Research Report

Robustness of Value-Added Analysis of School Effectiveness

Henry Braun

Yanxuan Qu

Catherine Trapani

April 2008

ETS RR-08-22

Robustness of a Value-Added Assessment of School Effectiveness

Henry Braun[1]
Boston College, Chestnut Hill, MA

Yanxuan Qu, and Catherine Trapani
ETS, Princeton, NJ

April 2008

Abstract

This paper reports on a study conducted to investigate the consistency of the results between 2 approaches to estimating school effectiveness through value-added modeling. Estimates of school effects from the layered model employing item response theory (IRT) scaled data are compared to estimates derived from a discrete growth model based on the analysis of transitions along an ordinal developmental scale. The data were extracted from the longitudinal records maintained in the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K) archive for students remaining in the same school from the beginning of kindergarten through the end of Grade 3. The results of different comparisons indicated that the estimates from the 2 approaches are moderately consistent.

Key words: Value-added modeling, school effectiveness, layered model, robustness

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Introduction

There is substantial interest in education policy circles in the possibility of using summary measures of student growth as the basis for making evaluations of the effectiveness of schools and teachers. Late in 2005, Margaret Spellings (U.S. Department of Education, 2005) encouraged states to submit proposals to incorporate measures of student growth (e.g., the average year-to-year change in students' test scores) into their calculations of adequate yearly progress. There is an important distinction between such statistics and those purported to estimate schools' (or classes') contributions to student learning. Measures of student growth rely on direct calculation and are typically judged against a standard expressed in absolute terms (e.g., Did the average score increase exceed 10 scale score points?), in relative terms (e.g., Was the average score increase 15% larger than last year's?) or in predictive terms (e.g., Are 70% of the students on track to achieve proficiency within 3 years?). Researchers using statistics intended to estimate schools' contributions to student learning, on the other hand, attempt to extract from the variation in students' score trajectories a component that can be attributed directly to enrollment in a particular school or class. These so-called *school* or *class effects* are typically normatively defined; that is, the units of analysis are only compared to one another (e.g., Is this school's average contribution to student learning significantly larger or smaller than the average contribution of the typical school in the district?).

Value-added modeling (VAM) is the generic name attached to the statistical machinery used to obtain estimates of such school or class effects.[2] Early work in this area is due to Sanders, Saxton, and Horn (1997) and Webster (2005). Econometricians also have weighed in with their own approaches to VAM. For an overview, see Sass and Harris (2006). McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) offered a technical review of VAM, whereas both Wainer (2004) and Lissitz (2005) provided surveys of current research in the area. For a nontechnical introduction, see Braun (2005a).

Despite the widespread enthusiasm for VAM, a number of concerns have been identified in employing VAM results, particularly in high-stakes settings. One such concern centers on the problem of drawing causal inferences from observational data (Braun, 2005a; Raudenbush, 2004).[3] Another focuses on the nature of the test score scale. Most VAMs require longitudinal student test data. Student performance is represented by a point on a score scale derived through vertical linking. The score scale is usually treated as if it were an interval scale. A natural question arises: How robust are the results to the choice of scale and the associated assumption of the interval scale property? (Note that this issue is more salient for value-added analyses than for straightforward reporting of growth along the scale.)

We describe a project undertaken to investigate this question, employing data from the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K; Pollack, Atkins-Burnett, Rock, & Weiss, 2005). We compare two different approaches to the estimation of school effects. One is the VAM proposed by Sanders et al. (1997) that is often referred to as the *layered model.* [4] The layered model makes use of test score data that are reported on a vertically linked scale. [5] The second approach (Braun, 2005b) makes use of transition probabilities, defined on a developmental scale, that represent student growth from one year to the next.[6] The transition approach requires neither vertically linked scales nor the interval scale assumption. This approach is called the developmental trajectory growth (DTG) model. Thus, the two approaches make rather different uses of the raw test data. The degree to which they yield similar rankings of schools (or classes) provides some evidence with respect to the question of robustness.

It is important to note that the robustness at issue here is statistical robustness against a specific assumption regarding the nature of the score scale as well as the particular statistical model applied to the data represented on that score scale. As such, this study provides a more stringent test of robustness than typically found in the literature, where different models are applied to the same data (McCaffrey et al., 2004; Tekwe et al., 2004).

However, a much deeper question is to what extent one is entitled to make causal inferences about schools or teachers' relative effectiveness on the basis of a statistical analysis of observational data, subject to selection biases of various sorts. Certainly a finding of a lack of statistical robustness would undercut the possibility of making such causal inferences. However, a finding of reasonable statistical robustness in this context is only the first step in laying a credible foundation for the kinds of inferences that policymakers would like to make on the basis of student score trends.

Data

The ECLS-K is a long-term longitudinal study funded by the U. S. Department of Education through the National Center for Education Statistics (Tourangeau et al., 2004). A nationally representative sample of schools was selected for the study, and a random sample of students from each school was selected to participate in the study. The ECLS-K, third-grade, public-use data file contains information from five waves of data collection: (a) kindergarten in fall (denoted as Fall-K), (b) kindergarten in spring (Spring-K), (c) first grade in fall (Fall-1), (d) first grade in spring (Spring-1), and (e) third grade in spring (Spring-3). See Table 1. Students took tests in both reading and mathematics in each wave. Since the third wave comprised only a 30% sample, we did not include it in the analysis. Thus, we are able to study the academic growth for a cohort of students over three transitions. Furthermore, because the study has a purely methodological focus, we restricted attention to those students who attended the same school for all four waves and had to be enrolled in a school with at least 10 such students. This reduced the sample size to 8,853 students in 619 schools, about half of the full sample. The analysis sample is similar to the full sample with respect to the distribution by gender, race, and test performance at Fall-K. The main difference is that the analysis sample has about 6% more White students and about 5% fewer Black, non-Hispanic students (see Appendix A for details).

