February 2012

Growth Model Comparison Study:
Practical Implications of Alternative
Models for Evaluating School Performance

Authored by Pete Goldschmidt, Kilchan Choi, & J.P. Beaudoin
for
Technical Issues in Large Scale Assessment (TILSA)
State Collaboratives on Assessments and Student Standards (SCASS)

CCSSO
Council of Chief State School Officers

Growth Model Comparison Study:
Practical Implications of Alternative Models for Evaluating School Performance

A paper commissioned by the
Technical Issues in Large-Scale Assessment
State Collaborative on Assessment and Student Standards
Council of Chief State School Officers

Authored By:
Pete Goldschmidt
Kilchan Choi
J.P. Beaudoin

Supported by:
Charlene Tucker, TILSA Advisor
Duncan MacQuarrie, Associate TILSA Advisor
Doug Rindone, Associate TILSA Advisor
Bill Auty, Education Measurement Consulting

This report was prepared for the Technical Issues in Large Scale Assessment (TILSA), a member of the system of State Collaboratives on Assessment and Student Standards (SCASS) supported by the Council of Chief State School Officers (CCSSO). The views expressed herein do not necessarily represent the positions or policies of CCSSO, its Board, nor any of its individual members. No official endorsement by CCSSO, its Board, nor any of its individual members is intended or should be inferred.

Table of Contents

Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance

Pete Goldschmidt, Kilchan Choi, and J.P. Beaudoin

INTRODUCTION

The Elementary and Secondary Education Act (ESEA) has had several tangible effects on education and the monitoring of education. There have been both intended and unintended consequences. ESEA's newer generation of federal programs, such as Race to the Top, and the recent ESEA flexibility guidelines, have continued to push development of methods to accurately and fairly monitor school (and more recently teacher) performance. One tangible result of the 2001 reauthorization of ESEA, titled No Child Left Behind (NCLB, 2002), is that there is considerable agreement that monitoring schools based on unconditional mean school performance or the percentage of students proficient does not hold schools accountable for processes for which they ought to be held accountable and tends to place diverse schools at a disadvantage (Novak & Fuller, 2003). Static average student performance measures are poor indicators of school performance and tend to reflect the input characteristics (i.e., student enrollment characteristics) of schools as much as they do actual school quality (Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank, & Williams, 2005; Choi, Goldschmidt, & Yamashiro, 2005; Meyer, 1996; Goldstein & Spiegelhalter, 1996) and capture factors outside of school control more than actual processes facilitated by schools (Hanushek & Raymond, 2003; Baker, Goldschmidt, Martinez, & Swigert, 2002; Meyer, 1996). This has prompted many to pursue incorporating growth models into accountability systems. There may be some debate as to what constitutes the optimal psychometric characteristics for assessments to be used in systems desiring to use growth models (Briggs & Weeks, 2009; Yen, 1986; Goldschmidt, Choi, Martinez, & Novack, 2010), but states are unlikely to step outside of their assessment development cycle for the sole purpose of basing accountability on student growth.

The purpose of this study is to compare several different growth models and examine empirical characteristics of each. This study differs from previous research comparing various models for accountability purposes in that the focus is broader — it is based on large scale assessment results from four states (Delaware, Hawaii, North Carolina, and Wisconsin) across two cohorts of students (each with three consecutive years of assessment results), and explicitly considers model results with respect to elementary and middle schools. Previous research has addressed statistical issues, and compared the effects of model specification (particularly with respect to student background characteristics) in some detail, and was often based on limited or simulated data (Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, & Resnick, 2004; Ballou, Sanders, & Wright, 2004; McCaffrey, Sass, Lockwood, & Mihaly, 2009; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Wright, 2010; ; Lockwood & McCaffrey, 2007; Wright, 2008).

These previous studies have provided significant guidance for the model selection and specifications we considered and examined in this study. We note that simulated data is optimal for identifying and testing specific aspects of models, but state standardized assessment results provide an opportunity for authentic evaluation of model performance and the variability that states may face in adopting a particular model. This applied study is well suited to address empirically whether certain models are more likely than others to provide accurate, fair, unbiased, precise, and consistent results[1]. This study addresses the following research questions regarding the performance of the different growth models:

1. Overall, does the model matter?
2. Do different models lead to different inferences about schools?
3. How accurately do models classify schools into performance categories?
4. Are models consistent in classifying schools from one year to the next?
5. How are models influenced by school intake characteristics (percent ELL, FRL, etc.)?
6. Do models perform similarly for elementary and middle schools?
7. Do models behave similarly across states?

We also note five additional considerations related to this analysis. One, we are not advocating using a single measure to evaluate school performance. Two, selection of which growth model to use needs to be driven by purpose and careful consideration of the

[1] We cannot strictly identify bias as these results are not based on simulation studies where a “true” model result can be compared to alternatives. Given we use existing data from multiple states, we examine deviations from results that appear to be consistent across models and/or context to determine inconsistencies, which could be loosely considered as bias.

underlying inferences that are desired (and are valid) from the results, i.e., driven by a theory of action (Marion, 2010). Three, one must recognize the extent to which transparency and complexity involve reasonable tradeoffs. Four, we make no attempt to rescale, improve the scale, or account for potential unaccounted for equating error. And five, the growth models that we compare are evaluated under the assumption that the model results are the basis for monitoring schools — that is, we do not examine the marginal impact of models under the Growth Model Pilot principles, nor any other current requirement related to NCLB. We reiterate that this is not a study of how these models might perform under the current ESEA guidelines, rather how these models might perform under an unconditional system. The role of growth models within an accountability system is not the focus of this analysis, rather how the models compare. How growth ought to be valued or incorporated into an accountability system is important and is discussed in *A Practitioner's Guide to Growth Models* (Ho and Castellano, in press) and is not the focus of this study. We do, however, provide some considerations that relate to incorporating growth model results into a broader accountability system.

While some states such as Tennessee (Sanders, Saxton, & Horn, 1997) and districts such as Chicago (Bryk, Deabster, Easton, Luppescu, & Thum, 1998) have used longitudinal models for accountability purposes for some time, interest in accountability models based on growth has increased nationally as a result of NCLB legislation and research demonstrating that cross-sectional accountability models provide weak indicators of school performance (Choi, Goldschmidt, & Yamashiro, 2005). The combined timing of states adopting the common core state standards, the assessment consortia[2] moving forward, the likely reauthorization of ESEA, and the ESEA flexibility program has resulted in further considerations of which model might be most appropriate to monitor school performance. The underlying question is whether there exists an optimal model that can consistently identify schools as performing well, or performing poorly. While the literature suggests that multiple measures ought to be considered, policymakers are interested in how well a single model can use student assessment results to hold schools accountable for facilitating

[2] There are two major consortia developing assessment systems for use in 2014-2015 that will assess the common core state standards in English language arts and mathematics, the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium. See http://parcconline.org/ for PARCC and http://www.k12.wa.us/smarter/ for Smarter Balanced.

learning. Specifically, there is increased interest in using growth models for school accountability. This study is unprecedented in that it presents results from a comparison of several models using data from four states and focuses on the practical implications of model choice.

THEORETICAL FRAMEWORK

Relevant criteria that must be established are the intended use and audience of the accountability model. We assume that the primary impetus for using growth models is to correctly identify the spectrum of school effectiveness in order to accurately monitor schools for rewards or sanctions, to ensure that all students have equal opportunities to succeed[3], and to ensure that schools are not disadvantaged by factors beyond their control. This focus is somewhat different from a system that provides information on which schools attain the highest achievement levels without consideration of potential confounding factors – which is what a parent is generally most interested in (Willms & Raudenbush, 1989). Hence, our concern is to first provide policymakers with results that afford valid inference related to school performance[4]. In practice, this relates to considering how status and growth combined might be used to form a comprehensive picture of school performance and provide desired motivations for performance — that is, depending on the values of the policymakers. Hence, considering what achievement status, or change in status, implies and how it relates to effectiveness and growth must be considered within both a value and technical framework to develop an accountability system.

Previous research has demonstrated that simply using school means (Aitkin & Longford, 1986) or year-to-year analyses of school performance (Linn & Haug, 2002) cause invalid generalizations about school quality. Moving away from school means and cross-sectional analyses toward longitudinal panel models[5] is the first step in generating potentially

[3] Growth models can, of course, serve other purposes. For example, they can monitor growth in specific student outcomes such as vocabulary, provide predictions of future performance, and evaluate programs based on student progress.

[4] Ultimately this information is suitable for parents and other stakeholders as well, but may require some communication as schools that are known to be “good” schools are often good only because they benefit from the enrollment characteristics of their students, and not due to school processes.

[5] Longitudinal models encompass several different sets of analyses. However, germane to school accountability are two varieties: panel models focusing on individual student achievement over time or school improvement, focusing on changes in performance of cohorts over time (e.g., 3rd graders in 2010 vs 3rd graders in 2011).

valid results. However, examining year-to-year changes in student achievement may not resolve problems with potential confounding factors (Campbell & Stanley, 1963) and present additional methodological problems, such as spurious negative correlations between initial status and achievement growth (Rogosa, 1995) at the individual student level. Further, the instability of year-to-year results at the school level (Linn & Haug, 2002) is also problematic in disentangling measurement error, changes in student composition, and school effects.

While the literature suggests that different models can provide very similar results, there is little systematic study of how models compare, not only against each other, but also in different contexts and over time. Models differ in purpose, approach, and assumptions. One basic difference is whether one assumes a fixed or random effects model. There is both a philosophical and a statistical rationale for each approach. Philosophically, if policymakers assume that schools are a random subset of a universe of schools, then a random effects model would be appropriate; however, if policymakers assume that the set of schools about which they intend to make inference represents all the schools of the relevant population, a fixed effects model would be appropriate[6]. This debate is similar to (and related to) the debate about including confidence intervals around school effects (whether based on status or growth[7]). Both fixed and random effects models produce school estimates with standard errors and could generate confidence intervals.

An important aspect that underlies using growth models is policymakers' conception of growth. Achievement growth can be considered in terms of greater depth and/or breadth of knowledge and ideally standards and assessments would be designed to explicitly consider student learning progression within a content area (Herman, Heritage, & Goldschmidt, 2011). If growth is defined as more mastery of skills, then this implies a model that uses gains or growth over time, whereas if growth is defined "next content topic," then this implies a covariate adjustment, or ANCOVA model.

Another key element for considering the use and interpretation of results based on growth models is that the outcome must have constant meaning over time (Raudenbush, 2001). Hence, the scale is important in drawing conclusions from individual growth curves

[6] It is likely that policymakers do not consider either of these assumptions, but rather a model is developed that carries assumptions that may be philosophically opposed to policymakers' views.

[7] This is not a focus of this paper, but the use of confidence intervals for growth is more nuanced than the use of confidence intervals around status. Confidence intervals should be considered when using growth models, especially projection models (Goldschmidt & Choi, 2007).

(Yen, 1986). Theoretically, the optimal metric to use when examining change is a vertically equated Item Response Theory (IRT)-based scale score that is on an interval scale and is comparable across grades (Hambleton & Swaminathan, 1987). Such scores represent content mastery on a continuum and may be used to measure absolute academic progress over time, and would be considered the metric of choice for an accountability model based on growth. However, different scaling methods affect results (Briggs & Weeks, 2011), and there is some concern that vertical equating using IRT does not guarantee an equal interval scale (Ballou, 2009). Also, equating is generally designed to compare contiguous grade pairs (Yen, 1986), and scales may be less meaningful as the grade span increases. Prior results indicate that inferences based on model results depend upon the metric used (Seltzer, Frank, & Bryk, 1994). Assessment and scaling procedures can impact inferences at the teacher and school level as well (Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007; Briggs & Weeks, 2011). At the school level, scaling techniques result in correlations of school effects ranging from .53 to .99 (Briggs & Weeks, 2011) but appear to be generally high. The percentage of schools identified in performance categories is unaffected (Briggs & Weeks, 2011), but the impact of scale on different school types is not known. Previous research also indicates that the metric may be less important for relative decisions and inferences about schools based on growth models, and more robust to missing student information, but are influenced by whether the assessment is a norm referenced test (NRT) or a standards based test (SBT) Goldschmidt, Choi, Martinez, & Novack, 2010).

DATA

The data for these analyses were compiled from four participating states: Delaware, Hawaii, North Carolina, and Wisconsin. Two of the states were relatively large (more than 500,000 students), and two fairly small (less than 200,000 students). The analyses use public elementary and middle school data. The dataset includes four years of data, and we focus the analyses on two elementary and two middle school cohorts (see Figure 1). In this way we based results for schools on students who potentially attended three years (although the impact of missing student data is addressed in the analysis). The total number of schools[8] in the analysis is 2,645[9]. The schools per state are identified in Table 1. It should be noted that the total number of schools in each state in the analysis may exceed the total number of schools that a state has in any given year because the total number includes all schools that may have existed in any one year (2007 through 2010).

	Grade					
Year	3	4	5	6	7	8
2007	C1*			C3		
2008	C2	C1		C4	C3	
2009		C2	C1		C4	C3
2010			C2			C4

* C1 refers to the 2009 elementary school cohort; C2 refers to the 2010 elementary school cohort; C3 refers to the 2009 middle school cohort; and C4 refers to the 2010 middle school cohort.

Figure 1: Data Structure

[8] We use the term schools to mean elementary and middle schools, or schools that contain grades 5 and 8. Students may change schools over the three year span; the student's current performance is attributed to school of record in the last year of a particular analysis (e.g., if the model examines gains from 2007 to 2008, the 2008 school is the school of record).

[9] One state (S3) was not able to provide middle school results for cohort 3. Also, we were unable to obtain SEMs (standard errors of measurement) from state S4, precluding us from estimating the true score gain model for that state. We used the last of the banked (students were allowed to retest up to three times for AYP purposes) set of scores from state S4 (allowing us to examine whether this policy relates to model performance).

Table 1:
Number of Schools in the Sample per State

State[*]	Frequency	Valid Percent
S1	279	10.5
S2	177	6.7
S3	252	9.5
S4	1,937	73.2
Total	2,645	100.0

[*] States are randomly assigned a number

In order to create some bounds on the extent of the analysis, we focus on a subset of student characteristics that have been shown to be related to school outcomes in prior research (Choi et al., 2005) and are generally the subject of debate in terms of whether school systems provide egalitarian results for all students, and whether these characteristics ought to be included in accountability models (Ballou, Sanders, & Wright, 2004). We restrict the analysis to minority status; ED (Economically Disadvantaged); ELL (English Language Learner); SWD (Students With Disabilities); and mobility. We do not control for variability in each state's classification process. We also consider school type (elementary and middle) and school size[10].

