *Putting education to the test: A value-added model for California.* San Francisco, CA: Pacific Research Institute.

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.

Doran, H. C., & Fleischman, S. (2005). Challenges of value-added assessment. *Educational Leadership, 63*(3), 85–87.

Florida Department of Education. (2003). Florida Comprehensive Assessment Test (FCAT): Assessment and School Performance.

Green, J. L. (2010). *Estimating teacher effects using value-added models*. University of Nebraska at Lincoln, Department of Statistics: Dissertations and Theses.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted*? Education Finance and Policy, 10*(1), 117–156.

Guzman, N. L. (2016). Review of rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability. *Education Review//Reseñas Educativas*, 23.

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

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review*, *21*, 1–17.

Manzi, J., San Martín, E., & Van Bellegem, S. (2014). School system evaluation by value added analysis under endogeneity. *Psychometrika, 79*, 130–153.

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

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

Olson, L. (2004, November 16). "Value added" models gain in popularity. *Education Week.* Retrieved from http://www.edweek.org/ew/articles /2004/11/17/12value.h24.html

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*, 121–129.

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

Raudenbush, S., & Bryk, A. (1988-89). Methodological advances in studying effects of schools and classrooms on student learning. In E. Z. Roth (Ed.), *Review of research in education* (pp. 423–475). Washington, DC: American Educational Research Association.

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

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

Sanders, W. L. (2000). Annual CREATE Jason Millman Memorial Lecture: Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*, 329–339.

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., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247–256.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Educational Assessment System (TVAAS): 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 Press.

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.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Washington, DC: Institute for Education Sciences.

Stewart, B. E. (2006). *Value-added modeling: The challenge of measuring educational outcomes.* New York, NY: Carnegie Corporation of New York.

Tekwe, C. D., Carter, R. L., Ma, C-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & 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.

Wainer, H. (2004). Introduction to a special issue of the journal of educational and behavioral statistics on value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 1–3.

Wiley, E. W. (2006). *A practitioner's guide to value-added assessment*. Retrieved from http://nepc.colorado.edu/files/Wiley_APractitionersGuide.pdf

GENİŞ ÖZET

Giriş

Son yıllarda, okulların etkililiği ve hesap verebilirliği konularına ilgide dünya çapında bir artış gözlenmektedir. Bu konulardaki ilk uygulamalar, öğrencilerin mevcut başarı durumlarını kullanmaya odaklanmıştır. Mevcut durum yaklaşımı, farklı kademedelerdeki öğrencileri tek bir zaman noktasında (genellikle dönem sonunda) karşılaştırmaya dayanmaktadır (Doran ve Izumi, 2004). Eğitimciler, bir seferlik test puanını kullanarak öğrencilerin performansı üzerindeki okul etkilerini tahmin etmenin çok doğru bir yol olmadığını düşünmektedir. Bu nedenle hesap verebilirlik sisteminde okul etkinliğini değerlendirmenin alternatif yolları aranmıştır. Bu yeni yaklaşımlar öğrencilere okulda yıl boyunca yapılan iyileştirmeler üzerine odaklanmaktadır. Araştırmacılar etkili okulları belirlemek için bireysel öğrenci başarı puanlarını zamanla beraber kullanmalarını sağlayan katma-değerli değerlendirme (KDD) fikrini geliştirmiştir. Tekwe ve diğerleri (2004)'e göre "Katma değer ifadesi bir yıldan diğerine okul ya da öğretmenin değerlendirilmesi yöntemlerini ifade eden ve daha sonra bu ölçütün bir performans değerlendirme sistemi için temel teşkil etmesinde kullanılan bir terimdir" (s.31). KDD'nin öncüleri, KDD'nin yalnızca bir yılın başarısını ölçen standart testlerden elde edilen sonuçlardan (mevcut durum yaklaşımı) daha adil ve daha doğru tahminler ürettiğini iddia etmektedir. KDD'nin birincil amacı öğretmenlerin veya okulların öğrencilerin gelişimine olan etkilerini belirlemektir (Raudenbush, 2004). KDD sistemi, okulları öğrencilerin bilgi düzeylerini nasıl arttırdıklarına göre değerlendirmeye dayanır.