Table 1

Selected Characteristics of the Early Childhood Longitudinal Study—Kindergarten Cohort, Third-Grade, Public-Use Data File

Data collection wave	Date of collection	Sample
1. Fall-K	Fall 1998	21,260 students in 866 schools (full sample)
2. Spring-K	Spring 1999	21,260 students in 866 schools (full sample)
3. Fall-1	Fall 1999	5,975 students in 293 schools (30% subsample of the total base-year schools)
4. Spring-1	Spring 2000	17,212 students in 866 schools (full sample)
5. Spring-3	Spring 2002	15,305 students in 866 schools (full sample)

Note. Waves abbreviated, e.g., Fall-K = fall kindergarten, Spring-1 = spring first grade.

Our goal was to compare two different approaches to estimating the value-added associated with each school. In the ECLS-K, student performance in a subject is represented in a number of different ways. We chose two. The first is an ordinal developmental scale, constructed especially for the ECLS-K. At each wave, a student is assigned to one of nine proficiency levels, based on the application of an algorithm devised by the developers of the test battery.[7] The second is a standard item response theory (IRT) scale constructed from the full battery of items employed from kindergarten through Grade 3. After each wave, students' IRT scale scores are derived and added to the database.

Analyses: Method 1

Method 1 is denoted as DTG, a preliminary version of which was suggested by Braun (2005b). The DTG approach is based on considering the conditional probabilities of students moving from one proficiency level to another over the course of a transition (e.g., from Fall-K to Spring-K). Again, since the focus here is on a comparison of methodologies, the confounding of summer growth with school contributions is not of primary concern.

To illustrate the DTG approach, let i = 1, 2, … , I denote the observed proficiency levels at Fall-kindergarten, and let j = 1, 2, … , J denote the observed proficiency levels at Spring-Kindergarten. The relevant matrix for the case I = J = 4 is presented in Table 2.

Table 2

Developmental Trajectory Growth (DTG) Matrix

		Spring-K			
		1	2	3	4
Fall-K	1				
	2			$P_{3\|2}$	
	3				
	4				

The conditional transition probability of a student moving from proficiency level i in Fall-K to proficiency level j in Spring-K is denoted as $P_{j|i} = P\{i \to j\} = \dfrac{P(i, j)}{P(i)}$, where the numerator denotes the joint probability of a student in cell *(i,j)* and the denominator denotes the marginal probability of a student being at level *i* at Fall-K. In what follows, these probabilities are estimated by the corresponding sample proportions.

Typically, the value-added measure of a school is defined normatively. For the DTG, we compare the pattern of transitions observed in a school to what would have been observed, given the row marginals, had the transition probabilities based on the aggregate experience of all schools in the sample been in operation in that school.[8] For each cell in the transition matrix, we compute a quantity equal to the deviation between observed and expected, weighted by the numerical label associated with the final level.[9] The weighted deviations are summed over the cells and standardized by dividing by the total number of students in the school. This statistic is denoted as the grand mean weighted deviation (GMWD). The formula for calculating the GMWD for school *k* is given by:

$$\text{GMWD (for school } k\text{)} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left\{j*\left(n_{ijk}-\frac{N_{ij}}{\sum_{j}N_{ij}}*\sum_{j}n_{ijk}\right)\right\}}{\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ijk}}.$$

Here,

n_{ijk} = number of students in school k who moved from proficiency level i to proficiency level j,

N_{ij} = total number of students in the sample (aggregated over all schools) who moved from proficiency level i to proficiency level j.

Then $\frac{N_{ij}}{\sum_{j}N_{ij}}$ is the corresponding total sample transition probability, $\sum_{j}n_{ijk}$ is the level *i* row total for school *k*, and their product yields the expected number of students in cell (*i*, *j*) for school *k* if the total sample transition probability were in force. The difference between n_{ijk} and the expected number of students in cell (*i*, *j*) is the deviation for the cell (*i*, *j*), which is then weighted by *j*. Thus, schools are given more credit for moving students to higher levels. A school with consistently large positive deviations at higher developmental levels will be assigned a large, positive GMWD. Note that with the DTG approach, a school's estimated value-added takes account of the developmental status profile of its student cohort at the start of the transition.[10]

Analyses: Method 2

Method 2 involves computing estimates of school value-added by employing the layered model of Sanders et al. (1997). The original layered model is a multivariate, longitudinal, mixed-effects model. In this implementation, average performance for each combination of season, grade, and subject is treated as a fixed effect, whereas the corresponding school effects are treated as random. The computations were carried out by a program developed by the RAND Corporation (Lockwood, McCaffrey, Mariano, & Setodji, 2006). The program takes as input students' scores on the vertically linked IRT scales for mathematics and reading and fits a bivariate, longitudinal model. The program relies on a Bayesian formulation and yields posterior distributions for each parameter of interest. The estimated value-added for a school for a particular subject-grade combination is the mean of the corresponding posterior distribution for that school. An estimate of the standard error of each school effect estimate is also provided. See Appendix B for the particular model employed.

Results

Results From Method 1: DTG

The population or aggregate transition matrices for both reading and math are fundamental to the analysis. Some students are missing scores in 1 or more years and are therefore not included in the corresponding transition matrix. As one would expect, the distribution of students across developmental levels shifts toward higher levels at higher grades. Typically, students move up one or more levels through a transition. Note that there is no correspondence in the meaning of a specific level (e.g., Level 4) across subjects. Inspection of the transition matrices indicates that, at least from a statistical point of view, the developmental scale is meaningful and appropriate, with only modest ceiling effects by the end of Grade 3. For illustrative purposes, we present two of these matrices.

Table 3 displays the transition matrix for reading for Spring-K to Spring-1. In Spring-K, there is a strong mode at Level 4, whereas in Spring-1, there is a mode at Level 5 with a near mode at Level 6. Notably, very few children remain at the same level or lose ground, especially if they start out at Level 5 or higher. At the same time, many children grow by three or more levels, especially if they start out at lower levels.

Table 4 displays the transition matrix for mathematics for Spring-K to Spring-1. In Spring-K there is a strong mode at Level 4, and in Spring-1 there is a strong mode at Level 5. Most students move up one level, although students who start out at lower levels typically move up two levels. Relatively few students remain at the same level or fall behind over this period.