[10] As noted, the focus of this evaluation is on between-model variation in results. Preliminary analyses indicated that including or excluding student background at the individual or the school aggregate level generally did not substantively change school results (as much as between model results).

Table 2a:

Descriptive Statistics of School Demographics

	State S1			State S2			State S3			State S4		
	Mean*	N	S.D.	Mean	N	S.D.	Mean	N	S.D.	Mean	N	S.D.
SWD	18.92	266	13.00	23.12	152	25.91	12.14	252	8.91	9.24	1854	8.52
ED	60.90	266	26.09	54.50	152	24.71	52.49	252	23.50	51.92	1854	24.86
ELL	8.99	266	12.79	3.55	152	7.12	8.96	252	10.63	9.32	1854	11.32
Minority	55.54	266	32.86	54.48	152	25.28	85.09	252	15.20	45.21	1854	30.11
C09same_schl 07_09[2]	64.06	270	25.75	41.81	166	34.63	15.91	168	15.75	80.57	1825	26.51
C09same_schl 08_09	81.30	270	21.44	58.26	166	37.28	73.32	252	24.96	89.97	1825	20.26
C09same_schl 07_08	73.46	270	22.58	63.78	166	27.91	64.80	168	24.72	88.72	1825	20.04
c10same_schl 08_10	61.34	266	24.50	47.01	152	32.99	67.65	252	26.36	79.73	1854	26.86
c10same_schl 09_10	75.54	266	23.53	66.30	152	32.25	88.62	252	12.47	89.25	1854	20.05
c10same_schl 08_09	70.38	266	20.67	57.44	152	29.56	73.32	252	24.96	88.67	1854	20.04
School N	459.74	266	1395.54	123.12	152	102.15	96.92	252	97.57	89.23	1854	65.55

* Means are in percent except school N; and 2) C09same_schl 07_09 through c10same_schl 08_09 refer to student stability (e.g., the 64.06 value for state 1 for C09same_schl 07_09 from 2007 thru 2009 64.06 percent of students of the 09 cohort remained in the same school over the three year span).

All the mean values in Table 2a are school percents except school N which is the average school size in the state. We calculated mobility by examining whether students had the same school ID in each year. While this does not capture within year mobility, it does capture between year mobility (although it could be within year if it occurred before assessments were given). This measure of school mobility is also highly correlated with FAY (Full Academic Year), r = .92.

We also utilized the two step cluster analysis feature of PASW™ (SPSS, 2001) to generate clusters of similar schools across states in order to examine the impact school characteristics have on model performance and allowing for the comparison of like schools across states. This process resulted in schools being classified into one of five clusters. The clusters are based on the two years of student background information (as presented in Table 2b) as well as mobility and school size. A summary of the classifications is presented in Table 3. The clusters were generated using the entire dataset. The classifications we present here are not intended to align with any single state's definitions and are classified to compare like schools across states.

The distinguishing features of school classifications are as follows:

- Disadvantaged I generally consists of a high proportion of ED, ELL, and Minority students;
- Disadvantaged II also tends to have a high proportion of ED and Minority students, but have fewer ELL students;
- Large schools are larger than average, but tend to be fairly average with respect to ED, ELL, and Minority student enrollment;
- Mobile schools tend to have both above average mobility and also above average proportions of SWD; and,
- Advantaged schools tend to have a below average percent of both ED and ELL students.

Table 2b:
Percent of Schools Within Each Category

	State				
	S1	S2	S3	S4	Total
Disadvantaged I:	15%	3%	10%	12%	11%
Disadvantaged II:	31%	27%	41%	19%	22%
Large:	14%	13%	15%	4%	6%
Mobile\SWD:	7%	23%	14%	4%	6%
Advantaged:	33%	34%	20%	62%	54%
Total N:	259	143	168	1,792	2,362

Table 3 summarizes the characteristics of schools for each cluster.

Table 3:
Student Characteristics Defining Clusters*,**

Characteristic	Disadvantaged I: ED/ELL/Minority	Large	Disadvantaged II: ED/Minority	Mobile\SWD	Stable Advantaged
SWD_09	10.33	10.54	13.07	20.17	9.47
ED_09	74.51	37.70	73.25	54.81	35.40
LEP_09	28.94	4.80	5.52	4.28	4.67
same_schl08_09	89.83	87.17	86.78	11.02	94.77
same_schl07_08	86.92	27.88	82.66	67.81	92.65
School size 09	74.74	714.16	69.86	49.11	107.39
minority_10	78.58	41.11	79.63	62.16	30.49
SWD_10	10.24	11.01	13.16	23.41	9.58
ED_10	78.77	40.64	75.83	58.23	39.33
LEP_10	32.28	5.46	6.31	4.97	5.80
c10same_schl09_10	89.20	87.61	84.58	48.78	94.49
c10same_schl08_09	86.25	48.90	83.60	35.51	92.94

* Table values are percent, except school size which is number of students.

**Green cells indicate that they influence the description of the clusters.

Brief Summary of Each State’s Assessments

Because each state’s assessment is unique, it necessarily confounds results related to state context and assessment effects. As noted, scale and assessment type impact school effect estimates. We are able to address this to some extent by using similar schools across states, which reduces the impact of context, and allows for inferences related to assessments based on differences in estimated effects by model within school category (e.g., Disadvantaged I and Large) by state. States S1, S2, and S4 are on a vertical scale, while state S3 is not on a vertical scale. In general each state with a vertical scale applied a different procedure that results in different scales (e.g., the state S1 scale has a broad span, while the state S4 scale is very narrow). However, this alone cannot isolate differences in assessments as several other critical elements impact score meaning (e.g., standards adopted, test blueprints, and standard setting processes). Moreover, there is mounting evidence that suggests that instability in scores is related to design effects and equating error (Phillips, Doorey, Forgione, , & Monfils, 2011). The extent to which these factors contribute to instability in school effect estimates is also unique to each state. These analyses do not include information about equating error, which implies that some of the inconsistencies observed in model performance may be linked to these additional sources of error.

Data Quality

The data quality approach for this study was comprised of five components. Each component was sequentially dependent and required minor variations to address the four data sets (DE, HI, NC, and WI) used in the student data file structure. The components were to

- normalize the raw files using common data elements;
- establish longitudinal structures via the unique student identifier (USI), unique school identifier (USchID), and unique district identifier (UDI);
- develop stored structured query language (SQL) procedures that created subject-specific, multi-year tables;
- screen data ranges within each element; and,
- produce descriptive data for each state.

MODELS AND METHODS

There are limitless modeling options related to monitoring school performance and we by no means capture all of them. However, we used as guidance the recent Council of Chief State School Officers' (CCSSO) brochure on growth models and the typology presented therein (CCSSO, 2010). Hence, we included a gain model (including simple gain and fixed effects); a covariate adjustment model (fixed and random effects); a measurement model (random effects); a layered model (mixed effects); a quantile regression model (student growth percentile as described by Betebenner, 2009); and a growth to standard model. Preliminary analyses indicated that, in general, the within model effect of specifications that included or excluded student background was substantially less than the differences within model by school level (elementary and middle) or subject (math and language arts), and substantially less than differences between models. Some of the growth models we utilize in this study would also be considered value added models.

A common approach for value added models is to assume that current student achievement, y_t, is a function of previous achievement, y_{t-1}, and that by including prior achievement, much of the cumulative process of school is incorporated (Hanushek, 1986; Ballou, Sanders, & Wright, 2004). Current achievement is also assumed to be a function of individual and contextual elements.

However, there is general interest across different specifications. As noted, each model type results in a different inference about schools. For example a gain score model holds schools accountable based on student gains in performance, whereas a covariate adjustment model holds schools accountable based on where students are currently, accounting for where they were on a prior occasion. Value added models generally take the latter approach, but may include multiple prior test scores, which has the benefit of producing more precise estimates as well as reducing bias.

More complex models include the layered model (Sanders, Saxton, & Horn, 1997), a mixed effects model that layers gains over years using all available assessment data. Another complex approach uses quantile regression. This latter model focuses on normative changes in performance over time.

As noted, an important concern is whether models produce biased results. Several of the above models use the term *random*, *fixed*, or *mixed* to describe effects. A fixed effects model requires fewer assumptions, but is less efficient if random effects model assumptions are tenable. Empirically, results indicate more flexibility in practice than in theory (; Lockwood & McCaffrey, 2007; Ballou, Sanders, & Wright, 2004). Results also suggest that the bias due to specifying a random effects model is significantly reduced given specific conditions of the data such as multiple prior assessments, and greater variability within schools (Lockwood & McCaffrey, 2007) and specification decisions when prior assessment results contain measurement error (Wright, 2008). A fixed effects model generally takes the form:

$$\Delta_i = (\alpha + \theta_j) + e_{it}, \text{ where } e_{it} \sim IID(0,\sigma^2) \qquad (1)$$

whereas a random effects model takes the form:

$$\Delta_i = \alpha + (\theta_j + e_{it}), \text{ where } e_{it} \sim IID(0,\sigma^2) \qquad (2)$$

Here, $\Delta_i = Y_{it} - Y_{i(t-1)}$ represents the gain in assessment Y for student *i* from Y_{t-1} to Y_t. The school effects, θ_j, are part of the intercept in the fixed effects model and part of the error term in a random effects model (McCaffrey et al., 2004). The key marginal assumption that random effects models assume is that residual school effect is uncorrelated with individual error e_{it}. As noted, in practice, random effects models generally converge with fixed effects estimates (Raudenbush, 2004). Random-effects models infer school effects from the school-level residuals in the model, while fixed effects models introduce a specific term for each school and directly estimate the school effects through the coefficients associated with those terms. This is particularly the case when multiple prior assessments are used as predictors and the assessments contain measurement error (Wright, 2008). In a fixed effects model, a school's score, (θ_j), represents the difference between school *j*'s effect and a reference school or the grand mean in the sample. A benefit of a fixed effects model is that the specific elements of what might contribute to success need not be specified and is captured by the school indicator variables. We estimate a fixed effects gain model as part of the analysis. A disadvantage of the fixed school effects model is that time invariant variables cannot be included in the model and it is less precise if, in fact, random effect assumptions tenable.

The goal of this evaluation is to determine the potential latitude states might have in choosing a growth model for school accountability. We note again that the key determinant for selecting a model is its intended purpose and the extent to which results provide for valid inferences of the type desired by policymakers. That is, it is consistent within a theory of action that underlies the accountability system. Below, we present a brief description of the models and highlight the underlying differences among them.

We present findings on 10 models that are estimated separately for each state. Generating a typology of models depends on the audience. We use CCSSO's *Achievement Growth and Accountability* brochure (CCSSO, 2011) to provide a framework that allows us to coherently place models into context.

Categorical models use change in student performance category placement from year to year as the growth indicator.

Gain Score models are based on the difference between a student's earlier score and a later score. Gains can provide a simple estimate of change, but may have low reliability.

Regression models can provide the most precise measure of growth. The assumptions of the model must match the associated policies. Calculations are complex and a vertical scale is generally required.

Value-added models are a complex type of regression model that take into account student or school characteristics. *Added Value* is defined as producing more than typical or expected growth given specific characteristics of the student or school.

Normative models compare changes in student performance to that of a norm group in order to determine whether the change is typical or abnormally high or low. A vertical scale is not required. The model does not directly address whether the observed growth is adequate to reach a defined standard[11].

Although the focus of this analysis is on comparing growth models, we include status as a transparent comparison for growth model results. Aligning the accountability policy requires an understanding of what would constitute valid inferences about a school based on that school's score on the model. A status model assumes that a student's performance is solely a function of current school processes and is not impacted (confounded) by additional factors that contribute to a student's score (Goldschmidt, Choi, Boscardin, Yamashiro, Martinez, & Auty, 2006). Status based on the original test metric is preferred to using percent proficient as too much information is lost by categorizing scores – not to mention potential misclassification errors, and loss of information due to aggregation (Thum, 2003; Choi et al., 2005).

[11] All of these model types can form the basis for determining whether a student is likely to reach some specific future performance standard.

We consider each growth model in turn. Much of the information related to model properties was developed focusing on teacher effects, but, in general, the properties established generalize to estimating school effects as well[12]. Table 4 presents the relationship between the CCSSO growth model typology and the models we examined in this study. We expand on the CCSSO typology by presenting two ways to consider each model type: one is the model's intent, or inferential intent; and two, is the general way the model would be estimated[13]. For example, a status (one point in time) model tends to be categorical (i.e., it uses the percent of students in a specific category such as proficient and above). Another model might be a simple gain model that affords the inference of the change in performance from one assessment occasion to another and is estimated by calculating a gain score (i.e., the simple difference between the current score and the previous score) for each student. A true score gain model has the same intent of a simple gain model, but is estimated using a (mixed effects) regression model.

In the following section we describe each of the models and model specifications in greater detail.

[12] There are important differences in terms of inferences and what effects may mean – even based on the same models. Also, there are some potential simplifications and complications as well. For example, teacher effect models often use mixed effects across classified models, while a school model could conceivably have a strict nesting structure. Estimated school effects are necessarily comprised of the effects of teachers, administrators, and other educators and staff in aggregate, and we are not concerned with its make-up or within school distribution. Teacher effects however may be confounded to some extent by general school effects. This issue is not the focus of our study, but it is important to note that, like models, it will mean different things when used for school or teacher accountability.

[13] VAM models are often estimated using ordinary least squares regression, but are different in that the focus of a VAM model will be on the unique contribution of schools to student performance, estimated either with a fixed parameter estimate or a random effect.

