

ARRS An Empirical Investigation of Growth Models

Ye Tong

Kimberly O'Malley

Pearson Educational Measurement

August 2006



rr0601

*Using testing and
assessment to
promote learning*

Pearson Educational Measurement (PEM) is the largest comprehensive provider of educational assessment products, services and solutions. As a pioneer in educational measurement, PEM has been a trusted partner in district, state and national assessments for more than 50 years. PEM helps educators and parents use testing and assessment to promote learning and academic achievement.

PEM Research Reports provide dissemination of PEM research and assessment-related articles prior to publication. PEM reports in .pdf format may be obtained at:

<http://www.pearsonedmeasurement.com/research/research.htm>



Introduction

With the recent legislation of NCLB, there has been an increasing interest to measure students' growth over the course of their schooling. In the meanwhile, policymakers are inquiring about the possibilities of incorporating the measurement of growth into the accountability system. The test scores earned one year provides status information (e.g., passing percents) on students, schools and districts. An adequate growth measure can provide information on how students, and hence schools and districts, have progressed throughout the year. Many growth models exist in the literature, from simple models focusing only on percentage change over years to complex models such as value added models that take into account various effects on students' achievement and growth. In this study, a few of these growth models were empirically compared in an attempt to inform researchers and practioners about relative strengths and weaknesses of the models.

The models investigated include the Status Model (SM), the Hybrid Success Model (HSM) (McCall, Kingsbury & Olson, 2004), the Ordinary Least Squares regression model (OLS), the Deviation from Passing model (DFP), and the Hierarchical Linear Model (HLM). Simulated data were used in this study, where the true growth is assumed known a priori. The research question was to investigate whether the various growth models were able to recover the true ranking of schools.

Methodology

Data Simulation

Simulated data were used in this study, where the parameters were known a priori. The models were compared in terms of their recovery of the parameter values. Students' test scores were generated for four consecutive years. A separate but similar linear model was used for each year. The linear model incorporated the following variables: gender, ethnicity, geographical region, SES, and teaching effect. These variables are often taken into account when evaluating school accountability. The following equation was used to simulate data:

$$Y_{ij} = Y'_{ij} + V_1 + V_2 + V_3 + V_4 + V_{5j} + \varepsilon_{ij}.$$

Y_{ij} refers to the test score for student i within school j for a given year; Y'_{ij} refers to the student's test score from the previous year; V_1 through V_4 are fixed effects, representing ethnicity, gender, social economic status and growth for a given year. V_{5j} represents the school effect and ε_{ij} refers to the random noise that goes into the test score for an individual.

Data were generated for 200 schools, each with 50 students. The demographic information in the data was generated based on the proportion of each category observed in one state's assessment data. With ethnicity, about 50% of the students were white, 40% of the students were African American and about 10% of the students were other. The ratio of males to females was about 1 to 1, and approximately 37% of the students in the data received free and reduced lunch. The above values were used as probabilities and

Bernoulli trials were generated to assign these demographic values to each of the students in the simulated data.

In Year One, because there were no scores from the previous year, random normal deviates were generated instead, with a mean of 300 and a standard deviation of 10. For each year, with ethnicity, three effects were used: if white, a score of 1 was added; if African American, a score of 0.2 was added; and if other, a score of 0.5 was added. For females, a score of .2 was added, and a score of .8 was added if not taking free or reduced lunch. The growth score was defined to be a score of 8 from Year One to Year Two, a score of 7 from Year Two to Year Three and a score of 6 from Year Three to Year Four, assuming some decelerated growth from lower to higher grades. School effects were random multivariate normal deviates with a mean of 1.2, a standard deviation of 0.4 and a correlation of 0.8 between years. The random deviates, ϵ_{ij} , were generated from multivariate normal distribution with mean of 2, standard deviation of 1 and correlation between any two years being 0.6. The relative magnitudes of these effects were based on previous research as well as observations of state assessment data (Sanders, Saxton, Schneider, Dearden, Wright, Paul & Horn, 1994; Sanders, Saxton, & Horn, 1997).