Next, we present descriptive statistics for the GMWD for each transition for both reading and math. These statistics were computed both for all the schools and for schools with at least 10 students. (The maximum number of students in a school was 23.) The results were quite similar, so we only present the latter set of results. For each subject-transition combination, the set of GMWDs are centered near zero and approximately normally distributed. Table 5 presents a few

Table 3

Population Transition Matrix: Spring-K to Spring-1, Reading

		1	2	3	4	5	6	7	8	9	Total
1	Frequency	18	61	88	117	81	12	2	0	0	379
	Row pct	4.75	16.09	23.22	30.87	21.37	3.17	0.53	0.00	0.00	
2	Frequency	2	44	110	311	534	157	28	0	0	1,186
	Row pct	0.17	3.71	9.27	26.22	45.03	13.24	2.36	0.00	0.00	
3	Frequency	1	6	42	236	820	449	104	10	0	1,668
	Row pct	0.06	0.36	2.52	14.15	49.16	26.92	6.24	0.60	0.00	
4	Frequency	0	2	15	157	1,024	1,350	443	58	6	3,055
	Row pct	0.00	0.07	0.49	5.14	33.52	44.19	14.50	1.90	0.20	
5	Frequency	0	0	0	2	53	357	237	55	18	722
	Row pct	0.00	0.00	0.00	0.28	7.34	49.45	32.83	7.62	2.49	
6	Frequency	0	0	0	0	7	63	115	53	14	252
	Row pct	0.00	0.00	0.00	0.00	2.78	25.00	45.63	21.03	5.56	
7	Frequency	0	0	0	0	2	6	27	23	13	71
	Row pct	0.00	0.00	0.00	0.00	2.82	8.45	38.03	32.39	18.31	
8	Frequency	0	0	0	0	0	0	5	2	3	10
	Row pct	0.00	0.00	0.00	0.00	0.00	0.00	50.00	20.00	30.00	
9	Frequency	0	0	0	0	0	0	0	0	2	2
	Row pct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	
Total	Frequency	21	113	255	823	2,521	2,394	961	201	56	7,345[a]
	Row pct	0.29	1.54	3.47	11.20	34.32	32.59	13.08	2.74	0.76	100.00

[a] 1,508 students are missing due to not having one or more reading scores.

summary statistics. Within subject, the dispersions across transitions are very similar. The standard deviations for the reading transitions are about one third larger than those for math.

Plots of GMWD against school sample size (for schools with 10 or more students) indicate that estimated school effects are weakly related to school size. Figure 1 illustrates the patterns for the Spring-K to Spring-1 transitions for reading and math, respectively. The correlations between GMWD and school size are displayed in Table 6.

Table 4

Population Transition Matrix: Spring-K to Spring-1, Math

		1	2	3	4	5	6	7	8	Total
1	Frequency	6	25	34	19	5	1	0	0	90
	Row Pct	6.67	27.78	37.78	21.11	5.56	1.11	0.00	0.00	
2	Frequency	3	50	182	321	251	20	2	0	829
	Row Pct	0.36	6.03	21.95	38.72	30.28	2.41	0.24	0.00	
3	Frequency	0	4	93	780	1,177	168	10	0	2,232
	Row Pct	0	0.18	4.17	34.95	52.73	7.53	0.45	0	
4	Frequency	0	2	12	464	1,974	839	95	4	3,390
	Row Pct	0.00	0.06	0.35	13.69	58.23	24.75	2.80	0.12	
5	Frequency	0	1	2	37	506	633	130	21	1,330
	Row Pct	0.00	0.08	0.15	2.78	38.05	47.59	9.77	1.58	
6	Frequency	0	0	0	0	10	117	74	8	209
	Row Pct	0.00	0.00	0.00	0.00	4.78	55.98	35.41	3.83	
7	Frequency	0	0	0	0	1	3	11	0	15
	Row Pct	0.00	0.00	0.00	0.00	6.67	20.00	73.33	0.00	
8	Frequency	0	0	0	0	0	0	0	1	1
	Row Pct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	
Total	Frequency	9	82	323	1621	3924	1781	322	34	8,096[a]
	Row Pct	0.11	1.01	3.99	20.02	48.47	22.00	3.98	0.42	100.00

[a] 757 students are missing due to not having one or more math scores.

Table 5

Summary Statistics for Grand Mean Weighted Deviation (GMWD) for Six Transitions

	Reading			Math		
Statistic	Fall-K to Spring-K	Spring-K to Spring-1	Spring-1 to Spring-3	Fall-K to Spring-K	Spring-K to Spring-1	Spring-1 to Spring-3
# schools	412	461	479	497	552	569
Mean	0.00	0.02	0.00	-0.01	0.00	0.00
SD	0.39	0.37	0.41	0.28	0.27	0.32
Skewness	-0.10	-0.50	-0.23	-0.15	-0.32	-0.02

Note. Schools have >= 10 students. K = kindergarten; Spring-1 = spring first grade, etc.

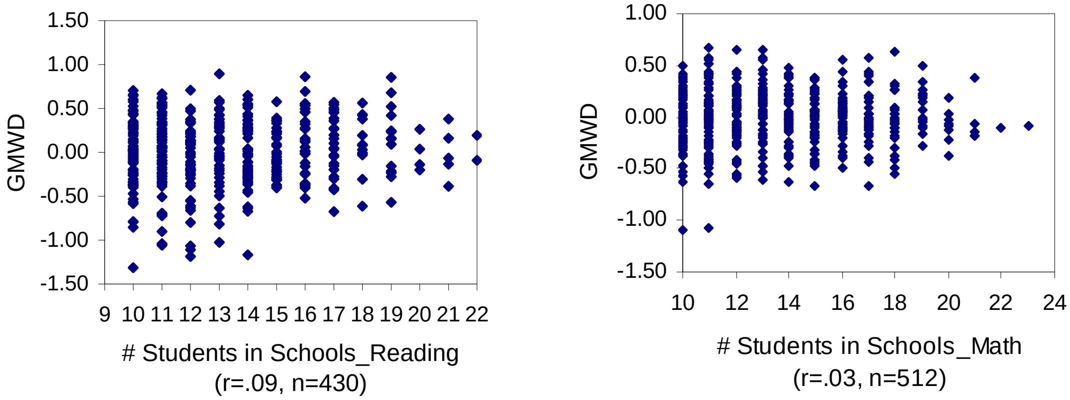

Figure 1. **Grand mean weighted deviation (GMWD) against school sample size for Spring-K to Spring-1.**