Table 4. Match in Intent (I) or Estimation (E) between Models in Study and Growth Model Typology

		Growth Model Typology				
Model in Study		Categorical	Gain	Regression	Value Added	Normative
Simple Gain	GAIN		I E			
Fixed Effects Gain	FEG		I		E	
True Score Gain	TSG		I	E		
Covariate Adjusted with School Fixed Effects	CAFE			I	E	
Covariate Adjusted with School Random Effects	CARE			I	E	
Simple Panel Growth	PANEL			I E		
Layered Model	LM				I E	
Student Growth Percentile (Quantile Regression)	SGP					I E
Growth to Standards	GTS		I E			

Simple Gain (Gain[14]): We use a simple gain model as gains are a direct measure of student growth, transparent, and scores are low inference, in that the average gain for a school is easily interpreted. A simple gain is calculated for each student and averages are calculated for schools. This approach ignores the clustering of students in schools and thus explicitly ignores school context (Burstein, 1980). A model that ignores the clustering of students within schools and simply aggregates individual student gains up to the school level potentially produces biased estimates of school effects (Raudenbush & Willms, 1995). This occurs because estimates that ignore the fact the students attend specific schools mixes within and between school estimates when the intraclass correlation is greater than zero (Aitkin & Longford, 1986). This applies to any model that is based on individual student scores that are aggregated up to a school. A gain is simply:

$$\mathrm{Gain}_i = \Delta_i = Y^s_{it} - Y^s_{i(t-1)} \qquad (3)$$

where Y^s_{it} is the assessment outcome for student *i* in subject *s* at time *t*. The inference is that a school's performance is based on average gains in its students' performance. Score meaning, in its simplest form, relates to a notion that students have more mastery of a given content. As noted, in order for gains to be meaningfully interpreted, assessment scores need to be on a vertical scale. However, in some cases researchers have normalized scores (z-scores) within grade levels under the assumption that performance standards are vertically moderated thus allowing for consistent meaning across grades at various anchor points. Such an approach moves away from a strictly absolute conception of growth to one that considers growth relative to standards. Simple gain score models do not account for differences in starting points and potentially suffer from the "growth to nowhere" criticism that concerns some policymakers.

Fixed effects Gain (FEG): A closely related model is a fixed effects gain model. This approach uses the calculated gain, Δ_i, as the outcome, but explicitly includes an indicator for each school in the system. This indicator captures the school effect.

There are some concerns with accountability based on gains because gains may not be sufficiently reliable, although there is some debate about this (Rogosa & Willett, 1983)

[14] The term in parentheses is the abbreviated name we use interchangeably with a model's full descriptive name.

and concerns about the spurious negative correlations between Δ_i and Y_{t-1} because Y_{t-1} is generally not an error free measure.

True Score Gain (TSG): One specific approach to avoid the spurious negative correlation between Δ_i and Y_{t-1} is to utilize a true score gain model that uses both Y_{it-1} and Y_{it} on the left hand side of the equation to estimate gains directly, thus avoiding any spurious relationship between gains and pre-test scores. We present this model in somewhat more detail because it conceptually provides a good framework for evaluating schools, but it has received less attention in the literature (Goldschmidt et al., 2005).

This model uses data from the current year *t* and a prior year *t-1*, and we explicitly model true student gains. The model is based on previous work by Bryk, Thum, Easton, & Luppescu (1998) and Choi (2007). At level one[15]:

$$A_{tij} = \alpha_{1tij}\pi_{1ij} + \alpha_{2tij}\pi_{2ij} + e_{tij} \qquad (4)$$

Here, for student *i*, with school *j*, at time *t*, the assessment scale score is denoted as A_{tij}. Time in this instance refers to the pre-test, $t = 0$, and the post-test $t = 1$. Eq. 4 estimates two parameters, student's initial status for the pre-test (π_{1ij}) and gain on the post-test (π_{2ij}). Given this parameterization, α_1 is coded as 1, and $\alpha_2 = 0,1$ for the pre- and post-test, respectively. The error, e_{tij}, is assumed to be $N\sim(0,\sigma^2)$.

This formulation is a basic growth model formulation which could include a time varying covariate (Raudenbush & Bryk, 2002). There are two concerns: one, the within-student error is assumed uncorrelated with other potential explanatory variables, which may not be a tenable assumption; and two, there may not be enough degrees of freedom available to estimate the random effects of interest. In order to model true initial status (pre-test) and true gain, as well as to begin to consider potential remedies, eq. 4 is reparameterized by scaling both sides of the equation by the inverse of SEMs, i.e., $1/S_{tij}$:

$$A^*_{tij} = \alpha_1{}^*_{tij}\pi_{1ij} + \alpha_2{}^*_{tij}\pi_{2ij} + e^*_{tij} \qquad (6)$$

[15] The growth model is based on the multilevel model framework presented in Raudenbush and Bryk (2002) where test occasions are nested within students, who are nested within teachers. Hence, level one is the within student model over time, level two is the between student model, and level three is the between teacher model. Education researchers often generate the error structure of the models by considering the "levels" or structure of the data. Given that this model is based on prior work by education researchers, we adopt this nomenclature for this model.

Thus e^{*}_{ijt} now becomes $N \sim (0,1)$, and π_{1ij} and π_{2ij} now estimate a student's true initial status and true gain, respectively (Bryk et al., 1998). By using a true score gain model at level one and including all achievement measures on the left-hand side of eq. 6, there is no spurious correlation between pre-tests and the error as could be the case in a standard pre-test as covariate model (Bryk et al., 1998). At level two, student covariates can be incorporated to account for between-student differences in true initial status and true gains. Hence at level two:

$$\pi_{1ij} = \beta_{10j} + X_{ij}\beta_{11j} + r_{1ij}, \qquad (7a)$$

$$\pi_{2ij} = \beta_{20j} + X_{ij}\beta_{21j} + r_{2ij}. \qquad (7b)$$

where X_{ij} represents a vector of time invariant student covariates, such as gender, language, and economic status. With this specification, the student's achievement gain, as measured by π_{2ij}, is allowed to vary randomly from student to student. The between-school model is:

$$\beta_{10j} = \lambda_{100} + u_{10j}, \qquad (8a)$$

$$\beta_{20j} = \lambda_{200} + u_{20j}. \qquad (8b)$$

This true score gain model developed by Bryk et al. (1998) is included in the set of models we consider. The interpretation of this model receives extensive treatment in Bryk et al. (1998), and we focus on the estimated true gain component as the indicator of school performance. In this way we can directly compare results of the true score gain model to a simple gain model and a fixed effects gain model.

While this model addresses measurement error in the pre-test, an extended form of this model attempts to mimic fixed effects estimates and further reduce bias. This has been accomplished in Goldschmidt, Tseng, & Goldhaber (2010) using specific centering strategies that can reproduce fixed effects estimates (Allison, 2009) while retaining advantages of random effects models. That is, through person specific centering (group mean centering – where at level 1, group refers to student) it is possible to include time invariant student covariates, account for the intraclass correlations, generate correct standard errors, and estimate cross level interactions (i.e., teacher by COV*x* interactions) that are not possible in the fixed effects framework.

Growth to Standard (GTS): We also examine another model based on gains, but one that explicitly eliminates the "growth to nowhere" concern. We estimate a growth to standard model that is similar to one of the state's Growth Model Pilot Study models. However, we apply the model not on the margin, rather as the single criteria by which schools are monitored. A student's score, *Y*, in year *t* is compared to the proficient score, Y_p, in year *t+2* and a required gain is calculated based on the difference between the student's actual year t score and the required year *t+2* score. The year *t* to *t+1* required gain is:

$$\text{Gain*} = (Y_{p(t+2)} - Y_t)/2 \qquad (9)$$

A school's score is the percent of students where actual gain ≥ Gain*. In year *t+1*, Gain* is reevaluated and adjusted up or down depending on the progress made between year *t* and *t+1*. This model explicitly addresses inferences on whether a student gained enough. Likely, school rankings based on this model will be correlated to status models as the results depend entirely on whether the current year score is sufficient to pass the set threshold[16]. The difference between this model and the general AYP model is that each student has an individual threshold.

Covariate Adjusted Model with School Fixed Effects (CAFE) or with School Random Effects (CARE): Fitzmaurice, Laird, &Ware (2004) argue that the choice between analysis of gain scores versus a covariate adjusted model depends on the research question. A covariate adjusted model tests how students differ on the post-test given that students started with the same pre-test score. Gain scores test how groups of students differ in gains, on average. We use both a fixed and a random effects covariate adjustment, or ANCOVA model. There are several key distinctions between ANCOVA models and gain models. One is that ANCOVA models are quite similar to gain models in that the model in eq. 1 can be thought of as

$$\Delta_I = Y^s_{it} - \delta Y^s_{i(t-1)} \qquad (10)$$

where δ =1. If we simply move the prior year score to the right hand side of the equation then:

$$Y^s_{it} = \delta Y^s_{i(t-1)} + \text{(fixed effects or random effects, covariates, and error)} \qquad (11)$$

which has the benefit of allowing δ, the relationship between the prior assessment and the outcome, to be estimated empirically. The fixed and random effects models are presented in eqs. 12 and 13.

[16] Students who are proficient but whose gain would put them below the proficient performance level (i.e., a negative gain) are not counted as proficient in the current year.

$$Y_{it} = (\theta_j + \mu_i) + \delta Y_{it-1} + e_{it}\text{ , where } e_{it} \sim IID(0,\sigma^2) \qquad (12)$$

And a random effects model takes the form:

$$Y_{it} = \alpha + \delta Y_{it-1} + (\mu_i + e_{it})\text{ , where } e_{it} \sim IID(0,\sigma^2) \qquad (13)$$

The ANCOVA model presented in eq. 13 likely produces biased results if the covariate is measured with error (McCaffrey et al., 2004). Using multiple prior assessments as predictors ameliorates much of the bias in estimates (Wright, 2008) and may be preferable (Ballou, Sanders, & Wright, 2004; Wright, 2008). For this reason we specified a fixed effects ANCOVA model with multiple prior assessments, including assessment in cross subjects (e.g., the mathematics model included both prior mathematics assessments and prior English Language Arts [ELA] assessments). We included two prior mathematics and two prior ELA assessments, along with a full set of control variables (student background). We use a "full model," one that includes multiple years of prior scores and student background, only in some instances as the correlation of school effect estimates between the fixed effects model results including and excluding student covariates are highly correlated >.97 in these datasets.

One advantage of the ANCOVA model is that it is more robust to usage with either vertical or non-vertical scales. An important, but subtle, distinction is that ANCOVA models do not provide results in terms of growth, rather they address current student performance explicitly accounting for differing initial performance.

Student Growth Percentiles (SGP): The notion of relative performance given prior performance is taken a step further in the SGP model. The SGP model does not address absolute growth in performance. However, it broadens the notion of robustness to scale by focusing on normative position based on student percentile ranks. The SGP model is fully detailed in Betebenner (2009). The SGP uses quantile regression to measure a student's progress from one year to the next in comparison with his or her academic peers with similar test score histories. For example, if a student's SGP score is 75 in fifth-grade mathematics, he grew as much or more than 75 percent of his academic peers with similar score histories. Thus, the SGP measure is interpreted in a normative sense. One cannot conclude that two students, each of whom obtained an SGP score of 75 but who had different prior year test scores, grew by the same absolute amount. All students with SGP scores of 75 experienced more growth than their peers with similar prior scores. In this way, the SGP model addresses normative changes in performance and is generally readily interpretable.

The SGP model proposed by Betebenner (2009) is estimated using quantile regression, which was introduced by Koenker and Bassett (1978). A quantile is a particular percentile point in a distribution. While typical Ordinary Least Square (OLS) regression models estimate the conditional mean of the response variable Y for each value of a covariate X, quantile regression extends the regression model to conditional quantiles of the response variable. For example, a quantile regression for the 75th percentile shows the score at the 75th percentile of Y for each value of X. The SGP model handles non-linearities in the relationships among the ranks quite well through the use of quantile regression models.

A student's growth percentile as a measure of student progress from one year to the next is calculated by estimating the conditional density associated with a student's score at time *t* (i.e., current year) using student's prior scores at times 1, 2, …, *t*-1 (i.e., prior years) as the conditioning variables (covariates). In other words, a student's current year score is situated normatively as a student's growth percentile by taking the student's past student performance into account. Hence, it is possible to model how the dependent variable responds to changes in covariates across the distribution (Hao & Naiman, 2007). Moreover, estimates using traditional regression models are sensitive to outliers or skewed distributions, whereas quantile regression model estimates are not.

The quantile regression allows more flexibly in modeling achievement growth across the student score distribution because the achievement growth coefficient (θ in the traditional fixed effects model) can vary by quantile. Quantile regression proceeds in exactly the same way as does an ordinary least squares minimization problem except in order to obtain estimates, the conditional quantile function in eq. 14 is minimized (Koenker, 2005).

$$\min \beta \in \Re^p \Sigma\rho_\tau (y_i - \xi(x_i,\beta)) \qquad (14)$$

The resulting minimization problem, when j(*x*, β) is specified as a linear function of parameters, can be readily solved with specific routines in R (R Development Core Team, 2009) and SAS^{TM}.

To obtain teacher-level, grade-level, and school-level SGP estimates, estimates of student-level growth percentile scores are aggregated to higher units (i.e., teacher, grade, and school). One of the key advantages of SGP is that after SGP is estimated at the student level, it is quite simple to combine them into higher level aggregates. To summarize teacher- and/or school-level SGP as a single number, the median of the SGP distribution for the teacher, grade, or school is usually used. The median represents the growth of a "typical" student in a given teacher's class or a given school.

Advantages of the SGP approach include its robustness to scale requirements and that the normative interpretation of student growth from one year to next is very understandable to a broad array of stakeholders. Also, it is easy to aggregate obtained student growth percentiles to higher units (e.g., teachers and schools). As previously noted, school effects estimated as simple aggregates provide a combined within and between-school effect estimate.

One disadvantage is that SGP models require a substantial amount of data in order to generate sufficient coverage across the percentiles. At the state level this should not be an issue. Also, SGP models generally include only contiguous prior test score histories. We estimate the SGP with all contiguously available, same subject assessment results.

Layered Model (LM): One method that is robust to missing data is an approach that is based on a mixed effects regression model. An example of a layered model is the model used in Tennessee, the Tennessee Value-Added Assessment System (TVASS). The "layered" model simultaneously models scores for multiple years in multiple subjects. Later years of teacher or school effects build upon estimated effects from earlier years (thus "layering" effects onto one another). This model has been applied in several states and districts. Covariates such as student background variables are typically not included.