After the data were simulated, performance categories needed to be assigned. Four performance categories were defined: Level One, Level Two, Level Three and Level Four. Level One represented the lowest performance level, and Level Four represented the highest performance. Level Three was treated as the proficiency level. Table 1 reports the cutscores and related percentages for each year. The cut scores were set based on records of one state's assessment data.

Table 1. Cutscores and Percent in Level for the Four Performance Levels.

	Level 1		Level 2		Level 3		Level 4	
	Cut	Percent In Level	Cut	Percent In Level	Cut	Percent In Level	Cut	Percent In Level
Year One		3.1		12.9		53.3		30.7
Year Two	283.91	2.1	292.23	14.3	306.32	61.5	314.42	22.1
Year Three	292.53	6.6	302.80	15.2	314.08	53.9	322.59	24.3
Year Four	307.66	7.1	320.93	16.1	329.99	53.9	339.34	22.9

When measuring students' growth in consecutive years, one of the challenging aspects is to track students' records over years. Due to various reasons, almost always longitudinal data analysis has to accommodate missing data. After comparing the growth models in their recovery of school ranks, these growth models were also compared to observe their robustness to missing data. Three missing datasets were created, one with 20% of the data missing, one with 50% of the data missing and one with 80% of the data missing. The SES variable was used to help create missing data. In this process, the students receiving reduced or free lunch had three times the probability of leaving the school than students not receiving reduced or free lunch. Random data points then were deleted from the data, where students with low SES variable values had three times the probability of being deleted.

After the missing data were created, the various growth models were applied both to the full data matrix and to the missing data matrix to observe robustness of the models.

Models

Status Model (SM)

The SM calculates the difference in percent of students in a given performance category in consecutive years. This model directly takes into account the state standards and is often used in state accountability workbooks to define AYP. Many states use this index. States can choose to include confidence intervals around estimates to account for measurement error. This index is straightforward and easy to calculate. No distributional assumptions are involved in this model.

Because typically stake holders are concerned with whether students have reached proficiency or not, therefore, in this study, the index is defined to be the difference between the percentages of students meeting the standards in consecutive years. For each school, the percentages of students meeting the standards were calculated for each year. Differences were taken between consecutive years. Next, the differences were added together and defined to be the SM index for the school. The overall school ranking was calculated based on this SM index.

Deviation From Passing (DFP) Model

The DFP uses a fixed score as a reference point to define growth. The metric is similar to the difference between two z scores, with the fixed reference point replacing the means. Usually the reference point is defined as the proficiency level. Growth indicates change in performance relative to the proficiency level. This model is simple to implement and explain. It addresses the primary concern—students' progress towards the goal of passing the test.

This model is defined in the following equation:

$$DFP = \frac{SS_{t+1} - Cut_{t+1}}{SD_{t+1}} - \frac{SS_t - Cut_t}{SD_t}$$

where SS_{t+1} and SS_t refer to the scale scores of a given student at two consecutive years and SD_{t+1} and SD_t their respective standard deviations; Cut_{t+1} and Cut_t are the cutscores for the proficiency level for the two years. This index defines how far in deviation units a student is from the proficiency level and compares that deviation from the year after to examine growth. The DFP index was computed for each student for each pair of consecutive years, assuming that the cut scores carry over the same meaning across different years. This assumption can be hard to meet. Means were taken across the students within each school and compounded to create the index for the school. The rank ordering of the schools were obtained based on DFP and compared with those true school effects used in the simulation.

Hybrid Success Model (HSM)

The HSM combines students' performance category and growth. For each student, a growth target is defined so that the student will reach or surpass the proficiency level in a set period of time. A school's success is calculated as the extent to which students in the school reach their growth targets. One advantage of this model is that students are expected to reach proficiency and the growth target is defined at the least as one that requires the student to reach proficiency in one or more years.

McCall, M., Kingsbury, G. & Olson, A (2004) provided a detailed description on how HSM index can be calculated. See that paper for the exact procedure adopted to calculate the value for this model. The basic idea was to define a growth target for each

student based on beginning performance level, average growth in that performance level and difference from the proficiency level. For example, the growth target for a given student who was at the basic performance level in the previous year can be defined as the amount of growth the student needs to demonstrate in the remaining years till 1012 to reach proficiency level. The student's actual growth score, which is available from the longitudinal data as the actual gain score, is then divided by the growth target. After some adjustment to the product, a student's HSM index can be obtained. In this study, the students' HSM values were aggregated within schools. Schools then were ranked ordered based on the magnitude of their HSM value and compared with the true school ranking.