Table 6

Correlations Between Grand Mean Weighted Deviation and School Size

Transition	Reading	Math
Fall-K to Spring-K	0.01 (*n* = 412)	0.15 (*n* = 497)
Spring-K to Spring-1	0.10 (*n* = 461)	0.06 (*n* = 552)
Spring-1 to Spring-3	0.26 (*n* = 479)	0.05 (*n* = 569)

Note. Schools have >= 10 students.

The relationships among sets of estimated school effects are also of interest. We found that within a subject, school effects across transitions are not correlated; for example, GMWDs for the Fall-K to Spring-K transition are uncorrelated with the GMWDs for the Spring-K to Spring-1 transition. This lack of correlation may be because students typically have different teachers in different grades. Table 7 displays the correlations, whereas Figure 2 presents illustrative scatter-plots.

On the other hand, for a particular transition, estimated school effects for math and reading are moderately correlated. The correlations are 0.43, 0.39, and 0.42 for the first, second, and third transitions, respectively. One would expect that within-transition correlations would be higher than across-transition correlations, since, in the former case, students are exposed to the same teacher.

Finally, we computed for each school the average GMWD for reading over the three transitions and the average GMWD for math over the three transitions. Figure 3 illustrates the strong correlation of 0.70 between the two sets of estimated average effects.

Table 7

Correlations of Grand Mean Weighted Deviation Between Transitions Within Subjects (Reading Below Diagonal and Math Above Diagonal)

Transition	Fall-K to Spring-K	Spring-K to Spring-1	Spring-1 to Spring-3
Fall-K to Spring-K		0.05 (*n* = 576)	0.18 (*n* = 576)
Spring-K to Spring-1	-0.07 (*n* = 575)		0.04 (*n* = 619)
Spring-1 to Spring-3	0.03 (n = 575)	0.12 (*n* = 618)	

Note. K = kindergarten; Fall-1 = fall first grade.

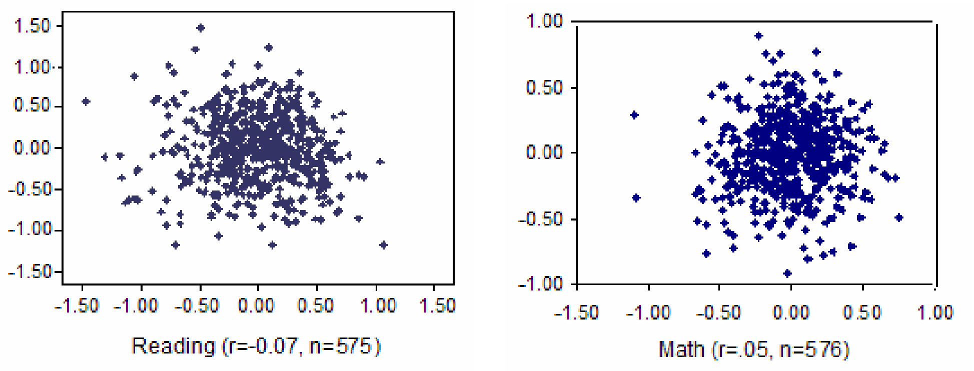

***Figure 2.* Grand mean weighted deviation (GMWD) Fall-K to Spring-K versus GMWD Spring-K to Spring-1.**

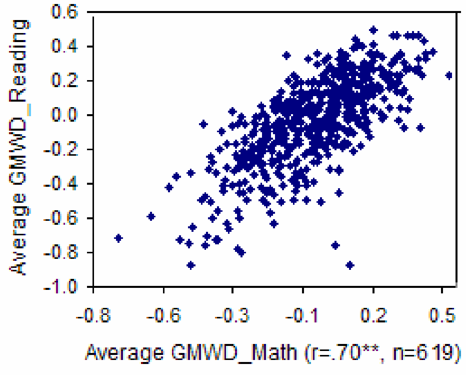

***Figure 3.* Average grand mean weighted deviation (GMWD) reading versus average GMWD math.**

One of the reasons for the interest in VAMs is the expectation, generally borne out empirically, that school value-added estimates are only weakly correlated with the characteristics of students and schools. Of course, that stands in contrast to a situation with statistics based on students' (current) status. Accordingly, we examined the relationship between schools' GMWD for a particular subject and transition and a measure of the schools' student proficiency distributions at the start of the transition. For the latter, we chose to use a weighted sum of the proportions of students at each developmental level, with the weights equal to the numerical labels attached to the levels (i.e., a type of mean). For school *k*, the measure is given the following formula and is referred to as input:

$$\sum_j j * \frac{\sum_j n_{ijk}}{\sum_i \sum_j n_{ijk}} = \sum_j j * \left(\frac{\text{\# of students in each proficiency level before transition}}{\text{total \# of students in the school}} \right).$$

In view of the weights employed in constructing the GMWD, one would expect the correlations to be somewhat positive. Indeed, the correlations were small, ranging from 0.14 to 0.23 in math and from 0.26 to 0.34 in reading. Figure 4 displays the scatter-plots for the transition from Spring-K to Spring-1 for reading and math, respectively. Note that for each level of input, the GMWD span a broad range of values.