This model is robust to missing data as it can include a variable number of non-contiguous prior test scores and explicitly uses assessment results across multiple subjects. The advantage to this is that it includes as much information as possible about each student. This approach has been demonstrated to effectively reduce bias (Wright, 2008) and substantively reduce the need for student background information (Ballou et al., 2004). This mixed effects approach layers gains over years and includes indicators for schools (or teachers for teacher level models). For a school effect:

$$Y_{ijklsn} = \mu_{ijkls} + e_{ijklsn} \qquad (15)$$

where Y represents an assessment for student *n* in subject *s* in year *k* and grade *l*. And the student attended school *j*, which is in school system *i*. The fixed mean for all students in the combination of subjects, grades, years, schools, and system is μ_{ijkls}; e_{ijklsn} is the random deviation for student *n* from the mean, μ_{ijkls}. (Sanders, Saxton & Horn, 1997). The complexity arises in the covariance structure of e_{ijklsn}, which, given the general use of five years of data (in TVASS) results in 120 elements. The layered model we employ uses the same structure but is limited to a maximum of three years and two subjects. This is likely a sub-optimal form of the layered model, but allows us to compare how the various models compare using three years of student data.

Two major advantages of the LM is its robustness to missing data and the inclusion of the maximum amount of student data available. A disadvantage is that, in essence, the layered model is based on the student’s gain in adjacent years/grades. Thus, it requires appropriately scaled scores. If vertically scaled scores are not available, it is recommended to convert them to NCE (noral curve equivalent) scores based on a statewide distribution.

Simple Panel Growth (Panel): The final model we consider is a simple longitudinal mixed effects model. Multilevel growth models monitor school performance by taking into consideration the nature of the data and by attempting to mitigate the effects of potential confounding factors (PCF). The effects of PCFs in non-randomized, cross-sectional designs (Campbell & Stanley, 1963) and limitations of pre-post designs (Bryk & Wesiburg, 1977; Raudenbush & Bryk, 1987; Raudenbush, 2001) in making inferences about school effects (i.e., change in student outcomes due to a hypothesized cause) leads many to consider the advantages of examining growth trajectories to make inferences about change (Rogosa, Brandt, & Zimowski, 1982; Willet, Singer, & Martin, 1998; Raudenbush & Bryk, 2002). Our model utilizes three years of data, but is able to include students whether they have one, two, or three assessments. These models theoretically require a vertical scale in order to maintain interpretable meaning, but can be estimated using normalized scores so long as the focus is on comparisons of schools rather than inferences about absolute growth. The notion is that assessment results are a function of time and the growth is directly estimated:

$$Y_{tij} = \gamma_{0tij} + \gamma_{1ti} T_{1ij} + e_{tij} \qquad (16)$$

where, Y_{tij} is the assessment score at time *t* of student *i*, who attends school *j*. In eq. 16, γ_{0tijti} is student *i*'s performance when *T*=0 (allowing γ_{0tijti} to be interpreted as student *i*'s initial status); γ_{1ti} is the estimated effect per time interval, or assessment occasion, that is, achievement growth. Specifically, the model we utilized is displayed in eq. 17. Similar to the layered model, the school effect is the deviation of school *k*'s trajectory U_{10j} from the average trajectory γ_{100}.

$$Y_{tij} = \gamma_{000} + \gamma_{100} GRADE_{tij} + r_{0ij} + r_{iij} GRADE + U_{00j} + U_{10j} GRADE_{tij} + e_{tij} \qquad (17)$$

It is common to code T in a manner that allows for specific interpretation of the intercept and growth. In this case Grade (T in eq. 16) is coded such that it equals 0 in grade 5[17] (for elementary school students) and in grade 8 (for middle school students). Other specifications are possible (e.g., centering a school's performance on the average grade Goldschmidt, Choi, Martinez, & Novack, 2010). The interpretation is that a school's performance is based on the school's growth evaluated in 5th and 8th grades. This approach ignores status performance, though that could be included either conjunctively or in a compensatory manner. Similar to the LM, one advantage of the Panel model is that it is robust to missing data, but like the LM, it requires a vertical scale for meaningful interpretation of growth, although relative comparisons among schools are still possible without a vertical scale.

Methods

The analyses include multiple comparisons that are intended to provide policymakers some guidance in choosing among the many potential model choices. As noted above, several studies have examined the relationships of some models and model specifications, but this evaluation differs in that we attempt to address practical consequences surrounding model choice. We emphasize again that the first consideration is the purpose of the model and the desired inferences about schools. As we outlined in the model section above, some models explicitly compare schools on gains or growth, others consider deviations from gains, while others consider current performance accounting for past performance, or current ranking considering prior rankings. In order to address the practical implications of choosing a model, we investigate the following questions:

[17] Grade 3 would equal -1 in grade 4 and -2 in grade 3.

1. Overall, does the model matter?
2. Do different models lead to different inferences about schools?
3. How accurately do models classify schools into performance categories?
4. Are models consistent in classifying schools from one year to the next?
5. How are models influenced by school intake characteristics?
6. Do models perform similarly for elementary and middle schools?
7. Do models behave similarly across states?

We examine a range of models that create indicators of school performance and evaluate how results overlap or differ and how, sometimes, disparate results can be combined. As noted we are guided by much of the previous research on estimating school and teacher effects using some variation of a growth model. There are a plethora of models to examine and many provide substantively similar results (Tekwe et al., 2004), or differences can be systematically accounted for (Choi, Seltzer, Herman, & Yamashiro, 2005; Ballou, Sanders, & Wright, 2004). Previous work on model specification and teacher effects indicates that teacher effects are fairly robust to specification (Lockwood & McCaffrey, 2007; McCaffrey et al., 2004), and these results likely hold for school effects as well. Empirical results between different modeling approaches (i.e., fixed effects, random effects, and layered models) at the school-level indicate that value added estimates[18] are highly correlated among approaches (Tekwe et al., 2004).

We begin by presenting static school descriptive information, related student background characteristics, and school size. This serves two purposes: one, to provide context; and two, to present some basic information that highlights differences and similarities among the states represented in the analysis. Beyond school descriptive statistics, variance decomposition is a straightforward way to summarize the proportion of variation in student outcomes that is attributable to schools[19] – which also provides an indication of how homogeneous students are within schools. It is important to bear in mind that only that

[18] Here we refer to value added as a subset of growth models that generate school effects based on either a random effect (deviation from average) or a fixed effect that directly estimates a school effect as a fixed parameter on a school indicator variable.

[19] Here we focus on within and between-school variation in student outcomes. Within school variation in student performance could be further decomposed (e.g., within and between teachers).

portion of the total variation attributable to schools will be amenable to policies addressing between-school differences in outcomes[20].

In order to address these questions we generate several school effects for each school. First, students are divided into cohorts (as displayed in Figure 1). That is, there is a 2007-2009 cohort and a 2008-2010 cohort. Analyses are conducted separately for each cohort. Each cohort's results are further divided by subject (mathematics and reading) and level (elementary and middle) and school effects are estimated separately by subject and level. Auxiliary models considered several specification options within each model related to covariates that included combinations of prior test scores and student background characteristics. This generated approximately 60 estimates per school — which, while informative, does not substantively add to the analyses presented here.

Models are estimated for each state separately. Because some states have vertical scales and some do not, all scores are normalized for those models that theoretically require a vertical scale[21]; specifically, these are the gain and Panel models.

In order to conduct a general test for model effects (discussed in more detail below), we generate normalized school effect estimates that allow for comparisons across states. These are estimated as:

$$\theta^*_{jkmcsg} = (\theta_{jkmcsg} - \theta_{..kmsg}) / SD_{kmsg}, \qquad (18)$$

where θ_{jkmcsg} is the school effect for school *j* in state *k* estimated with model *m* for cohort *c* in subject *s* in school level *g*; $\theta_{..kmsg}$ is the estimated mean effect for state *k* estimated with model *m* in subject *s* in school level *g*; and SD_{kmsg} is the standard deviation of $\theta_{..kmsg}$. We use this effect size primarily to estimate a fixed effects model that simultaneously tests for state, model, subject, cohort, and joint effects. We estimate the model in eq. 18 separately for elementary and middle schools. We use estimated precision (squared) as the regression weights as is typical if we conceptualize this specific analysis as a quasi meta-analysis. In this case we estimate:

$$\theta^*_{jkmcsg} = u_{kmcs} + \Sigma_1^K \delta(\text{state}_k) + \Sigma_1^M \gamma_{1m}(\text{model}_m) + \gamma_2(\text{cohort}_c) + \gamma_3(\text{subject}_s) +$$

[20] Of course some policies aimed at schools may also decrease within school variability and improve overall performance.

[21] This is an example of preliminary analyses we conducted that examined the impact on estimated school effects when estimating gains on actual scale scores and on normalized scores. Again, this effect was considerably less relevant than the ones we present here and is consistent with previous research Goldschmidt, Choi, Martinez, & Novack, 2010).

$$\Sigma_1^L \lambda(\text{school cluster}_l) + e_{jkmcsg} \qquad (18b)$$

The F-test for γ_{1m}s tests whether the model significantly impacts a school's estimate, while the test of γ_2 tests whether the school effect is consistent from one year to the next (i.e., does a school's estimate differ from one cohort to another, *ceteris paribus*). Similarly, γ_3 tests for subject effects (i.e., does a school's effect differ when estimated using math or language arts, *ceteris paribus*). A series of indicator variables are used as fixed state effects, δ, and are included to capture unobserved state circumstances; since assessments are unique to states, δ captures both unobserved state characteristics and variation among states in the psychometric properties of the assessments. Including school cluster, *l*, provides some standardization among states in school characteristics (recall that clusters are not nested within state and are fully crossed with state). The test of λs tests whether school effects systematically differ by cluster. We also expand eq. 18 to include model, by state, cohort, subject, and school cluster interactions. Interactions (or joint effects) are important to examine whether there are differential effects of the models by the covariates included above[22]. The joint test of the interaction terms of model and cluster tests whether the model results differ by cluster or conversely whether cluster results differ by model. If the null hypothesis (of no effect) can be rejected, then this implies that models behave differently among the clusters.

The bulk of the analysis consists of the growth model comparisons. The goal of these comparisons is to examine the effect of various models on inferences about school performance, specifically focusing on issues related to precision, reliability, stability, and bias. Two cohorts consisting of three years of panel data are used to examine the relationship among various school accountability models. As noted, we conduct comparisons separately for elementary (grades 3-5) and middle schools (grades 6-8) and separately for mathematics and ELA. High schools are excluded as they do not include contiguous state test results.

Estimated precision is one of the key statistics of school estimates as it allows us to determine how accurately we can evaluate a school's performance. We use the estimated standard error (SE) of the estimate for each model and school to examine various facets of precision. We note that the magnitude of the SEs vary by test metric and model, and hence simply comparing the magnitude does not provide a good judge of model precision. Instead

[22] This also changes the interpretation of the main effects, as they pertain to the left-out category of the interacted variable.

we evaluate precision by examining how well models classify schools into performance bands.

Reliability is directly linked to precision and is often incorrectly used interchangeably with precision. Reliability estimates intend to provide some information related to the ability of the model to detect differences in true school performance. However, there has been some debate related to the role of reliability, particularly when models are based on growth and the intended use is school accountability (Rogosa, 2002). This is consistent with other cautions about interpreting reliability when considering growth (Raudenbush & Bryk, 2002). We consider reliability in terms of the shared variance in consecutive years of school effect estimates per model. This conception is based on a regression of the year_2 estimate on the year_1 estimate from which R^2 provides an estimate of common variance, and can be interpreted as an estimate of reliability. Empirical results lend credence to the notion that reliability in the context of growth is not necessarily a good indicator of model performance because this is often confused with the model's ability to "accurately" measure growth. As Rogosa (2002) points out, reliability is neither a necessary nor a sufficient condition to guarantee actuate indicators or performance.

We also examine stability, by placing schools into performance quintiles in each of the two years and checking how many schools remain in the same quintile. This measure is imprecise because it confounds true stability with true school movement, but is a common analysis undertaken to evaluate the stability of teacher effect estimates (see Koedel & Betts, 2007, for example).

We also compare correspondence in performance band classifications across models. Comparing how models classify schools in performance quintiles is substantively more relevant since this is generally what state accountability systems intend to accomplish with school estimates (e.g., A-F school grades and school categories such as average, above average).

Bias is best addressed through simulations as this is the only way to know exactly what constitutes truth and what constitutes bias. Since we are using actual student data, we examine this concept indirectly by examining the influence of student background characteristics on school effect estimates and by comparing results among models. While a simulation study isolates a particular facet and allows a specific test to determine whether that facet is adequately reproduced or accounted for, these analyses have no such control. Comparing any two models would not unequivocally identify the correct model and the incorrect model. Given there is no true model, in a strict sense this is not identifying bias, rather whether among models there is consistency in results. That is, are models unequivocally interchangeable, ignoring that alternative theories of action can and should lead to differences in how schools are classified?

Specifically, we conduct the following analyses to address each of the research questions posed above.

Overall, does the model matter? As described above, we model all normalized school effects simultaneously along with fixed effects for model, state, cohort, subject, and school level (as well as the interactions of model and the state, cohort, subject, and school level variables).

Do different models lead to different inferences about schools? States often make inferences about schools by placing them in performance categories. We divided the schools into five equally sized groups (quintiles) based on their growth scores and examined how well the models placed schools into those categories. Based on the size of the interval and on the standard deviation of the school effect, we estimated how likely a school is to fall within the interval. Additionally we correlate model results.

How accurately do models classify schools into performance categories? Schools are placed into performance quintiles. Comparing how models classify schools in performance quintiles is substantively more relevant since this is generally what state accountability systems intend to accomplish with school estimates (e.g., A-F school grades and school categories such as average, above average).

Are models consistent in classifying schools from one year to the next? We can look at both the shared variation in school estimates using the prior year school effect estimate as an independent variable and the current year as the dependent variable (i.e., $R^{2)}$. We estimate stability using the correlation between the prior year estimate and the current year estimate. These two methods are essentially the same since the square root of R^2 is r.