Ordinary Least Squares (OLS) Regression Model

The OLS utilizes a regression approach to compare the actual score for Year 2 and the predicted score for Year 2. If the student's actual score is greater than or equal to the expected score, the student is considered to have grown. School and district growth is calculated by aggregating students' actual minus predicted scores. This approach allows the inclusion of other related variables, such as demographic variables. Because of the peculiarities of regression-based methods (regression to the mean, linearity assumptions, etc.), the growth measure will need to be interpreted taking into account these limitations. The general equation for OLS model is as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots .$$

In this study, OLS predicted scores were calculated for each student. The differences between the predicted and the actual value from the longitudinal data were obtained. These values then were aggregated over schools so that each school was

associated with an OLS index. Rank ordering of the schools thus could be obtained and compared with the parameters.

Hierarchical Linear Model (HLM)

The HLM, similar to OLS, is a linear approach. In addition, it allows more complex investigation of various factors contributing to students' change score from one year to the next. It accounts for the hierarchical structure of assessment, students, schools and districts. This model also provides unbiased parameter estimates when data are missing at random or missing completely at random (Bryk & Raudenbush, 1992). Due to its complexity, however, the model specification can be problematic and is potential to numerous sources of bias. Compared with the other models described above, HLM is relatively more complex and harder to interpret. Little evidence has been presented to demonstrate that the benefit of these models outweighs the costs of implementing them.

In this study, a three-level HLM model was specified to fit the data: level 1 was the time level, level 2 was the student level, and level 3 was the school level. The model specifications are listed below:

$$\text{Level 1: } Y_{ij} = \pi_{0ij} + \pi_{1ij}(\text{Time}) + e_{ij};$$

$$\text{Level 2: } \pi_{0ij} = \beta_{00j} + \beta_{01j}(\text{BLACK}) + \beta_{02j}(\text{HISP}) + \beta_{03j}(\text{GEN}) + \beta_{04j}(\text{SES}) + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + r_{1ij};$$

$$\text{Level 3: } \beta_{00j} = \gamma_{000} + u_{00j}$$

$$\beta_{01j} = \gamma_{010} + u_{01j}$$

$$\beta_{02j} = \gamma_{020} + u_{02j}$$

$$\beta_{03j} = \gamma_{030} + u_{03j}$$

$$\beta_{04j} = \gamma_{040} + u_{04j}$$

$$\beta_{10j} = \gamma_{100} + u_{10j}.$$

At level 1, students' test score is modeled to be a linear function of initial achievement as the intercept and of growth trajectory as the slope. At level 2, the person level, initial achievement is a linear function of person-level demographic variables and error; the growth trajectory is an unconditional function of school effect and error. The level 3 model represents the variability among schools. The student level growth parameter, based on how this model was defined in this paper, is the π_{1ij} parameter from level 1. But to compare schools in terms of their effectiveness, our focus was on the β_{10j} parameter estimate for each school, the individual effect each school has on students' achievement. For each school, a separate β_{10j} school effect was estimated. Schools were rank ordered by their β_{10j} parameter estimates. The ordering was compared with the parameter values.

Results

Table 2 reports the mean, standard deviation, minimum and maximum values for each of the four years' simulated data. The variability of the scores increased across years, as can be expected from the way the data were simulated.

Table 2. Statistics for the Simulated Data.

	Mean	Standard Deviation	Minimum	Maximum
Year One	302.57	10.00	264.02	339.26
Year Two	313.04	10.25	272.95	353.18
Year Three	322.52	10.67	279.96	365.20
Year Four	330.98	11.21	287.36	377.86

As described in the methodology section, indices based on SM, DFP, HSM, OLS and HLM were calculated for each for the four datasets: full data, data with 20% missing, data with 50% missing, and data with 80% missing. The 200 schools were ranked ordered based on each of the growth models for each dataset. Correlations were calculated between the rank order of the schools in their parameter school effect and the rank order of the growth model results. Tables 3 to 6 report these correlation matrices.