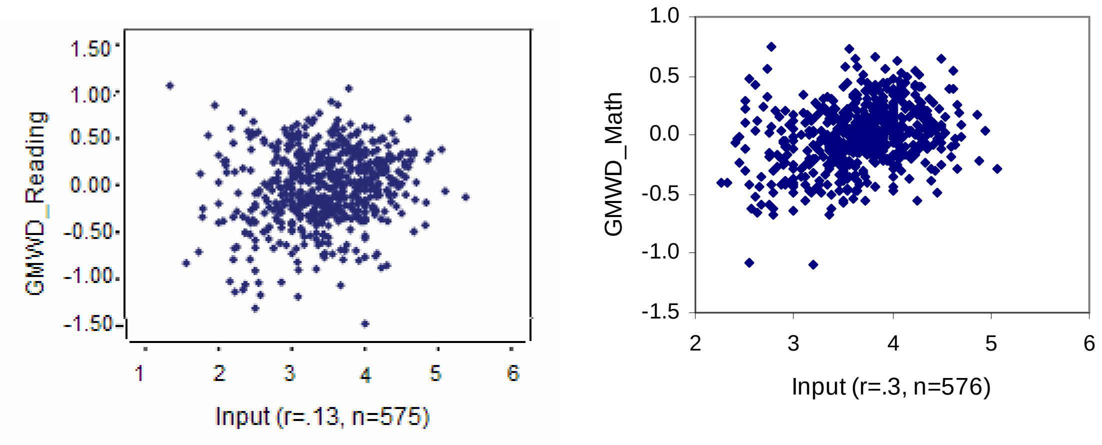

***Figure 4.* Grand mean weighted deviation (GMWD) versus input status from Spring-K to Spring-1.**

We also examined the relationship between estimated school effects and school poverty. Poverty was categorized by two levels: less than or equal to 50% of students eligible for free lunch, and more than 50% of students eligible for free lunch. We ran an analysis of covariance on school effects for each combination of subject and transition, with input status as the covariate and poverty level as the discrete predictor. With the exception of reading (Fall-K to Spring-K), the poverty contrast was significant, but the regression on input status was not. The range of R^2, however, was from 0.04 to 0.14. Practically speaking, then, the relationship between this measure of value-added and these school characteristics is rather weak.

Results From Method 2: The Layered Model

We now turn to the results from fitting the layered model to the data. We first present descriptive statistics along with the corresponding histograms. For each subject transition, the ensemble of estimates is centered at zero and approximately normal. See Table 8.

Table 8

Summary Statistics for the Layered Model for Six Transitions

	Reading			Math		
	Fall-K to Spring-K	Spring-K to Spring-1	Spring-1 to Spring-3	Fall-K to Spring-K	Spring-K to Spring-1	Spring-1 to Spring-3
Mean	-0.00	0.02	-0.01	-0.00	0.01	-0.00
SD	2.18	3.98	3.39	1.59	2.58	2.79
Skewness	0.36	0.09	-0.15	0.46	0.01	-0.13

Note. N = 337; only 337 schools have at least 10 students with both reading and math scores for all four waves. K = kindergarten; Spring-1 = spring first grade, etc.

The graphs in Figure 5, termed *caterpillar plots,* depict the estimated school effect (the mean of the posterior distribution) plotted against the rank of the mean. An error bar extending two standard deviations (also based on the posterior distribution) in each direction is superimposed on the graph. It is evident that only a small proportion of the estimated effects were statistically significantly different from zero.

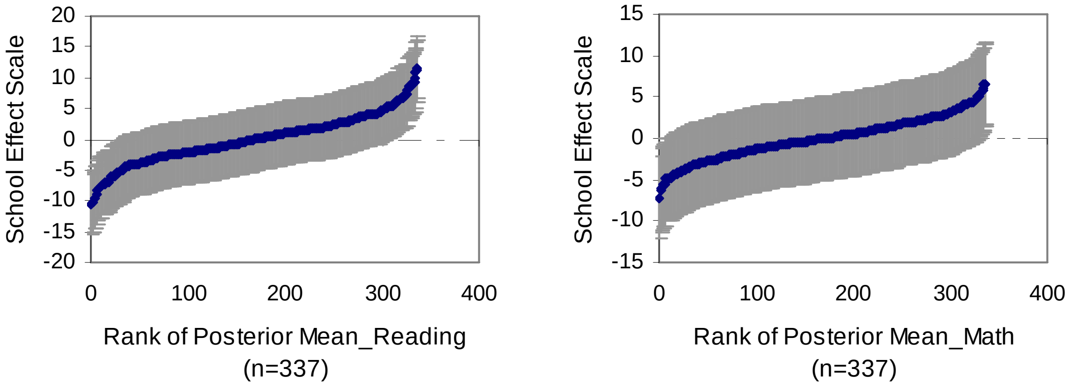

***Figure 5.* Confidence intervals for layered model estimates for each school from Spring-K to Spring-1.**

As was the case for GMWDs, we found that within a subject, school effects across transitions were not correlated when estimated by the layered model. Table 9 displays the correlations.

Table 9

Correlations of Layered Model Estimates Between Transitions Within Subjects (Reading Below Diagonal and Math Above Diagonal)

	Fall-K to Spring-K	Spring-K to Spring-1	Spring-1 to Spring-3
Fall-K to Spring-K		0.05	-0.10
Spring-K to Spring-1	-0.04		0.22
Spring-1 to Spring-3	-0.12	0.14	

Note. N = 337. K = kindergarten; Spring-1 = spring first grade, etc.

On the other hand, estimated school effects for math and reading based on the layered model were moderately correlated within transitions (correlations range from 0.48 to 0.62). When the subject-specific school effects were averaged over the three transitions, the between-subject correlation was quite strong (r = 0.73), and very similar to the correlation of 0.70 that we found for the average GMWD estimates (Figure 6).