Bugüne kadar, okul etkililiğinin değerlendirilmesinde basit gelişim puanlarından karmaşık karma modellere kadar değişen çeşitli alternatif modeller (Katma-Değerli Modeller; KDM) önerilmiş olmasına rağmen bu modelleri karşılaştıran sınırlı sayıda çalışma bulunmaktadır (McCaffrey vd., 2003; Tekwe ve diğerleri, 2004; Weiss, 2006). Hesap verebilirlik sistemlerinde yeni KDD yaklaşımlarını benimseyerek problemlere çözüm bulmak adına hangi modelin en etkili ve hangi modelin en kolay uygulanabilir olduğunun gösterilmesinin uygulamacılar adına faydalı olacağı düşünülmektedir. Herhangi bir KDD modeli geleneksel yöntemlerden daha üstün olmakla birlikte devletlerin hesap verebilirlik sistemlerinin (karmaşıklığından ötürü) KDM'leri kullanmadaki

isteksizliği gözlenmektedir. KDM'ler açısından daha basit modellerin daha karmaşık modeller kadar etkili olduğunu gösteren çalışmalara uygulamacıların fikrini değiştirmek adına ihtiyaç duyulmaktadır.

KDM'lerin geleneksel yöntemlerden daha etkili olduğu görüşünün yanında bu modellerin ve dayandığı istatistiksel uygulamaların doğru ve güvenilir sonuçlar üretmediğini ileri süren çalışmaların olduğu da unutulmamalıdır (Guarino, Reckase ve Wooldridge, 2015; Manzi, San Martín ve Van Bellegem, 2014; Schochet ve Chiang, 2010; Wiley, 2006).

KDD kapsamında geliştirilen modeller farklı varsayımlara dayanarak okul katma değerlerini hesaplamaktadır. Örneğin, bazı modeller okula ait ve okul dışı diğer değişkenleri hesaba katmazken, bazı modeller bu faktörleri istatistiksel düzenlemelerle kontrol etmeye çalışmaktadır. Florida Comprehensive Assessment Testi'nden (FCAT) elde edilen sınav puanlarını kullanarak; bu çalışmada, okula ait ve okul dışı faktörlerin okul düzeyindeki katma-değer puanlarına etkileri araştırılmış ve KDM'lerin karmaşıklığı konusu üzerine ışık tutulmaya çalışılmıştır. Bu iki konu bağlamında uygulayıcılar ve eğitim yöneticileri için en faydalı modeli/modelleri belirlemek amacıyla en yaygın olarak kullanılan üç katma-değerli model incelenmiştir. Bu çalımada cevaplanmaya çalışılan temel soru: "Okul etkinliğinin katma-değerli değerlendirilmesi için karmaşık istatistiksel modellere gerçekten ihtiyacımız var mı, yoksa daha basit modellerle daha karmaşık modellerle olduğu kadar okul etkililiğini etkin bir şekilde değerlendirebilir miyiz?"

Yöntem

Bu çalışmada, 2003 yılında Florida eyaletinde bulunan orta okul (6-8. sınıflar) öğrencilerine ait verilerin ayrı ayrı analizleri yapılmıştır. Öğrencilerin FCAT matematik ve okuma testlerinden 2002 ve 2003 yıllarında aldıkları puanlar büyük bir bölgedeki 44 okulun katma-değerlerini tahmin etmek için analiz edilmiştir. Analizlerde sadece standart müfredatı takip eden öğrenciler kullanılmıştır; özel eğitim öğrencileri ve sınırlı İngilizce yeterlik programında iki veya daha az yıl geçiren öğrenciler de analizlerin dışında tutulmuştur. Bu çalışmada toplam 60.718 öğrenci bulunmaktadır. Yoksulluk durum bilgisi, bir öğrencinin ücretsiz öğle yemeği alıp almayacağına bağlı olarak belirlenmiştir. Diğer okul dışı değişken, etnik köken değişkeni olarak tanımlanmıştır. Bu çalışmada, okul etkililiği bağlamında üç popüler KDM (basit sabit etki modeli (SFEM) ve iki hiyerarşik doğrusal model (düzeltilmiş HLM: AHLMM ve düzeltilmemiş HLM: UHLMM)) incelenmiştir.