In Table 3, it can be observed that HLM yielded the highest correlation with the true school ranking, followed by DFP, OLS, HSM and finally by SM. HLM, DFP, HSM and OLS all yielded reasonably high correlations, whereas SM yielded a relatively low correlation coefficient. Among the better performing growth models, DFP and HSM are the most straightforward models. All the information required to calculate the growth index was students' scale scores for all four year and cutscores for the proficiency levels. Both models directly reference the performance standards. HLM is the relatively more complicated model and involves assumptions on normality, variability of error terms, etc. Strong correlations can be observed among DFP, OLS and HLM.

Table 3. Correlation Matrix for Full Data.

	TRUE	SM	DFP	HSM	OLS	HLM
TRUE	1.000	0.596	0.840	0.794	0.813	0.862
SM		1.000	0.673	0.559	0.678	0.686
DFP			1.000	0.885	0.984	0.969
HSM				1.000	0.900	0.938
OLS					1.000	0.973
HLM						1.000

In Table 4, when there were 20% of the data missing, the correlations for all the models decreased, but only by 0.02-0.04. SM again was the weakest model, correlating only about 0.57 with the true rank ordering of the schools.

Table 4. Correlation Matrix for Data with 20% Missing.

	TRUE	SM	DFP	HSM	OLS	HLM
TRUE	1.000	0.572	0.804	0.751	0.774	0.858
SM		1.000	0.621	0.506	0.629	0.658
DFP			1.000	0.860	0.979	0.929
HSM				1.000	0.881	0.889
OLS					1.000	0.927
HLM						1.000

Table 5 reports correlation matrix when there was 50% of the data missing. As expected, all the correlations dropped. HLM again produced the highest correlation, with a drop of only 0.01 on correlation coefficient metric. SM again produced the lowest correlation, accounting for less than 25% of the variance. DFP, HSM and OLS still

correlate reasonably high with the true rank ordering of schools, especially with 50% of the data missing.

Table 5. Correlation Matrix for Data with 50% Missing.

	TRUE	SM	DFP	HSM	OLS	HLM
TRUE	1.000	0.469	0.740	0.723	0.732	0.844
SM		1.000	0.597	0.414	0.602	0.553
DFP			1.000	0.796	0.981	0.875
HSM				1.000	0.831	0.843
OLS					1.000	0.881
HLM						1.000

When 80% of the data are missing, as shown in Table 6, the correlation coefficients dropped. SM still had the lowest correlation, with a value of only 0.25; DFP, HSM and OLS had similar correlations, in the neighborhood of 0.50; and HLM again produced the highest correlation, a 0.79 value, not too much lower compared with the correlation obtained with the full data matrix (.86).

Table 6. Correlation Matrix for Data with 80% Missing.

	TRUE	SM	DFP	HSM	OLS	HLM
TRUE	1.000	0.253	0.503	0.546	0.465	0.792
SM		1.000	0.478	0.285	0.520	0.329
DFP			1.000	0.713	0.967	0.617
HSM				1.000	0.747	0.670
OLS					1.000	0.620
HLM						1.000

Across the results from all the data, it appears that OLS and DFP were quite similar in terms of ranking schools in their effectiveness. Comparing the two models, DFP is relatively more straightforward, only taking into account the deviation score of a student from the proficiency level and does not require demographic variables. OLS, on the other hand, needed information on demographic variables as part of the independent variables.

To demonstrate how well HLM performed and how poorly the SM model performed, Figures 1 through 4 were constructed to demonstrate the true values contrasted with the estimated values. Figures 1 and 2 were constructed to contrast the scatter plot of SM versus true ranks for both full data matrix and data with 80% missing data. The horizontal axis represents the true rank of a given school in terms of its school effect; the vertical axis lists the rank of a given school using SM index. In Figure 1, even with the full four years of students' data, the relationship between the true ranks and the SM ranks was fairly weak. Extreme cases can be observed, such as very low true ranks but very high ranks according to SM index, and vice versa. In Figure 2, with 80%

missing data, SM failed to pick up the differences among schools. The rank order of schools based on SM appeared to be almost independent of the schools' true ranks. This index does not seem to work effectively when substantial amounts of data are missing.