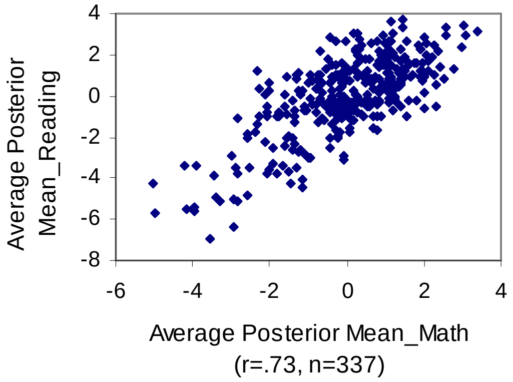

***Figure 6.* Layered model estimates for reading and math.**

Estimated school effects from the layered model are only weakly correlated with schools' mean input distribution, based, as before, on the proportions of students at each developmental level. The correlations range from 0.15 to 0.30 for math and from 0.02 to 0.28 for reading. Scatter-plots for transition from Spring-K to Spring-1 are displayed in Figure 7. We also ran an analysis of covariance on school effects for each combination of subject and transition, with input status as the covariate and poverty level as the discrete predictor. The results were very similar to those for GMWD, and, as before, we conclude that the relationship between this value-added measure and these school characteristics is rather weak.

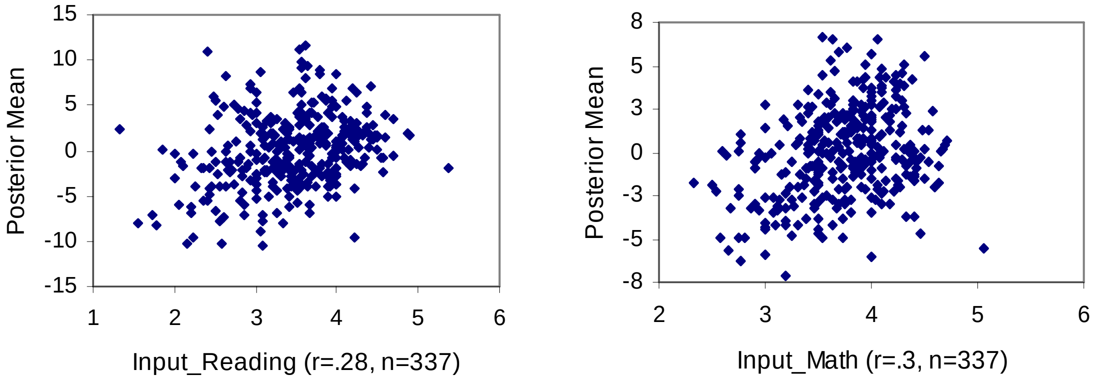

***Figure 7.* Layered model estimates versus input status from Spring-K to Spring-1.**

Comparison Between DTG and Layered-Model Methods

Our chief interest lay in comparing the estimated school effects from the two methods. This could be accomplished most directly through a series of scatter-plots, one for each subject-transition combination. As Figures 8 and 9 indicate, the estimates were highly correlated. In math, they ranged from 0.75 to 0.82, and in reading from 0.77 to 0.86 (Figure 8). When estimated school effects were averaged across transitions within subject, the correlations were 0.84 for math and 0.88 for reading (Figure 9).

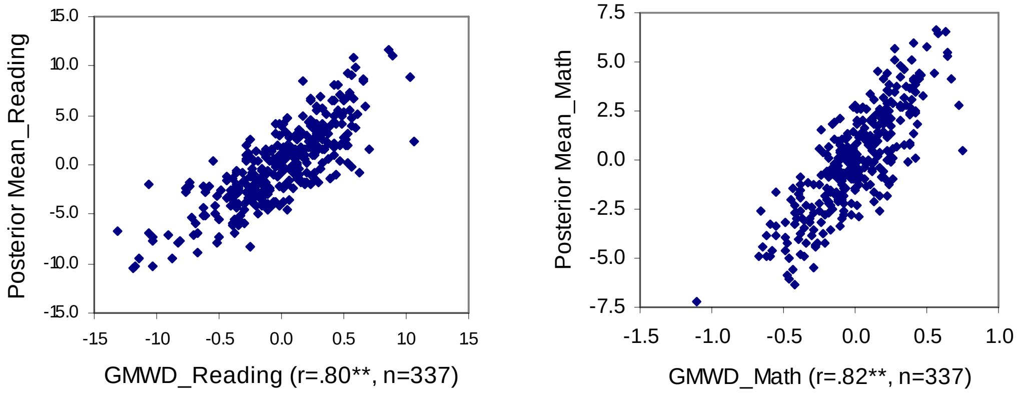

***Figure 8.* Layered model estimates versus developmental trajectory growth estimates from Spring-K to Spring-1. GMWD = grand mean weighted deviation.**

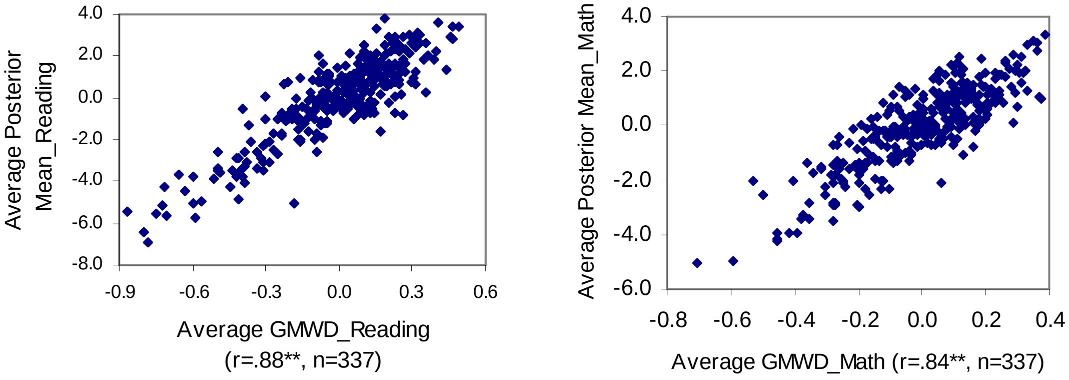

***Figure 9.* Average layered model estimates versus average developmental trajectory growth estimates. GMWD = grand mean weighted deviation.**

Another way to compare the two methods is to examine the consistency of the classifications by schools within transition, restricting attention to the 337 schools used for the layered model analysis. For this model, we identified a school as statistically different from the average if the probability that the school effect was greater than zero was at least 0.9. Those schools are labeled with a "+". Similarly, schools whose estimated effect had at least a probability of .9 of being less than zero were labeled with a "-". The remaining schools were labeled with a "0". The categorization was done separately for reading and math.