How are models influenced by school intake characteristics? We use Ordinary Least Squares regression of normalized school effects on aggregate school inputs: Students with Disabilities (SWD%); Minorities (Minority%); Economic Disadvantage (ED%); School

Stability (Stability%); School Size (School N). We evaluate whether intake characteristics have linear and/or non-linear effects.

$$\theta_{jkmcg} = u + \beta_1(\text{SWD\%}) + \beta_2(\text{Minority\%}) + \beta_3(\text{ED\%}) + \beta_4(\text{Stability\%}) + \beta_5(\text{School N}) + \beta_6(\text{SWD\%})^2 + \beta_7(\text{Minority\%})^2 + \beta_8(\text{ED\%})^2 + \beta_9(\text{Stability\%})^2 + \beta_{10}(\text{School N})^2 + e_{jkmcg}$$

Do models perform similarly for elementary and middle schools? Using the methods described above we compare how models produce results for elementary schools and compare them to how they produce results for middle schools. For example, we may observe that the estimated stability (year-to-year correlation of school effects) for a particular model is .6 for elementary schools, but may be .5 for middle schools. We subjectively compare results across models and school level to see whether we can identify specific areas where policymakers might need to use caution in applying a model. Again, the omnibus test for whether the school level matters is addressed in question one with the fixed effects model on standardized school effect estimates.

Do models behave similarly across states? Similar to the previous question, the overall test for this is conducted with the fixed effects model on normalized school effects. But we use the other analyses outlined above to examine more subjectively whether models behave equally across states (which, as noted, confounds potentially several state unique contextual elements).

The analyses outlined above can take many levels of disaggregation (i.e., model A can be compared to model B, overall, or by school level, by cohort, by subject, and by state — or any combination of these). Many additional analyses are possible, and while potentially interesting, would tend to obfuscate the general results.

RESULTS

We present results systematically, addressing each of the research questions noted above utilizing specific elements of the comparisons noted above. We attempt to report results as concisely as possible and place many of the detailed tables into Appendix 3.

While not strictly addressing how models compare, considering how states differ in the structure of school performance provides evidence for the notion that states are demonstratively different. Table 2 provides descriptive information for each state, and the classification process of finding similar schools across states also reveals that states differ substantively in terms of the enrollment characteristics of the schools. One aspect to consider up front is the ICC (intraclass correlation), or the proportion of variation in student achievement that is attributable to schools[23]. The ICC is also a measure of how homogeneous students are within a school on a particular outcome. Table 5 below summarizes the results by state, subject, and grade level. Previous research indicates that ICCs tend to range from about .1 to around .3 (Raudenbush & Bryk, 2002). The results in Table 5 generally corroborate those findings. There are some instances of more extreme values. For example, in state S2, there are some cohort/subject/school levels where the ICC is approximately .4, implying a fair amount of homogeneity among students within schools. On the other extreme, state S3 demonstrates considerable within-school heterogeneity. This provides further reason to classify schools on a common metric in order to reduce confounding among state specific factors. The ICC is important because it provides an upper-bound as to what proportion of student achievement could be accounted for by variation in school processes.

[23] Assuming a random effect model, the ICC is $\tau_j/(\tau_j + \sigma_{ij}^2)$. See Raudenbush and Bryk (2002) where τ_j is the variance of the school random effects and σ_{ij}^2 is the variance of student level (generally measurement) error.

Table 5:

Variation in Student Performance attributable to schools - grade 5

	S1				S2				S3				S4			
	C09[1]		C10[2]		C09		C10		C09		C10		C09		C10	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA
2007	0.22	0.16			0.21	0.19			0.07	0.10			0.20	0.16		
2008	0.22	0.19	0.21	0.17	0.20	0.20	0.19	0.18	0.05	0.11	0.07	0.10	0.20	0.17	0.19	0.18
2009	0.16	0.14	0.19	0.17	0.21	0.27	0.22	0.21	0.05	0.12	0.05	0.11	0.21	0.18	0.20	0.18
2010			0.16	0.15			0.19	0.23			0.05	0.11			0.21	0.18

Variation in Student Performance attributable to schools - grade 8

	S1				S2				S3				S4			
	C09[1]		C10[2]		C09		C10		C09		C10		C09		C10	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA
2007	0.29	0.24			0.29	0.21			n/a	n/a			0.11	0.17		
2008	0.30	0.22	0.30	0.25	0.28	0.32	0.28	0.30	n/a	n/a	0.04	0.10	0.22	0.17	0.22	0.17
2009	0.24	0.21	0.27	0.23	0.32	0.43	0.30	0.34	n/a	n/a	0.04	0.08	0.22	0.17	0.19	0.17
2010			0.19	0.17			0.37	0.40			0.02	0.08			0.22	0.17

1) C09 = the 2009 cohort. 2) C10 = the 2010 cohort.

Overall, does the model matter?

We examine this question by using the estimated standardized school effect size estimates in a quasi-meta-analytic model described above (eq. 18). The results are summarized in Table 6. A detailed elaboration of results is displayed in Appendix 3.

Table 6. Overall Impact of Model on School Effect Estimates
Type III Tests of Fixed Effects

	Elementary School			Middle School		
Source	Num df	F	Sig.	Num df	F	Sig.
Intercept	1	0.71	0.40	1	0.45	0.50
Model	10	47.24	0.00	7	13.17	0.00
State	3	0.37	0.77	3	0.80	0.49
model * state	29	22.39	0.00	20	2.94	0.00
Subject	1	0.35	0.56	1	0.11	0.73
Cohort	1	0.67	0.41	1	0.00	0.98
model * subject	10	60.38	0.00	7	7.02	0.00
model * cohort	10	62.06	0.00	7	2.77	0.01
Compared to Schools in Advantaged Category:						
Disadvantaged I	1	0.04	0.84	1	1.53	0.22
Large	1	0.01	0.92	1	0.43	0.51
Disadvantaged II	1	0.03	0.87	1	6.75	0.01
Mobile	1	8.30	0.00	1	5.59	0.02
model * Disadv I	10	4.04	0.00	7	0.88	0.52
model * Large	10	17.72	0.00	7	12.56	0.00
model * Disadv II	10	14.02	0.00	7	15.11	0.00
model * Mobile	10	42.68	0.00	7	90.55	0.00

Note: Numerator = Advantaged. *df* elementary = 84,783; middle 54, 102
AIC: **Elementary School,** Null model 526983, Full Model 456748; **Middle School** Null model 235233, Full Model, 233107.

The results in Table 6 imply that school effect estimates do, in fact, vary by model in elementary school (*p* <.01) and in middle school (*p* < .01). Moreover, the results imply that estimated school effects vary jointly with model and state, which means that different models generate school effects differently in each state. The interaction effects related to model by subject and model by cohort imply that individual school effect estimates vary by model and subject and cohort. The results in Table 6 also indicate that similar schools within a cluster will be rated differently by different models.

Overall, the results in Table 6 confirm that there will be differences in school effect estimates and that these differences depend on both school context and the model used to estimate the effect. While we rejected the null hypothesis that models are statistically equal, we next disaggregate results more carefully to determine whether these differences relate to practical differences that policymakers ought to take into account when selecting a growth model.

Do different models lead to different inferences about schools?

We examine this question by looking at the between model correlations of estimated school effects and by calculating how often models place schools into the same performance category. Tables 7 through 10 present the unconditional relationships among models. In order to capture the variability across states, we present both the observed maximum and minimum correlations of school effect estimates. Consistent with expectations, results based on status are fairly highly correlated with models also basing inferences on current or conditioned current performance. For example, the results in Table 7 indicate that the estimated correlation between schools' performance estimated by a status model and schools' performance, GTS can be virtually perfectly correlated and very highly correlated with the CAFE model. On the other hand, status provides a different picture than simple GAIN (or FEG) models, which would be expected.

Table 7:

Correlations[1] among models - Elementary ELA

	Status		Gain		FEG		CAFE		TSG		CARE		Panel		SGP		LM		GTS	
	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
Status	1.00	1.00	0.23	-0.50	0.24	-0.42	0.93	0.71	0.71	0.13	0.62	0.29	0.64	-0.53	0.60	0.37	0.63	-0.40	1.00	0.27
Gain			1.00	1.00	0.96	0.88	0.30	-0.55	0.90	-0.30	0.79	0.14	0.64	-0.26	0.79	-0.16	0.86	0.38	0.42	-0.49
FEG					1.00	1.00	0.30	-0.31	0.87	-0.13	0.76	0.35	0.64	-0.02	0.79	-0.01	0.87	0.38	0.35	-0.39
CAFE							1.00	1.00	0.75	0.19	0.65	0.34	0.76	-0.56	0.72	0.43	0.94	-0.35	0.94	0.84
TSG									1.00	1.00	0.90	0.64	0.70	0.57	0.84	0.75	0.78	-0.14	0.70	0.15
CARE											1.00	1.00	0.84	-0.07	0.83	0.65	0.83	0.25	0.62	0.30
Panel													1.00	1.00	0.76	0.23	0.60	-0.11	0.66	-0.53
SGP															1.00	1.00	0.80	0.04	0.59	0.35
LM																	1.00	1.00	0.85	-0.42
GTS																			1.00	1.00

[1] Represent the correlations among school effect estimates

Table 8:

Correlations[1] among models - Elementary Math

	Status		Gain		FEG		CAFE		TSG		CARE		Panel		SGP		LM		GTS	
	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
Status	1.00	1.00	0.35	-0.40	0.32	-0.49	0.88	0.75	0.69	0.06	0.52	0.30	0.54	0.01	0.57	0.32	0.64	-0.18	1.00	0.33
Gain			1.00	1.00	1.00	0.85	0.28	-0.46	0.95	0.01	0.88	0.29	0.66	0.10	0.84	0.16	0.94	0.52	0.32	-0.41
FEG					1.00	1.00	0.28	-0.36	0.91	0.05	0.88	0.37	0.62	0.12	0.84	0.18	0.94	0.54	0.32	-0.49
CAFE							1.00	1.00	0.63	0.03	0.61	0.35	0.63	0.07	0.51	0.33	0.67	-0.22	0.95	0.79
TSG									1.00	1.00	0.87	0.71	0.69	0.51	0.88	0.73	0.73	0.32	0.68	0.01
CARE											1.00	1.00	0.81	0.59	0.91	0.74	0.93	0.46	0.52	0.31
Panel													1.00	1.00	0.80	0.56	0.67	0.25	0.57	0.05
SGP															1.00	1.00	0.89	0.43	0.55	0.42
LM																	1.00	1.00	0.64	-0.18
GTS																			1.00	1.00

[1] Represent the correlations among school effect estimates

Table 9:

Correlations[1] among models - Middle School ELA

	Status		Gain		FEG		CAFE		TSG		CARE		Panel		SGP		LM		GTS	
	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
Status	1.00	1.00	0.23	-0.50	0.31	-0.62	0.94	0.25	0.53	-0.10	0.53	0.02	0.28	-0.35	0.59	0.07	0.51	-0.27	1.00	0.27
Gain			1.00	1.00	0.96	0.69	0.49	-0.62	0.67	-0.08	0.60	0.29	0.48	0.18	0.78	-0.01	0.53	0.16	0.42	-0.49
FEG					1.00	1.00	0.43	-0.73	0.71	-0.21	0.64	0.30	0.44	0.00	0.78	0.03	0.76	0.40	0.41	-0.58
CAFE							1.00	1.00	0.67	-0.11	0.59	0.16	0.90	-0.37	0.76	0.41	0.98	-0.61	0.97	0.83
TSG									1.00	1.00	0.91	0.60	0.80	0.60	0.75	0.43	0.69	0.03	0.60	-0.09
CARE											1.00	1.00	0.73	-0.01	0.73	0.48	0.75	0.39	0.53	0.02
Panel													1.00	1.00	0.82	0.23	0.91	0.00	0.85	-0.35
SGP															1.00	1.00	0.76	0.11	0.71	0.33
LM																	1.00	1.00	0.87	-0.29
GTS																			1.00	1.00

1) Represent the correlations among school effect estimates

Table 10:

Correlations[1] among models - Middle School Math

	Status		Gain		FEG		CAFE		TSG		CARE		Panel		SGP		LM		GTS	
	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
Status	1.00	1.00	0.35	-0.40	0.54	-0.50	0.87	0.26	0.22	-0.06	0.42	0.12	0.38	0.00	0.66	0.22	0.41	-0.22	1.00	0.33
Gain			1.00	1.00	0.94	0.76	0.51	-0.54	0.65	0.33	0.73	0.33	0.50	0.14	0.80	0.14	0.60	-0.02	0.32	-0.41
FEG					1.00	1.00	0.51	-0.58	0.71	0.36	0.77	0.40	0.51	0.02	0.81	0.00	0.79	0.31	0.36	-0.50
CAFE							1.00	1.00	0.62	-0.03	0.57	0.18	0.76	0.02	0.73	0.48	0.97	-0.39	0.97	0.70
TSG									1.00	1.00	0.93	0.76	0.83	0.46	0.82	0.47	0.66	0.28	0.55	-0.09
CARE											1.00	1.00	0.89	0.49	0.87	0.53	0.92	0.50	0.47	0.13
Panel													1.00	1.00	0.85	0.53	0.78	0.24	0.67	0.00
SGP															1.00	1.00	0.84	0.29	0.62	0.33
LM																	1.00	1.00	0.87	-0.22
GTS																			1.00	1.00

1) Represent the correlations among school effect estimates

Mixed effects models and conditional growth models tend to be fairly highly correlated as well. In general the potential for different models to provide relatively similar results is high, as evidenced by the maximum correlations presented.

Overall, the patterns are relatively consistent across subject and school level. The patterns presented in Tables 7 through 10 vary by state, which further demonstrates that models do not unequivocally perform well under all circumstances. This variability is summarized by comparing the maximum and minimum correlations.

Perhaps a more informative and practical approach is to examine how the different models classify schools. Correlations imply a linear relationship between models (and some, like the SGP may produce estimates that are not linear with respect to another model's school effect estimates, for example). We next turn to using performance quintiles. This division is somewhat arbitrary, but consistent with much of the literature attempting to place schools in performance bands. Table 11 consists of several parts that consider selected model comparisons using mathematics (reading does not substantively alter interpretations). In Table 11, *Exact* refers to the percent of time two models place schools into the same performance band (quintile). So in Panel A, which focuses on gain models, a simple gain model and percent proficient model would place schools in the same performance band 26% of the time. These two models would be off by one performance category about 32% percent of the time. Hence, percent proficient and a simple gain model would place schools within one performance band about 58% of the time. In 22% of cases, the classifications would be off by two bands, and 6% of the time, the two models would place the schools on opposite extremes (best and worst). The results in Table 11 indicate that simple gains and fixed effects gains (FEG) produce very similar placements (as expected), but simple gains and growth to target (GTS) do not. Panel B focuses on conditional change models. The CARE and SGP models have highly correlated results, while the CAFE model tends to classify schools consistently with the others, but the classifications occur at a somewhat lower rate. Panel C focuses on growth models. All of the models tend to classify schools similarly except for the growth to standard (GTS) model which behaves more like a status model.