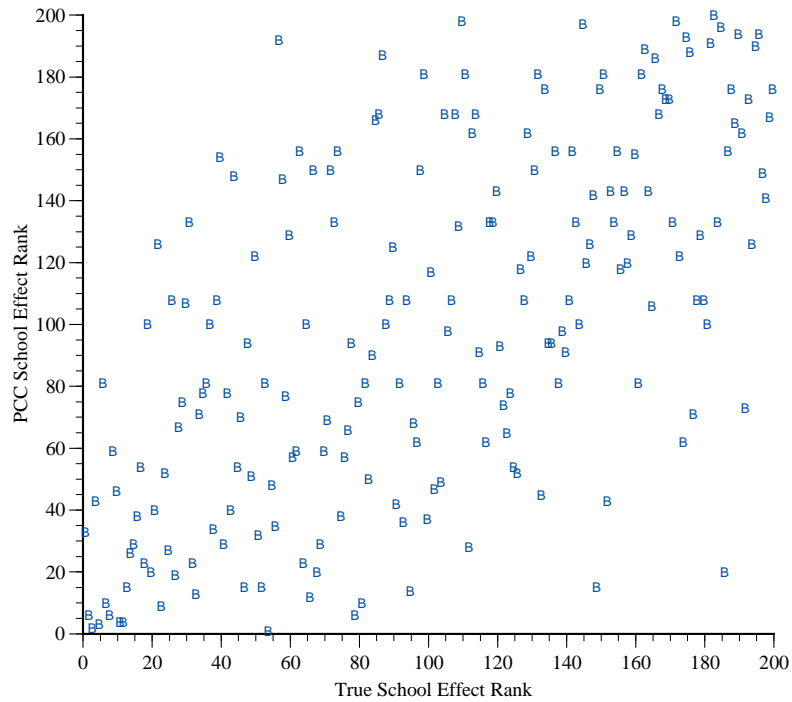


Figure 1. Contrast of Rank Order of Schools, SM Model, Full Data.

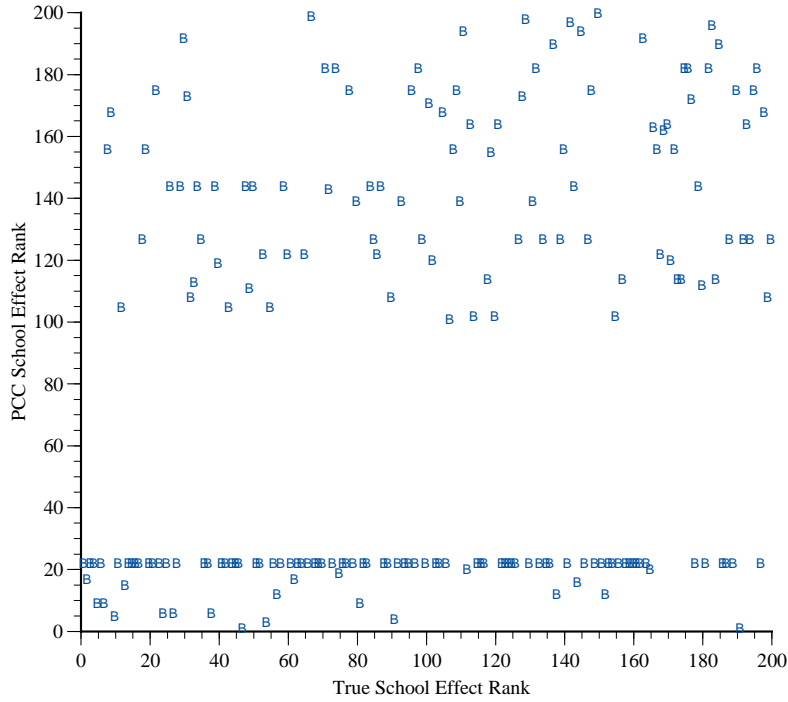


Figure 2. Contrast of Rank Order of Schools, SM Model, 80% Missing Data.