Since we had not derived estimated variances for the GMWD, we had to adopt a different strategy to designate schools with extreme estimated effects. As it happens, the layered model identified approximately 10% of schools as "+" and 10% as "-". Accordingly, we simply classified schools by the decile of the distribution of DTG-estimated school effects for the particular subject transition. For each subject-transition combination, we then cross-classified schools into 30 categories based on their estimated effects from each method. The results are presented in Table 10 and are a reasonable basis for comparing the consistency of the results of the two methods, given that the marginal numbers of schools in both sets of end categories are similar.

We were particularly interested in whether schools that were designated significantly different from the average by the layered model also fell in the corresponding tails of the distribution of estimated effects from the DTG model. In fact, this was the case about 75% of the time, if the tail is defined to be the extreme two deciles. In general, the correspondence was slightly better in math than in reading. Finally, for the Fall-K to Spring-K and Spring-K to Spring-1 transitions, about 15% of schools labeled "0" under the layered model fell in each tail of the DTG distribution. For the Spring-1 to Spring-3 transition, the statistic was closer to 20%.

Conclusions

A critical issue in considering the use of value-added estimates is the sensitivity of these estimates to various model assumptions. One set of assumptions is related to the construction of the score scale from the raw data as well as the metric properties of that scale. The DTG approach was introduced as a simple alternative to the more complex layered model—one that makes different assumptions about the characteristics of students' test scores. From a practical point of view, the DTG is not "ready for prime time": It is inefficient because it does not "borrow strength" across grades or subjects, and at this point there are not estimated standard

Table 10

Classification of Schools by Developmental Trajectory Growth (DTG) Model and Layered Model Within Each Transition for Each Subject

		DTG										
		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Total
Reading: Fall-K to Spring-K												
Layered	-	21	8	8	3	1	0	0	0	0	0	41
model	0	12	26	26	31	32	34	29	29	20	14	253
	+	0	0	0	0	0	0	5	5	14	19	43
Reading: Spring-K to Spring-1												
Layered	-	20	11	4	2	1	0	0	0	0	0	38
Model	0	12	23	32	30	31	30	33	25	24	13	256
	+	0	0	0	0	1	2	2	7	11	20	43
Reading: Spring-1 to Spring-3												
Layered	-	19	4	2	5	4	0	0	0	0	0	34
Model	0	15	30	30	31	27	33	32	31	30	23	282
	+	0	0	0	0	0	1	3	2	5	10	21
Math: Fall-K to Spring-K												
Layered	-	11	8	2	0	0	1	0	0	0	0	22
Model	0	22	26	32	34	33	33	32	28	29	13	282
	+	0	0	0	0	0	1	2	4	6	20	33
Math: Spring-K to Spring-1												
Layered	-	20	13	2	1	0	0	0	0	0	0	36
Model	0	13	22	30	33	30	40	34	27	23	13	265
	+	0	0	0	0	0	0	0	4	11	21	36
Math: Spring-1 to Spring-3												
Layered	-	18	4	3	2	0	0	0	0	1	0	28
Model	0	15	29	32	31	33	32	32	29	28	17	278
	+	0	0	0	0	0	3	3	4	4	17	31

Note. Abbreviations as follows, e.g.: Spring-K = spring kindergarten; Fall-1 = fall first grade.

errors to attach to the value-added estimates. Nonetheless, our analyses show that the DTG approach has reasonable properties and serves a methodological purpose of generating a plausible comparison set of estimates.

The results of this study indicate a reasonable degree of robustness in the estimates of value-added. That is, the correlations between the two sets of estimates are quite respectable. This is the case although both sets of estimates are based on the analysis of a single cohort. Ideally, results should be averaged over (say) three cohorts to enhance the stability of the estimates. [11] In fact, when the school value-added estimates are averaged over transitions, the correlations are quite high.

In practical applications, interest often centers on schools with value-added estimates at the extremes of the distribution (e.g., the lowest and highest deciles). The stability of decile location across methods is another aspect of robustness that merits consideration. In our version of this analysis, we find a moderate level of stability. In sum, these results give some support to the judicious use of value-added estimates in school improvement efforts.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Betebenner, D. W. (2007, January). *Growth as a description of process*. Paper presented at the 2007 festschrift to the life and work of Professor Robert L. Linn, Los Angeles.

Braun, H. I. (2005a). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: ETS.

Braun, H. I. (2005b). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.)*, Value-added models in education: Theory and applications* (pp. 19-38). Maple Grove, MN: JAM Press.

Lissitz, R. (Ed.). (2005). *Value-added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2006). Bayesian methods for scalable multi-subject value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(21), 67-101.

Pollack, J., Atkins-Burnett, S., Rock, D., & Weiss, M. (2005). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K), psychometric report for the third grade* (NCES No. 2005-062). Washington, DC: National Center for Education Statistics.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement? The ninth annual William H. Angoff Memorial Lecture*. Princeton, NJ: ETS.

Sanders, W. L., Saxton, A., & Horn, B. (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 evaluational measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Sass, T. R., & Harris, D. (2006). *Value-added models and the measurement of teacher Quality.* Unpublished manuscript.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-35.

Tourangeau, K., Brick, M., Byrne, L., Le, T., Nord, C., West, J., et al. (2004). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K), third grade methodology report* (NCES No. 2005-018). Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2005, November 18). *Secretary Spellings announces growth model pilot, addresses chief state school's officers' annual policy forum in Richmond.* Retrieved April 7, 2008, from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

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), 1-2.

Webster, W. J. (2005). The Dallas school-level accountability model: The marriage of status and value-added approaches. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 233-271). Maple Grove, MN: JAM Press.