Table 11. Comparisons of Classification into Mathematics Quintiles

Panel A	**Status (% prof)** compared to	**Fixed Effects Gain** compared to					
	GAIN	**GAIN**		**TSG**		**GTS**	
		ES	MS	ES	MS	ES	MS
Exact	26%	94%	74%	49%	36%	26%	28%
within 1	32%	5%	20%	28%	44%	33%	30%
Exact +1	58%	99%	94%	77%	80%	58%	57%
within 2	22%	1%	5%	13%	12%	22%	21%
Extremes	6%	0%	0%	2%	3%	6%	8%

Panel B	**Student Growth Percentile** compared to				**Layered Model** compared to			
	CARE		**CAFE**		**SGP**		**CAFE**	
	ES	MS	ES	MS	ES	MS	ES	MS
Exact	55%	49%	33%	37%	52%	48%	26%	29%
within 1	40%	43%	37%	35%	38%	36%	36%	31%
Exact +1	95%	91%	70%	72%	90%	84%	62%	60%
within 2	4%	8%	20%	17%	7%	11%	22%	21%
Extremes	0%	0%	2%	2%	0%	1%	4%	5%

Panel C	**Simple Panel Growth** compared to							
	LM		**SGP**		**GTS**		**TSG**	
	ES	MS	ES	MS	ES	MS	ES	MS
Exact	38%	36%	47%	41%	26%	20%	40%	38%
within 1	39%	39%	40%	41%	35%	35%	37%	38%
Exact +1	76%	75%	87%	82%	61%	56%	77%	77%
within 2	18%	17%	11%	13%	23%	22%	18%	15%
Extremes	1%	2%	0%	1%	4%	6%	1%	1%

Overall, models from which similar inferences can be made are more likely to place schools into similar performance bands than models that differ fundamentally, which is most readily apparent when comparing gains and percent proficient. The fundamental basis differs, and not surprisingly, they rate schools very differently. Although models place schools into different performance quintiles, there is a reasonable amount of consistency, and except for status, none of the models demonstrates extreme variations in school ratings. That is, it is unlikely that one model would rate a school as a top performer (Grade of A), while another model would rate the same school as a poor or very poor performer (Grade of D or F). Variations among school ratings by model generally follow expectations. Simple Gain produces results that are somewhat different from other growth models, except for Gain and FEG models, which are virtually identical. Conditional status models SGP and CARE, both of which are current performance based on previous performance, perform similarly. Panel model results are consistent with SGP, LM, and TSG implying that as more data are used to estimate effects, results tend to converge.

How accurately do models classify schools into performance categories?

The previous section examined how similarly models placed schools into performance bands. We believe this is an important element to focus on and leads us to further examine how accurately models place schools into those bands. This depends on precision. The results in Tables 12a and 12b indicate that there is considerable variation among states and models in terms of how well they place schools into performance bands. Part of the variability has to do with the models and being able to differentiate schools based on the underlying attribute (e.g., status, gains, and growth) and partly it relates to properties of the schools in terms of the ICC and size. For example, the SGP model generally performs quite well with accuracy nearing 100% (denoted as 100% in the table due to rounding); it should be noted that we would expect the SGP model to place schools into performance bands with high accuracy since the bands are based on quintiles — which is somewhat tautological for the SGP model. Still, in states with smaller average school size, the SGP is only 60% likely to place a school in the correct band. Simple Gain models tend to perform poorly and the TSG model appears to rely on vertical scales that are well suited for this type of model. Also, it is clear that the layered model (LM) does not perform well here and this is likely, at least partially, due to the limited amount of data we used to estimate a model that is designed to incorporate substantially more data. Clearly, however, Tables 12a and 12b again imply that models behave very differently in different states and wholesale adoption does not guarantee success.

Table 12a. Proportion of Estimates Likely to Fall into Performance Correct Band*

	Elementary School								Middle School							
	ELA				Math				ELA				Math			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Status	60	54	57	60	58	38	55	60	60	54	57	60	58	38	55	60
	45	43	44	39	48	51	29	39	45	43	44	39	48	51	29	39
	40	15	27	30	45	26	23	28	40	15	27	30	45	26	23	28
	31	12	34	30	35	20	18	27	31	12	34	30	35	20	18	27
	42	19	53	45	48	26	59	32	42	19	53	45	48	26	59	32
GAIN	64	66	48	54	61	63	57	56	64	66	48	54	61	63	57	56
	25	20	22	18	24	31	16	30	25	20	22	18	24	31	16	30
	27	20	18	15	31	25	14	26	27	20	18	15	31	25	14	26
	36	22	24	18	39	26	15	31	36	22	24	18	39	26	15	31
	56	53	51	54	55	59	65	57	56	53	51	54	55	59	65	57
FEG	63	70	48	55	57	69	42	54	61	70	60	61	64	71	36	59
	20	20	18	21	20	26	15	25	27	21	12	25	19	16	17	12
	20	23	16	17	22	23	18	22	29	13	11	20	27	10	17	12
	27	22	22	20	42	24	18	27	45	18	16	25	39	23	12	14
	53	56	54	54	54	59	61	56	65	68	57	57	65	72	25	59
TSG	98	99	78		99	100	81		100	75	95		89	90	78	
	70	65	47		78	76	41		49	31	44		46	48	31	
	50	62	39		72	58	31		44	47	39		48	57	20	
	52	49	63		66	77	30		50	57	29		76	80	33	
	94	95	94		100	100	78		84	74	94		100	85	72	
CAFE	65	87	64	65	63	72	61	65	66	73	87	81	67	62	75	82
	49	47	39	48	51	52	33	49	46	67	42	51	54	62	27	51
	50	48	37	44	48	46	26	47	46	46	40	41	45	55	20	44
	39	35	42	44	41	43	25	49	36	49	30	42	40	59	23	47
	58	69	61	62	59	66	62	63	60	85	88	63	59	75	75	67
CARE	49	55	38	44	52	56	42	53	48	39	39	41	52	45	39	54
	13	23	16	19	18	24	14	29	12	20	17	18	15	16	14	26
	10	27	12	16	12	22	14	27	9	12	11	16	11	20	6	22
	14	22	18	19	16	29	12	31	11	19	12	18	22	26	14	32
	40	56	50	44	51	54	40	51	33	42	40	47	52	43	37	54
PANEL	100	97	85	62	100	100	81	76	100	91	64	53	100	78	92	71
	69	47	39	30	72	51	37	41	49	70	46	23	68	38	25	32
	67	49	37	28	59	69	35	34	41	48	38	23	47	72	29	35
	46	50	49	35	52	57	33	38	46	71	41	21	51	79	38	38
	90	100	86	74	99	98	73	77	91	90	64	53	91	63	87	78
LM	28	68	46	41	34	58	40	50	61	67	57	37	57	60	28	51
	5	36	19	15	10	15	12	28	18	60	28	13	17	48	13	26
	3	40	17	13	8	15	14	24	12	25	25	10	19	28	50	21
	5	33	18	15	11	17	14	30	23	33	33	13	19	38	35	28
	17	63	48	40	42	48	30	51	55	30	56	37	46	32	48	54
SGP	100	100	100	99	100	100	99	100	100	100	100	100	100	100	100	100
	77	82	80	77	89	90	69	90	77	92	75	76	90	79	60	91
	82	83	69	74	79	85	73	86	61	82	66	68	73	73	71	85
	61	83	74	78	83	87	73	87	82	79	80	74	91	90	69	91
	99	100	100	98	100	100	99	99	100	97	97	100	100	99	100	100
GTS	100	100	100	100	100	100	96	100	100	100	100	100	100	100	96	100
	95	94	76	85	96	97	69	84	95	94	76	85	96	97	69	84
	88	81	71	77	92	86	65	75	88	81	71	77	92	86	65	75
	84	73	72	75	83	81	69	73	84	73	72	75	83	81	69	73
	91	92	95	92	96	97	94	77	91	92	95	92	96	97	94	77

*Accuracy (High to Low) is color coded within cells on a scale ranging from "Red" (Low) to "Green" (High)

Table 12b. Proportion of Estimates Likely to Fall into Performance Correct Band*

	State 1				State 2				State 3				State 4			
	Elementary		Middle		Elementary		Middle		Elementary		Middle		Elementary		Middle	
	ELA	MATH	ELA	MATH	ELA	MATH	ELA	MATH	ELA	MATH	ELA	MATH	ELA	MATH	ELA	MATH
Status	60	58	60	58	54	38	54	38	57	55	57	55	60	60	60	0
	45	48	45	48	43	51	43	51	44	29	44	29	39	39	39	39
	40	45	40	45	15	26	15	26	27	23	27	23	30	28	30	28
	31	35	31	35	12	20	12	20	34	18	34	18	30	27	30	27
	42	48	42	48	19	26	19	26	53	59	53	59	45	32	45	32
GAIN	64	61	64	61	66	63	66	63	48	57	48	57	54	56	54	56
	25	24	25	24	20	31	20	31	22	16	22	16	18	30	18	30
	27	31	27	31	20	25	20	25	18	14	18	14	15	26	15	26
	36	39	36	39	22	26	22	26	24	15	24	15	18	31	18	31
	56	55	56	55	53	59	53	59	51	65	51	65	54	57	54	57
FEG	63	57	61	64	70	69	70	71	48	42	60	36	55	54	61	59
	20	20	27	19	20	26	21	16	18	15	12	17	21	25	25	12
	20	22	29	27	23	23	13	10	16	18	11	17	17	22	20	12
	27	42	45	39	22	24	18	23	22	18	16	12	20	27	25	14
	53	54	65	65	56	59	68	72	54	61	57	25	54	56	57	59
TSG	98	99	100	89	99	100	75	90	78	81	95	78				
	70	78	49	46	65	76	31	48	47	41	44	31				
	50	72	44	48	62	58	47	57	39	31	39	20				
	52	66	50	76	49	77	57	80	63	30	29	33				
	94	100	84	100	95	100	74	85	94	78	94	72				
CAFE	65	63	66	67	87	72	73	62	64	61	87	75	65	65	81	82
	49	51	46	54	47	52	67	62	39	33	42	27	48	49	51	51
	50	48	46	45	48	46	46	55	37	26	40	20	44	47	41	44
	39	41	36	40	35	43	49	59	42	25	30	23	44	49	42	47
	58	59	60	59	69	66	85	75	61	62	88	75	62	63	63	67
CARE	49	52	48	52	55	56	39	45	38	42	39	39	44	53	41	54
	13	18	12	15	23	24	20	16	16	14	17	14	19	29	18	26
	10	12	9	11	27	22	12	20	12	14	11	6	16	27	16	22
	14	16	11	22	22	29	19	26	18	12	12	14	19	31	18	32
	40	51	33	52	56	54	42	43	50	40	40	37	44	51	47	54
PANEL	100	100	100	100	97	100	91	78	85	81	64	92	62	76	53	71
	69	72	49	68	47	51	70	38	39	37	46	25	30	41	23	32
	67	59	41	47	49	69	48	72	37	35	38	29	28	34	23	35
	46	52	46	51	50	57	71	79	49	33	41	38	35	38	21	38
	90	99	91	91	100	98	90	63	86	73	64	87	74	77	53	78
LM	28	34	61	57	68	58	67	60	46	40	57	28	41	50	37	51
	5	10	18	17	36	15	60	48	19	12	28	13	15	28	13	26
	3	8	12	19	40	15	25	28	17	14	25	50	13	24	10	21
	5	11	23	19	33	17	33	38	18	14	33	35	15	30	13	28
	17	42	55	46	63	48	30	32	48	30	56	48	40	51	37	54
SGP	100	100	100	100	100	100	100	100	100	99	100	100	99	100	100	100
	77	89	77	90	82	90	92	79	80	69	75	60	77	90	76	91
	82	79	61	73	83	85	82	73	69	73	66	71	74	86	68	85
	61	83	82	91	83	87	79	90	74	73	80	69	78	87	74	91
	99	100	100	100	100	100	97	99	100	99	97	100	98	99	100	100
GTS	100	100	100	100	100	100	100	100	100	96	100	96	100	100	100	100
	95	96	95	96	94	97	94	97	76	69	76	69	85	84	85	84
	88	92	88	92	81	86	81	86	71	65	71	65	77	75	77	75
	84	83	84	83	73	81	73	81	72	69	72	69	75	73	75	73
	91	96	91	96	92	97	92	97	95	94	95	94	92	77	92	77

*Accuracy (High to Low) is color coded within cells on a scale ranging from “Red” (Low) to “Green” (High)

Overall, the results in Tables 12a and 12b are consistent with previous results we report — models vary in how well they place schools into the performance bands. The SGP model appears to be the standard, but that result is somewhat tautological since performance bands are based on quintiles. The forgoing analysis highlights an important issue: inference about school performance is very much dependent upon how performance bands are defined and any model's ability to consistently place schools into defined bands. Figure 2 provides a sense of the performance that is required for a school to be placed into a performance band. The GTS model has fairly linear equally spaced bands and it is clear from Figure 2 how performance varies among states. This linearity may be a desirable property for some policymakers.