Figures 3 and 4 were constructed to contrast the scatter plots of HLM versus true school ranks for both full data matrix and data with 80% missing data. Compared with the scatter plots based on SM, the relationship between true and estimated ranks was much stronger. Some extreme cases still existed, but not as extreme or as many as the results observed with the SM model. Even with 80% of the data missing, HLM still appeared to perform well at preserve school ranking. However, as can be observed from the figures, the ranks at two extremes were better preserved than the ranks in the middle. It appears that the schools on the two ends were better discriminated than the schools in the middle.

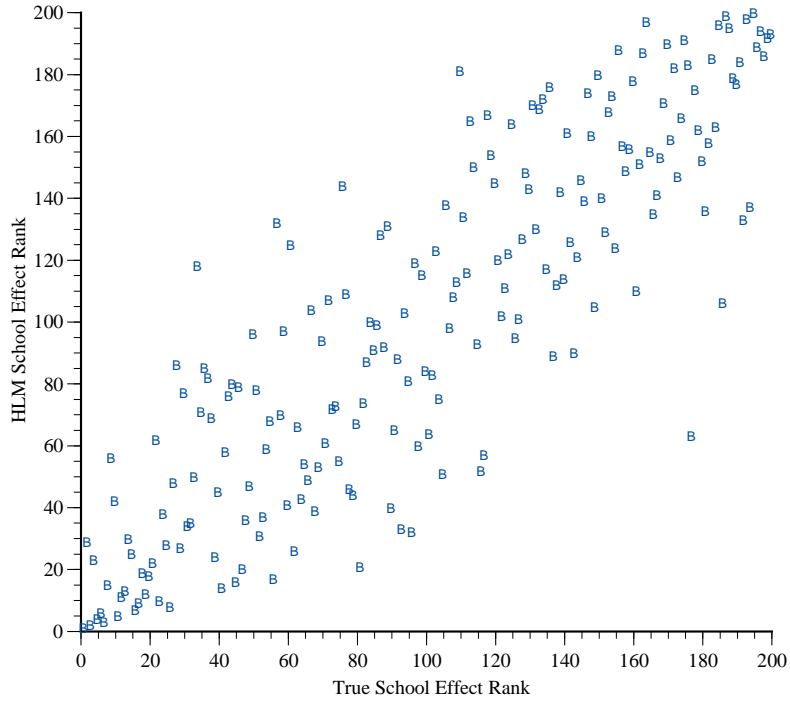


Figure 3. Contrast of Rank Order of Schools, HLM Model, Full Data.

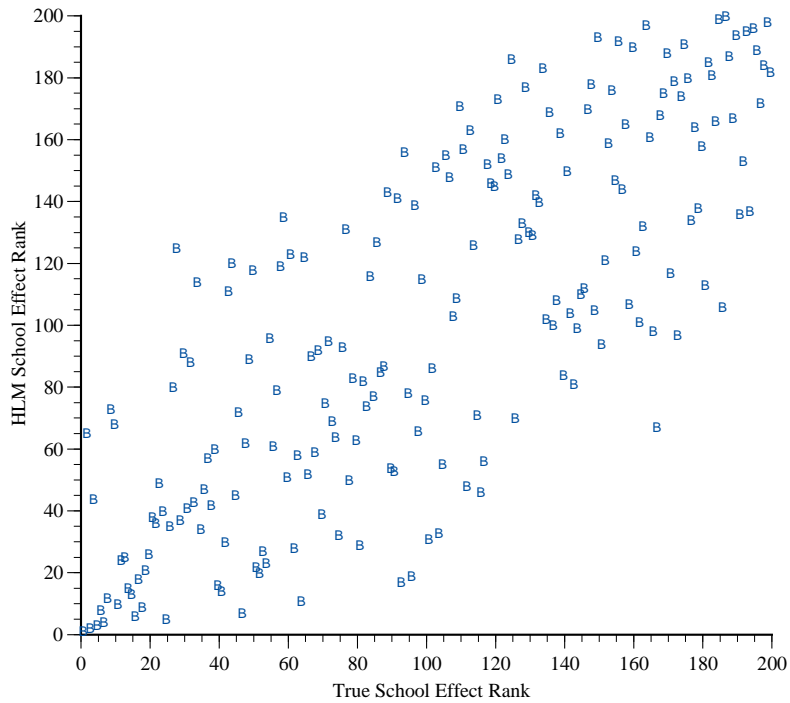


Figure 4. Contrast of Rank Order of Schools, HLM Model, 80% Missing Data.