Notes

[1] Much of the work reported here was done while Henry Braun was a distinguished presidential appointee at Educational Testing Service, which supported this research.

[2] Typically, results of value-added analyses at the classroom level are denoted as *teacher effects*. However, strictly speaking, the statistical analysis yields estimated effects due to classroom assignment. Inferences that these are due to the teacher in the classroom involve yet other technical issues. Accordingly, we will use the term *class effect*.

[3] Phrases such as *school effect* or *class effect* are not intended to invoke a causal interpretation of the estimates obtained through statistical analysis. Use of the term *effect* is traditional in statistics and harkens back to the analysis of agricultural experiments, for which a direct causal interpretation was indeed plausible. In nonrandomized studies, interpreting school effects as measures of school effectiveness is more problematic.

[4] The commercial product based on the layered model is known as the educational value-added assessment system (EVAAS) and is marketed by SAS Incorporation.

[5] It is not strictly necessary for the data to be on a single scale, but most applications of the layered model do involve such data.

[6] Betebenner (2007) also has investigated the use of transition matrices to model student growth.

[7] The defined developmental levels are numbered 2–8. Level 1 is reserved for those students who have not reached the first developmental level, Level 2. See Pollack et al. (2005).

[8] This is usually known as indirect standardization.

[9] This weighting gives more influence or weight to transitions into higher developmental levels. This is reasonable if each higher level corresponds to a qualitatively more complex skill standard. Other choices of weights are certainly possible. For example, the transition from *i* to *j* could be weighted by a factor (j - i). Such a weighting scheme treats changes in level as basic. There is an implicit value judgment in the choice of a weighting scheme.

[10] However, two schools with the same set of transition probabilities would not necessarily be assigned the same GMWD irrespective of the input distribution.

Appendix A

Sample Characteristics

Table A1

Distribution of Gender in the Gull Sample and the Analysis Sample

	Frequency		Percentage	
Gender	Full sample[a]	Analysis sample	Full sample[a]	Analysis sample
Male	10,866	4,448	51.14	50.24
Female	10,381	4,405	48.86	49.76

[a]Missing = 13.

Table A2

Distribution of Race in the Full Sample and the Analysis Sample

	Frequency		Percentage	
Race	Full sample[a]	Analysis sample[b]	Full sample[a]	Analysis sample[b]
White, non-Hispanic	11,741	5,454	55.41	61.7
Black or African American, non-Hispanic	3,210	887	15.15	10.03
Hispanic, Race specified	1,757	699	8.29	7.91
Hispanic, Race not specified	2,005	706	9.46	7.99
Asian	1,364	593	6.44	6.71
Native Hawaiian, other Pacific Islander	220	123	1.04	1.39
American Indian or Alaska Native	379	160	1.79	1.81
More than one race, non-Hispanic	514	218	2.43	2.47

[a]Missing = 70. [b]Missing = 13.

Appendix B

Description of the Layered Model

Let *i* index students, *j* index transitions, and n_i index the school attended by student *i*. Then, the bivariate model is of the form

$$\left(y_{ij}, z_{ij}\right) = \left(\mu_j, \gamma_j\right) + \sum_{k \le j} \left(\theta_{n_i k}, \varphi_{n_i k}\right) + \left(\varepsilon_{ij}, \delta_{ij}\right); \quad (j = 1, 2, 3)$$

where y_{ij} represents the student's reading score; z_{ij} represents the student's math score; μ_j represents the average reading score over the whole population; γ_j represents the average math score over the whole population; $\theta_{n_i k}$ represents a school effect in reading; $\varphi_{n_i k}$ represents a school effect in math; and ε_{ij} and δ_{ij} are the random error terms in reading and math, respectively.

The parameters $\{\mu\}$ and $\{\gamma\}$ are assumed fixed, whereas the parameters $\{\theta\}$ and $\{\varphi\}$ are assumed random and jointly independent. Let $\underline{\varepsilon_i} = \left(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}\right)$ and $\underline{\delta_i} = \left(\delta_{i1}, \delta_{i2}, \delta_{i3}\right)$, then $\left(\underline{\varepsilon_i}, \underline{\delta_i}\right)$ are assumed to follow a multivariate normal distribution with mean vector zero and an unstructured positive definite covariance matrix. Conditional on the other parameters in the model, $\left(\underline{\varepsilon_i}, \underline{\delta_i}\right)$ are assumed to be independent across students.

For the three transitions, the model can be represented as

$$\left(y_{i1}, z_{i1}\right) = \left(\mu_1, \gamma_1\right) + \left(\theta_{n_i 1}, \varphi_{n_i 1}\right) + \left(\varepsilon_{i1}, \delta_{i1}\right);$$

$$\left(y_{i2}, z_{i2}\right) = \left(\mu_2, \gamma_2\right) + \left(\theta_{n_i 1}, \varphi_{n_i 1}\right) + \left(\theta_{n_i 2}, \varphi_{n_i 2}\right) + \left(\varepsilon_{i2}, \delta_{i2}\right);$$

$$\left(y_{i3}, z_{i3}\right) = \left(\mu_3, \gamma_3\right) + \left(\theta_{n_i 1}, \varphi_{n_i 1}\right) + \left(\theta_{n_i 2}, \varphi_{n_i 2}\right) + \left(\theta_{n_i 3}, \varphi_{n_i 3}\right) + \left(\varepsilon_{i3}, \delta_{i3}\right).$$

The layered model is sometimes referred to as a *persistence model* because the school effects at one transition are carried over to succeeding transitions. Lockwood et al. (2006) dealt with a general model, termed a variable persistence model, in which the school effect is dampened in succeeding transitions. They employed a fully specified set of Bayesian priors for all model parameters and explained how to carry out the requisite computations to obtain Bayesian estimates of the parameters. Lockwood et al. also made provision for including a vector of student covariates. However, in the opinion of the developers of the layered model (Ballou, Sanders, & Wright, 2004), it is unnecessary to adjust for differences in student characteristics, because the model exploits the covariances within transitions across subjects and between transitions within subjects, so that students are effectively treated as their own controls.