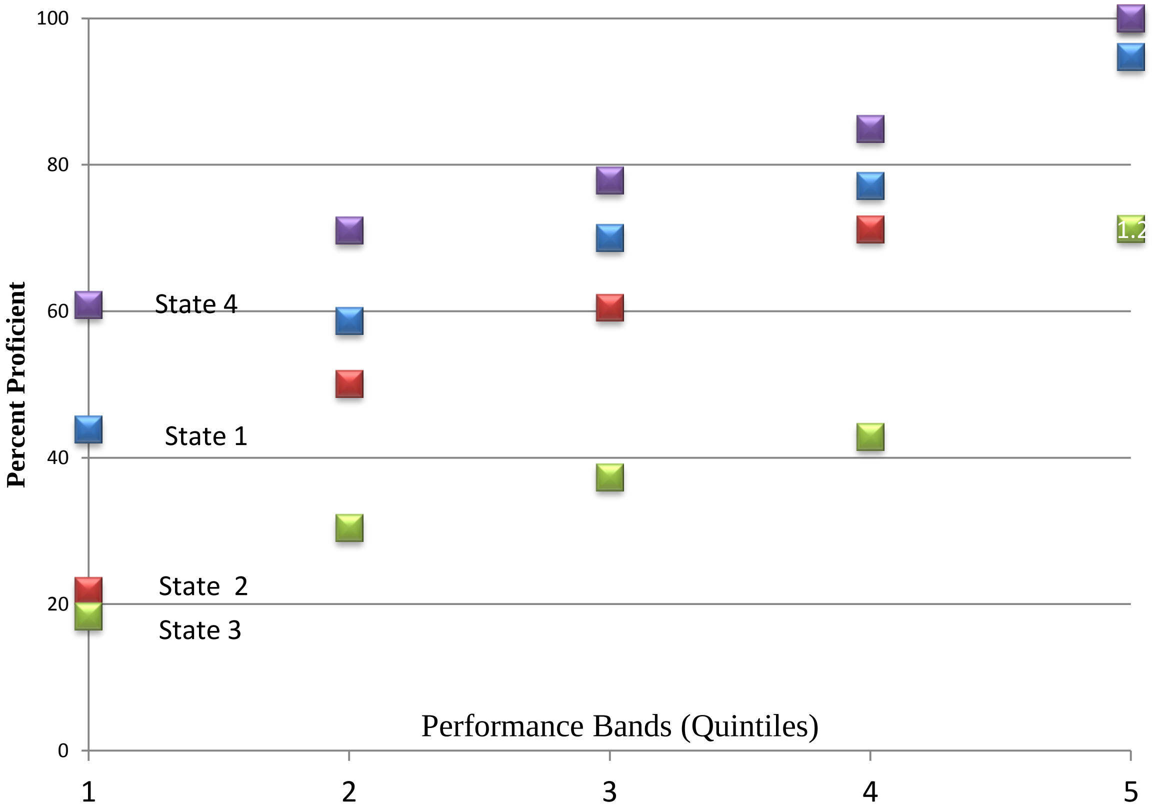

Figure 2: Performance Bands: Growth to Standard Model by State (Mathematics)

In contrast, Figure 3 presents the performance bands for the CARE model. It is important to note that all of the states converge on the middle performance band, as the estimated effect at the median is 0[24]. However, the bands are somewhat non-linear in both directions toward the extremes (toward very good and very poor performers).

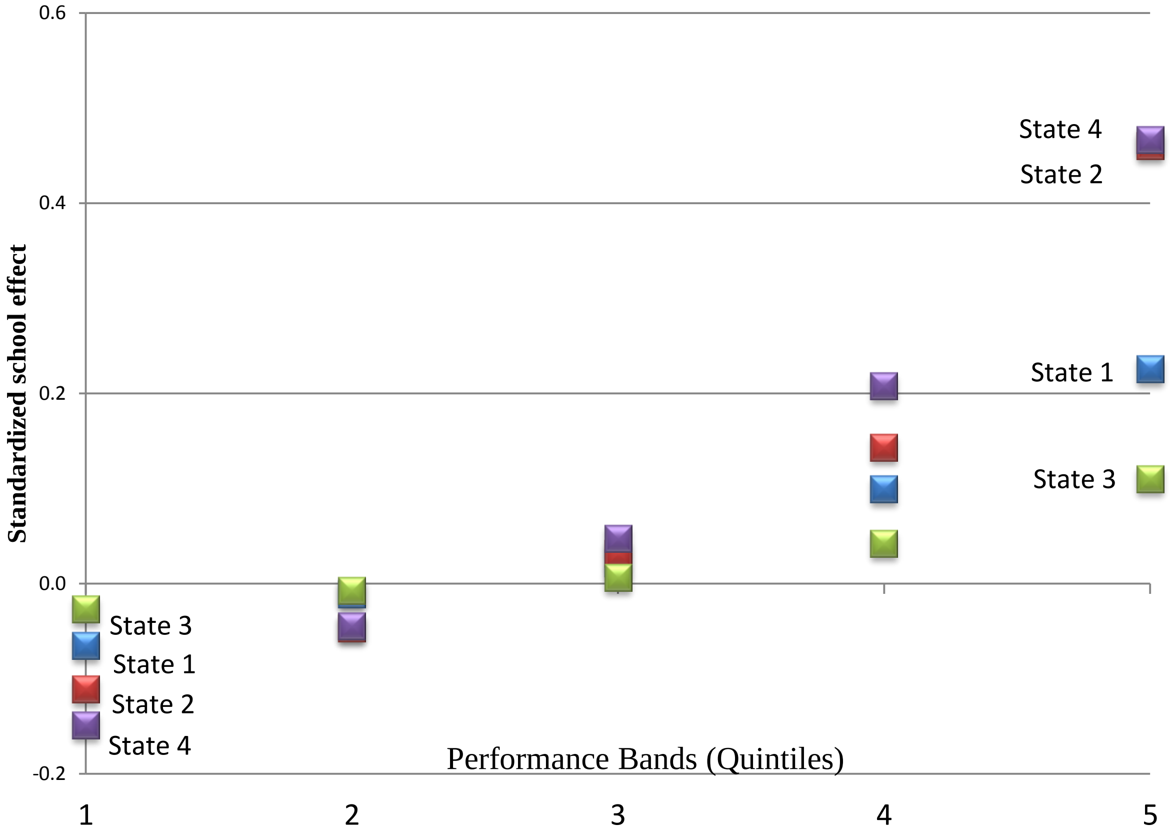

Figure 3: Performance Bands: Covariate Adjusted Random Effects Model (Mathematics)

Again, the properties around a CARE model may be what policymakers hope to accomplish — good schools receive positive values and poor performing schools receive negative values. For schools to receive top ratings, for example, policymakers may want exponentially higher performance to denote those bands — it is important to reiterate that the bands are based on the performance quintiles based on the model estimates.

Are models consistent in classifying schools from one year to the next?

We examine this question by considering our notion of reliability. The results are based on the shared variance of 2009 and 2010 school effect estimates. Another way to

[24] Value added models generally provide estimates that have a mean of 0.

conceptualize this notion is how well the 2009 school effect predicts the 2010 school effect. This notion is important given policymakers want to base decisions about schools on school effect estimates. A school's performance in one year, all else equal, ought to be a good indicator of its performance the following year. The results in Table 13a imply that these models are moderately consistent in estimating the school effects in subsequent years. Consistent with previous research, simple Gains are highly volatile from year to year.

Overall, these models provide a fair bit of information about subsequent performance, but the results should be a caution and provide further justification for rating schools on multiple indicators.

Table 13a
Reliability of School Effect Estimates

	Math ES	Reading ES	Math MS	Reading MS
Status (pct proficient)	0.49	0.44	0.49	0.44
Gain	0.06	0.08	0.06	0.08
FEG	0.21	0.14	0.14	0.08
CAFE	0.62	0.65	0.62	0.52
CAFE w/SBG*	0.62	0.63	0.61	0.51
TSG	0.18	0.26	0.26	0.09
CARE	0.27	0.19	0.28	0.06
Panel	0.37	0.37	0.19	0.21
SGP	0.21	0.10	0.21	0.11
LM	0.14	0.31	0.50	0.62
GTS	0.62	0.63	0.62	0.63

*SBG = student background characteristics

Another approach to considering the consistency of school effect estimates from one year to the next is to consider stability, which can be estimated with the correlations between each year's school estimates (Lockwood & McCaffrey, 2007). First, in Table 13b, we present some contextual information related to stability that indicates how the percent of students with a specific classification correlate from one year to the next. For example, overall, the percent of students who are economically disadvantaged (ED) at a school is correlated .85 from one year to the next. Another aspect to consider, not shown in Table 13b, is that the correlations of the percent of students who are proficient or above from one year the next varies from .25 to .90 (depending on state and subject). The results in Table 13c describe the correlations (stability) between the estimated school effects and are simply the square-roots of the reliability estimates in Table 13a. The results indicate that models that tend to be more closely related to end-status (CARE, CAFE, GTS) tend to be more stable. Consistent with expectations, simple gains are not stable, while more complex growth models (less aligned with status) are more stable than simple gains. In general, results are more stable for elementary schools than for middle schools (except for the layered model, which is more stable for middle schools).

Table 13b
Stability of Selected Background Characteristics

		2009	
2010	ED	SWD	ELL
ED	.85		
SWD		.81	
ELL			.78

Table 13c
Stability in School Effect Estimates

	Math ES	Reading ES	Math MS	Reading MS
Status (pct proficient)	0.70	0.66	0.70	0.66
Gain	0.24	0.28	0.24	0.28
FEG	0.46	0.37	0.37	0.28
CAFE	0.79	0.81	0.79	0.72
CAFE w/SBG[*]	0.79	0.79	0.78	0.71
TSG	0.42	0.51	0.51	0.30
CARE	0.52	0.44	0.53	0.24
Panel	0.61	0.61	0.44	0.46
SGP	0.46	0.32	0.46	0.33
LM	0.37	0.56	0.71	0.79
GTS	0.79	0.79	0.79	0.79

[*] SBG = student background characteristics

Overall, the models show moderate consistency in classifying schools from one year to the next. Of course, schools are in fact changing true performance as well, so we would not expect results to be perfectly correlated. Simple Gain models are the least consistent and the True Score Gain (TSG) model provides some additional stability. This can be seen quite readily in Figure 4, which plots school effects for the simple Gain model (left panel) and the TSG model (right panel). Taking measurement error into account substanitially improves a Gain model, but does so at the expense of tranparency and simplicity – however, the interpretation is consistent, and this may be sufficient for policymakers wishing to use gains as an indicator of growth.

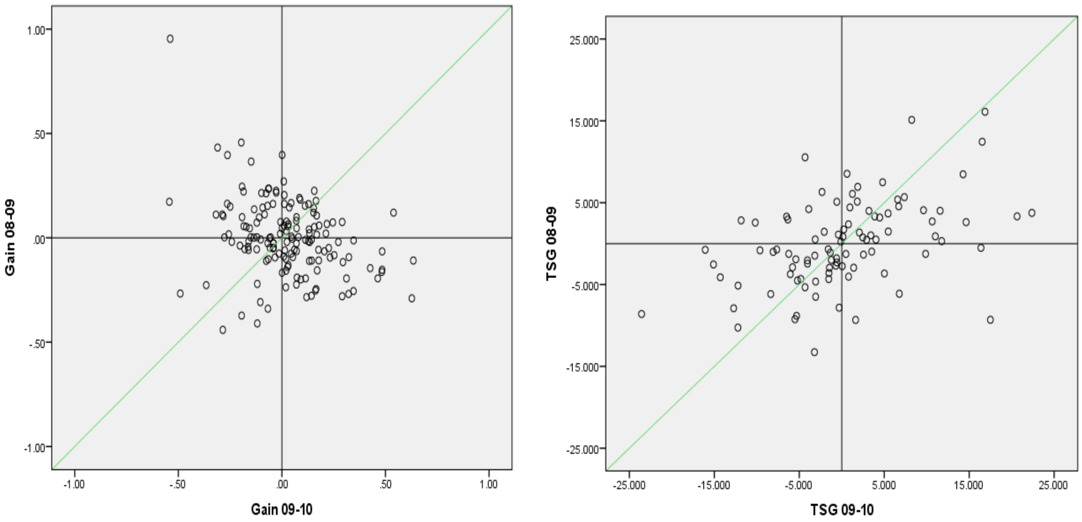

Figure 4: Stability for Gain and TSG Models Compared

Another common concern among policymakers is the impact of school size. Figure 5 highlights the impact of school size on four models. The green schools are schools with estimates based on 30 students or fewer. The results indicate school effects are substantially less stable when schools are small using Gain and SGP models than when using CARE or Panel models. School effects based on the Panel model are substanitvely unaffected down to school sizes of 30.

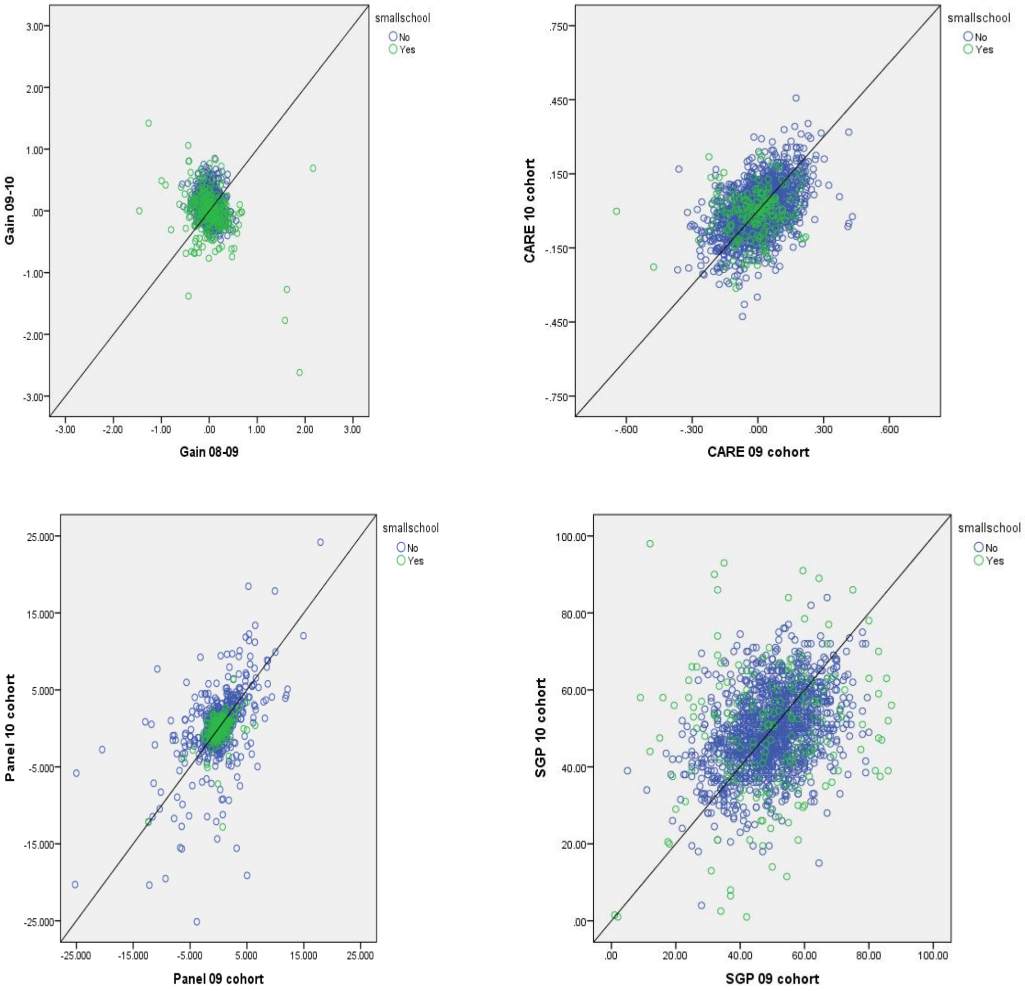

Figure 5: The Effect of School Size on Stability Estimates

How are models influenced by school intake characteristics?