Conclusion and Discussion

Compared with the random error in the simulation model, school effect was relatively small. Even so, most of the growth models investigated were able to preserve the rank order of schools reasonably well. Based on results using both full and missing data, HLM appeared to have performed the best, followed by the three models DFP, HSM and OLS. The SM appeared to have performed the worst. The simulation approach used in this study meets most of the major assumptions associated with each model.

The SM model appeared to be the worst-performing model according to the results from this study. Even with the full data matrix, the correlation between SM results and parameters was only around 0.60 and a fair number of school ranks were substantially different from their parameter ranks. The disadvantages of this model include that it does not accommodate growth occurring away from the proficiency line and it defines growth as increased percentages of different cohorts of students. Students that start quite below the proficiency level and end with barely below proficiency will be treated the same as students that start and end with barely below proficiency. As the results indicated from the study, this model may not be optimal for states to model students' growth.

DFP performed reasonably well with both full data matrix and data with missing records (except for data with 80% missing). This model does not require a vertical scale, takes into account the performance standards of the state, and addresses the fundamental interest—students are expected to pass the standards. On the other hand, with the DFP model, if standards are not of similar difficulty, students may be expected to grow differently from one year to the next. Measurement error is not accounted for in this

model. This approach, or a modification of this approach, might be considered as states are choosing an appropriate growth model.

HSM performed reasonably well with this simulation study. It produced similar correlations with DFP and OLS models. However, in the case of the extreme missing data (80%), the correlation yielded by HSM dropped significantly less than DFP and OLS. HSM is easy to implement, takes into account state standards and performance category, and is defined directly related to passing the standards. Measurement error is not accounted for in this model. Overall, the HSM showed some promise from the results of this paper. This approach, or some modified approach, might be considered as states are choosing an appropriate growth model.

OLS uses regression technique to compare students' predicted and actual growth. The results in this study showed that OLS produces relatively stable school-level results. Only when a large amount of data missing does this model fail to perform well. Growth targets are set using the empirical information from the designated base year on all students, instead of a selection of students. In this study, the OLS model assumed linearity in growth; however, nonlinear terms could be added. This model does not reference the standards; therefore students might pass and not show growth or students might show large growth and not pass. One major problem with this model is innate—regression towards the mean. As a result, high performing students are typically expected to have lower scale scores the next year, whereas low performing students are typically expected to have higher scale scores the next year. A high correlation between OLS and DFP was also observed from all data results. Between the two models, OLS is less straightforward and harder to implement.

HLM accounts for nesting effects, accommodates for missing data and can directly estimate school effects using a linear model. When the model is correctly identified, the results in this study showed that HLM performed with precision even when there was large amount of data missing, even data not missing at random. On the other hand, HLM is a quite complex model that requires stringent assumptions, the specification of the model can be challenging, the cost tends to be high, and the results can be hard to interpret. In addition, a vertical scale is typically required for the implementation of this model. Whether its benefits outweigh its costs, the states have to decide on their own when they are considering HLM as an option for modeling student growth.

Reference

- McCall, M., Kingsbury, G. & Olson, A. (2004). *Individual growth and school success*. Portland, OR: Northwest Evaluation Association.
- O'Malley, K. J., Vansickle, T., Housson, S., & Meyers, J. (2005). *Measuring Growth: Where Policy Makers and Psychometricians Meet*. Paper Presented at the CCSSO Conference, San Antonio, TX.
- Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee Value-Added Assessment System (TVAAS): A quantitative, outcomes-based approach to educational assessment. In Millman, J. (ed.), *Grading Teachers, Grading Schools*, Thousand Oaks, CA: Corwin Press.
- Sanders, W.L., Saxton, A.M., Schneider, J.F, Dearden, B.L., Wright, S. Paul, & Horn, S.P. (1994). Effects of building change on indicators of student academic growth. *Evaluation Perspectives*, 4(1) pp 3 and 7.
- Wright, S.P., Horn, S.P., & Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11(1), 57-67.