Bulgular ve Sonuç

Bu çalışmada kullanılan katma-değerli modellerden elde edilen okul katma değer tahminleri, modellerin farklı özelliklerinin bu tahmin değerleri ve okul etkililiğini belirlemedeki etkilerini görmek için incelenmiştir. Birincil soru, SFEM gibi daha basit modellerin okul sıralaması açısından AHLMM gibi daha karmaşık olan modeller kadar etkili olup olmadığını araştırmaktı. Önceki araştırmalar, basit ve karmaşık modellerin sonuçları arasında çok az farklılıklar olduğunu bulmuştur (Tekwe ve diğerleri, 2004). Bu çalışmadaki analizlere göre SFEM ve AHLMM arasındaki korelasyon ,55 ile ,85 arasında değişmektedir. Bu çalışmanın sonucu Tekwe ve diğerleri (2004) bulgularıyla kısmen tutarlılık göstermektedir. Basit model (SFEM) okul sıralaması açısından AHLMM ile benzer sıralamalar üretmiştir. Ayrıca, basit modellerin karmaşık modeller kadar etkili olduğunu ve bu basit modelin (SFEM) çalışmada ele alınan daha karmaşık modellerin (AHLMM ve UHLMM) yerine geçebilecekleri sonucuna varılmıştır. Uygulamacılar arasında daha basit istatistiksel modelleri kullanma isteği olduğundan, bu sonuçlar önceki araştırmalara ek olarak basit modellerin de karmaşık modeller kadar hesap verebilirlik sisteminde etkili olabileceğini göstermektedir.

Bu çalışmada okula ve öğrenciye ait değişkenlerin katma-değer tahminleri üzerine etkisi de incelenmiştir. Modeller arasında sadece AHLMM okul tahmin değerlerini etkileyebilecek bu

karıştırıcı değişkenleri kontrol edebilen istatistiksel düzeltmelere sahiptir. Tekwe ve diğerleri (2004), KDD analizlerinde bu değişkenlerin modele doğrudan dahil edilmesinin veya tamamıyla göz ardı edilmesinin, okullar hakkında yanlı tahminler elde edilmesine yol açtığını belirtmektedir. Araştırmacılar bunu yapmak yerine bu değişkenleri istatistiksel olarak kontrol edebilen modellerin kullanılmasını tavsiye etmektedir. Çalışmamızda AHLMM'de bu değişkenlerin etkisini görmek için karşılaştırmalar yapılmıştır. AHLMM'nin sonuçları ile UHLMM'nin ve SFEM'nin sonuçları arasında belirgin bir farklılık bulunamamıştır. Genel olarak, bu eş değişkenlerin dâhil edilmesinin, katma değerli tahminler üzerinde büyük bir etkisi olduğu sonucuna varabiliriz. Bu sonucun aynı zamanda Tekwe ve diğerleri (2004)'ün yorumlarıyla da uyumlu olduğu görülmektedir. Sonuçlara dayanarak, öğrencilerin farklı arka planlara sahip olduğu durumlarda diğer VAM'lara nazaran AHLMM'nin tercih edilmesini tavsiye edebiliriz.

Çalışmamızın sınırlılıklarından birisi de okul değerlendirmesinde sıklıkla kullanılan LMEM'nin kullanmış olduğumuz veri yapısından dolayı çalışmaya dâhil edilmemiş olmasıdır. LMEM, okul, konu ve yıl açısından çoklu durumları dikkate alan güçlü bir modeldir. İki yıllık veri ve istikrarlı öğrenciler nedeniyle LMEM'nin gerçek etkisini çalışmamızda göremeyeceğimiz düşüncesiyle analizler arasına eklenmemiştir. Çok değişkenli yöntemin okulun etkinliği üzerindeki etkisini görmek için daha fazla araştırmanın farklı veriler kullanarak yapılması önerilir.

Sonuç olarak, hesap verebilirlik sistemini şekillendirmede KDM'lerin önemli bir rolü olduğu bu çalışmada gösterilmeye çalışılmıştır. Bu çalışmada elde edilen bulgular Florida Eyaletinde uygulanan FCAT sınavından elde edildiği için, çalışmanın bulgularının diğer eyaletlere ya da ülkelere genellenip genellenemeyeceği kesin olmamakla beraber alan yazında KDD modellerinin kullanıma dair ek kanıtlar sunduğu açıktır. Ayrıca bu çalışmada katma-değerli değerlendirme yaklaşımı ve uygulanmasında kullanılan modeller tartışıldığı için çalışmanın okul etkililiği üzerine çalışan yöneticiler ve eğitimciler için faydalı olacağı düşünülmektedir. Yurt dışında birçok ülkede tercih edilen ve okul performansının değerlendirilmesinde kullanılan bu modellere ait ayrıntılı açıklamalar içeren bu çalışmanın ülkemizde bu modelleri uygulamak isteyen araştırmacılara yardımcı olacağı düşünülmektedir.

Appendices

Appendix A. Grade 6 Math Estimates

Table 5. Estimates of the School Effects Obtained from Three VAMs Based on Grade 6 Math Results

	SFEM		UHLMM		AHLMM	
Rank*	Estimate	School ID	Estimate	School ID	Estimate	School ID
1	54.126	34	47.539	25	43.963	25
2	50.883	10	45.621	13	22.978	19
3	46.729	41	41.297	19	19.502	6
4	43.380	39	39.370	34	18.071	22
5	32.629	42	29.861	41	17.040	13
6	32.055	14	28.624	6	14.723	32
7	31.476	16	27.787	10	14.085	37
8	25.660	11	23.992	14	11.225	23
9	24.598	13	22.195	11	10.468	31
10	24.186	1	21.982	8	10.446	41
11	24.036	28	21.971	42	10.196	4
12	22.312	6	19.878	12	9.751	27
13	21.740	26	19.528	37	9.287	12
14	19.308	12	17.133	39	9.253	34
15	10.485	8	9.613	29	7.079	11
16	10.254	7	9.203	36	7.000	42
17	9.985	33	8.964	4	6.872	30
18	8.621	2	7.843	28	5.885	8
19	7.804	43	6.678	22	4.198	38
20	6.766	36	5.929	24	3.468	10
21	3.580	30	3.196	26	2.105	29
22	1.377	38	1.181	38	1.489	39
23	0.695	18	0.552	5	1.349	36
24	-4.603	24	-4.034	15	1.149	14
25	-6.922	19	-6.003	31	1.092	9
26	-9.650	15	-8.645	17	-1.146	24
27	-9.718	29	-8.779	32	-3.735	17
28	-10.151	25	-9.158	23	-4.549	5
29	-10.366	31	-9.277	2	-4.838	3
30	-13.212	23	-11.659	7	-5.679	28
31	-13.489	37	-12.163	40	-6.250	20
32	-18.218	20	-16.900	1	-6.839	18
33	-19.810	40	-17.065	27	-7.952	15
34	-20.228	17	-18.407	9	-9.202	44
35	-21.194	32	-19.345	20	-9.377	7
36	-21.681	35	-19.563	30	-10.226	26
37	-24.274	4	-22.656	16	-12.559	35
38	-32.380	5	-28.937	3	-14.295	40
39	-33.237	3	-29.809	21	-15.309	2
40	-34.008	22	-30.654	18	-17.635	21
41	-50.386	44	-41.325	44	-23.673	16
42	-53.935	21	-45.737	43	-26.979	1
43	-58.680	9	-49.734	33	-39.858	43
44	-	27	-50.094	35	-42.581	33

Note. Only the school rankings based SFEM estimates were presented in the table. Estimate represents fixed effect estimates for SFEM while random effects estimates are presented for UHLMM and AHLMM. SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

Appendix B. Grade 6 Reading Estimates

Table 6. Estimates of the School Effects Obtained from Three VAMs Based on Grade 6 Reading Results

	SFEM		UHLMM		AHLMM	
Rank*	Estimate	School ID	Estimate	School ID	Estimate	School ID
1	60.918	25	44.015	25	-24.438	36
2	48.315	10	35.401	10	-23.498	33
3	36.844	13	27.928	13	-16.870	43
4	27.022	34	21.091	34	-14.180	1
5	24.526	39	19.427	14	-13.321	21
6	23.067	14	18.243	28	-11.845	42
7	22.844	28	18.239	39	-10.927	35
8	22.459	8	17.915	8	-10.611	38
9	17.803	26	13.995	26	-9.648	24
10	15.573	19	12.195	11	-8.487	6
11	15.465	2	11.993	2	-6.880	9
12	15.173	11	11.402	19	-6.373	16
13	12.764	29	10.336	29	-5.099	4
14	11.373	27	7.840	20	-4.015	41
15	10.610	20	7.753	27	-4.011	17
16	10.071	12	7.536	12	-2.407	18
17	8.149	5	6.051	5	-1.944	40
18	6.386	32	4.879	32	-1.771	15
19	5.408	40	4.064	40	-0.895	5
20	3.358	23	2.388	23	-0.093	37
21	3.189	15	2.180	15	0.115	7
22	2.815	31	1.899	31	0.587	12
23	1.086	7	0.756	7	2.018	3
24	0.765	37	0.560	37	2.194	22
25	-1.526	41	-1.245	41	2.909	19
26	-6.618	22	-4.619	22	3.048	44
27	-7.420	16	-6.164	16	3.343	26
28	-9.302	17	-6.806	17	3.457	29
29	-12.984	4	-9.036	44	3.732	2
30	-13.007	3	-9.660	24	3.736	11
31	-13.095	24	-9.730	3	4.714	30
32	-14.541	1	-10.055	4	5.163	14
33	-16.027	30	-11.841	1	6.266	8
34	-17.307	6	-12.910	30	6.417	34
35	-19.570	42	-13.058	6	6.687	28
36	-22.921	18	-14.798	42	7.251	39
37	-24.321	43	-18.471	18	8.767	20
38	-25.337	38	-18.894	43	9.931	31
39	-26.937	21	-19.247	21	10.142	23
40	-29.138	9	-19.764	38	11.042	13
41	-35.153	33	-22.641	9	13.489	27
42	-44.516	36	-28.362	33	14.161	32
43	-54.027	35	-34.359	36	15.269	10
44	-	-	-36.434	35	32.868	25

Note. Only the school rankings based SFEM estimates were presented in the table. Estimate represents fixed effect estimates for SFEM while random effects estimates are presented for UHLMM and AHLMM. SFEM = Simple fixed effects model, UHLMM = Unadjusted hierarchical linear model, AHLMM = Adjusted hierarchical linear model.

Appendix C. SAS Codes Used for Model Estimations

***/ SAS Code for Model1 (SFEM)*/;**

```
proc glm data=GRADE6;
model cahangem = S1 - S43/solution; run;
```

***/ SAS Code for Model2 (UHLMM)*/;**

```
proc mixed data=GRADE6;
class student;
model changem =;
random intercept / type = un sub = school solution;
repeated /type = un sub = student; run;
```

***/ SAS Code for Model3 (AHLMM)*/;**

```
proc mixed data=GRADE6;
class student min pov;
model changem = Z1M Z2 V1 min pov;
random intercept/type= un sub = school solution;
repeated/type = un sub = student; run;
```