We next turn to a critical aspect that provides some evidence related to bias. A model should afford students equal opportunity to succeed, and by extension, schools' accountability ranking should not be overly influenced by school intake characteristics – that is, student enrollment characteristics. Table 14 summarizes results based on the model presented above in eq. 18 that explicitly tests whether intake characteristics are related to school effect estimates in a linear or non-linear fashion.

The results indicate that there is variability in the influence of student intake characteristics on model results. Most notably, models most closely related to status tend to be most influenced by intake characteristics, and much of the stability observed with these models is based on the stability of school enrollment.

Table 14
Effect of Student Background on School Effects

	Significant Effects (L=Linear; N=Non-Linear)					
	SWD %	Minority %	ED %	ELL %	Stability %	School N
Status (% proficient)	L N	L N	L N	L N	L N	L N
GAIN	L	L	L	L	L	
FEG			L	L	L	
TSG		L	L	L	N	L
CAFE	N			L	L	
CAFE *	N			L	L	
CARE	N		L	L	L	
PANEL				L	L	
LM				L	L	
SGP		N				
GTS	L N	L N	L	L N	L N	L N

*Fixed Effect Covariate Adjusted with Student Background Characteristics

Overall, status and growth models correlated to status (e.g., GTS) are more susceptible to influence from intake characteristics beyond school control. Consistent with expectations, models that use more prior information (Panel, LM, and SGP) are least influenced by intake characteristics beyond school control. Simple Gain and the TSG models tend to be related to several intake characteristics. Models that are not simple aggregates (e.g., Gain, GTS, and Status) can incorporate intake characteristics explicitly, thus potentially reducing the effects of intake characteristics. This requires a tradeoff between statistics and policy. It is also important to note that models behave differently among the states, so these results, when disaggregated by state vary.

Another aspect that should be considered for practical applications of these growth models is the extent to which the precision is related to school size. Figure 6 plots the correlations of school size against the estimated standard errors for school estimates.

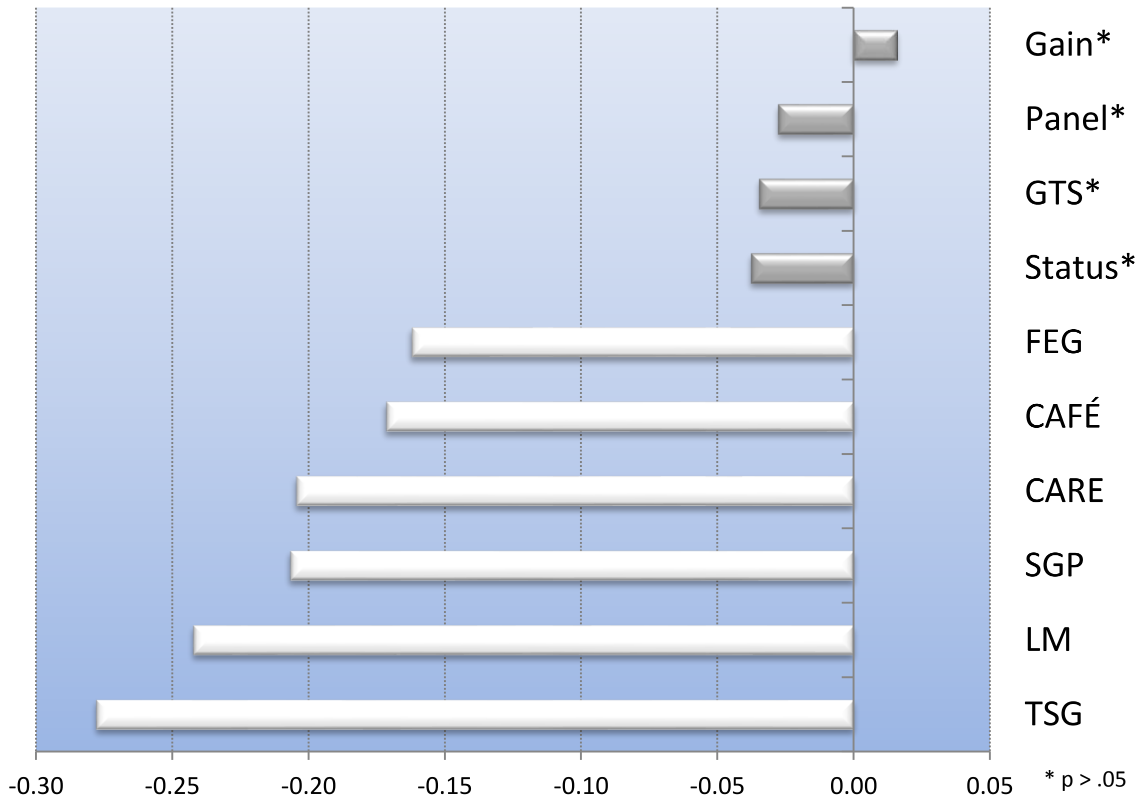

Figure 6: Correlations Between the Standard Error and School Size

The results from the four states indicate that the standard errors associated with school effect estimates are related to school size except for Gain, Panel, GTS, and Status models.. Again, this points to considering the context in which the models will be applied.

Do models perform similarly for elementary and middle schools?

Overall, the models we examined in this study tend to work relatively similarly in elementary school and in middle school. Middle school models are sometimes less accurate and less stable. Growth models, in general, appear to be fairly robust to school organization. Continued decisions related to K-2, single grade schools, or untested grade school (e.g., grade 9, 10) will be necessary for attribution purposes.

Do models behave similarly across states?

This question has been addressed throughout this study. The meta-analytic model provided the general test of state effects and those results indicated that models perform differently among the states. There can be 40 point swings in reliability and large swings in precision as well for a model across states. Part of the difference is due to differences in schools among the states, and partially this is due to other factors, such as the psychometric properties of the test or testing rules. Ultimately, model results vary in substantively important ways and preclude any notion of a single model meeting all the needs (just as assessment cannot possibly support the myriad of inferences desired by policymakers). It is clear however, that models based on growth, especially when using as much data as possible, provide the soundest indicators of school performance.

This notion is clearly presented in Figure 7. The top panel of Figure 7 highlights how a model may provide very good differentiation among schools (in this case performance bands) in one state, but not another. The bottom panel addresses some of the cause of the variability in model performance among states, by disaggregating performance plots by school type. Clearly, the ability of a model to distinguish school performance depends on the context. For example the GTS model does quite well with advantaged schools and substantially less well with disadvantaged schools.

We emphasize that no model will fit all situations and models have intended uses that should be considered before their application. Even after (or during) model selection,

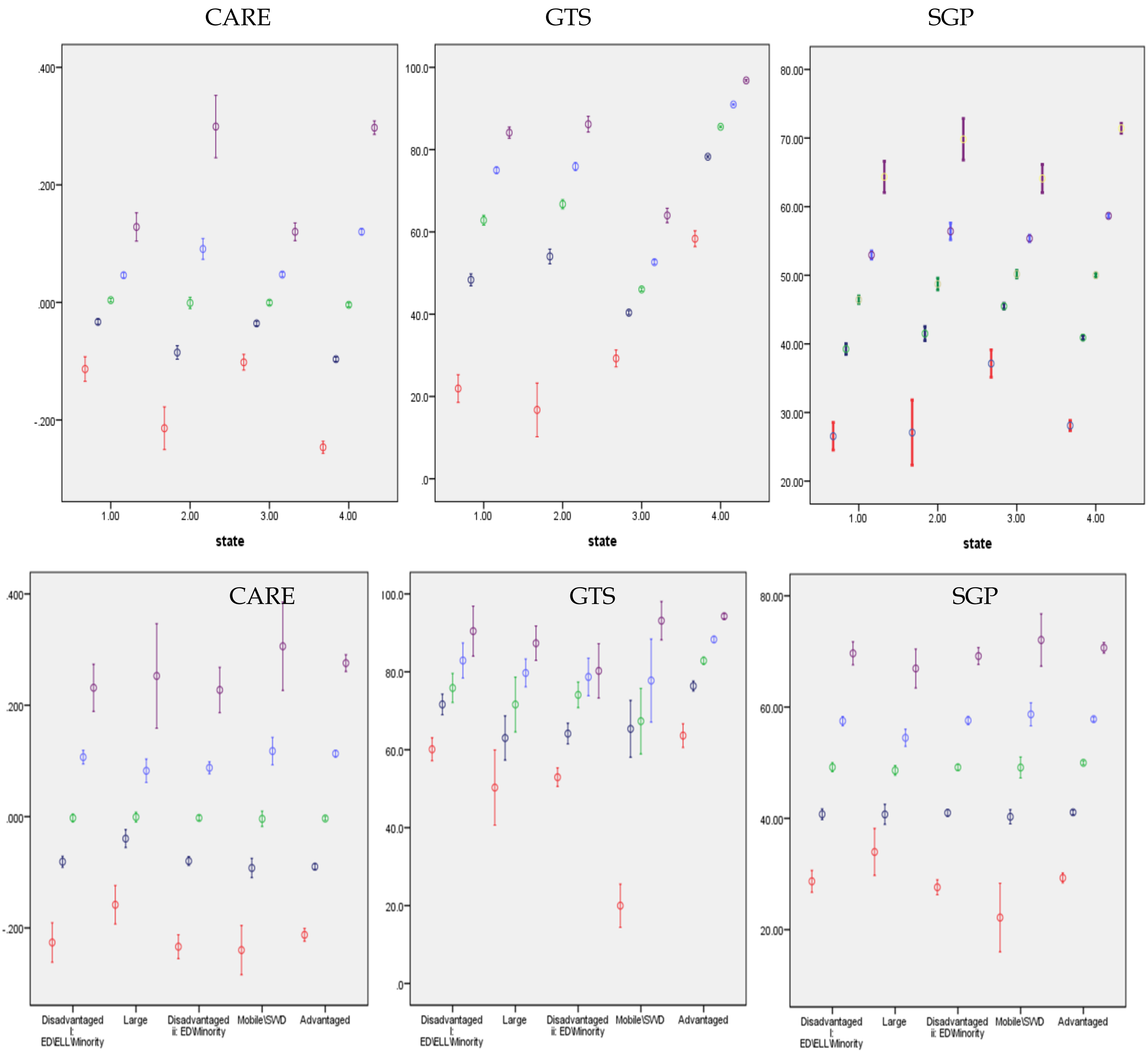

Figure 7: How Models Distinguish Among Schools by State and School Type Model

analyses such as those carried out here should be considered to evaluate whether models produce results that will be amenable for inferences in high stakes environments. We detail all of the models for each state and school type in appendix four and five.

SUMMARY/DISCUSSION

The results of these analyses confirm that no single model can unequivocally be assumed to provide the best results. This is not possible for two reasons: one, different models address different questions about schools; and two, the empirical results indicate that context matters when examining models. By context we mean that the state in which the model will be run affects how the model may work. State affects include several pieces that are confounded. These include tests scales, testing procedures, student characteristics, and school characteristics. An accountability model should not be unduly influenced by factors outside of schools' control and models clearly differ in this respect. Distinguishing between a school's ability to facilitate learning and a school's performance as a function of advantageous (or challenging) student enrollment characteristics is where statistical machinery provides its biggest benefit. Models that condition on additional information (student background or prior performance, for example), such as Panel, SGP, LM, and CARE or CAFE models with multiple prior assessments) are clearly better able to attribute student learning to schools. Of course the tradeoff is technical complexity and true understanding of the inferences afforded by model results. We briefly consider and summarize the research questions below.

We begin with whether, overall, the model matters. In addressing this question we first note that the models we examined were all placed into an existing typology. In general models within typology category allow for similar inferences, but may be estimated differently. But we note that this is not exclusively so. For example, covariate adjustment models and student growth percentile models are similar in terms of inference about schools, but are different in terms of how they are estimated. The layered model is estimated very differently than a TSG or Panel model, but the inferences are similar.

The results in Table 6 imply that school effect estimates do, in fact, vary by model in elementary school ($p < .01$) and in middle school ($p < .01$). Moreover, the results imply that estimated school effects vary by model by state, which means that different models generate school effects differently in each state. The joint effects related to model by subject and model by cohort imply that individual school effect estimates vary by model and subject and cohort. The results in Table 6 also indicate that similar schools within a cluster will be rated differently by different models.

The results in Table 6 confirm that there are differences in school effect estimates and that these differences depend on both school context and the model used to estimate the effect. We rejected the null hypothesis that models are equal, statistically. Consistent with expectations, models within typology categories provide more similar results than models between categories.

We next consider whether different models lead to different inferences about schools. That is, would we rate a school as an A under one model, but a C under another model? Again, models from which similar inferences can be made (generally models in the same typology classification, but also including the exceptions noted above) are more likely to place schools into similar performance bands than models that differ fundamentally in their inferential intent. This is most readily apparent when comparing gains and percent proficient.

We also addressed how accurately models classify schools into performance categories. Models vary in accuracy and this depends largely on school size, test scale, and potentially the ICC.ICCs provide information about whether we can distinguishes differences in true school performance. We presented ICCs for status, but each approach (model) needs to examine how much variability there is between schools. For example, we may be able to estimate individual student growth precisely, but if, students all demonstrated very similar growth rates, it would be difficult to differentiate schools based on growth.

Whether models are consistent in classifying schools from one year to the next is an important question as unstable results would lead to less credibility. Most straightforward is the correlation between results from one year to the next as an indicator of stability. Stability varies by model and school level, and to a lesser extent, subject. Models that are conditional status models (e.g., SGP covariate adjustment models) tend to be more stable than gain based models (which is consistent with prior research) such as gains, true score gains (although the TSG model provides substantial improvement over the simple Gain model), and the LM model. Elementary school results tend to be more stable than middle school results. Also, results based on mathematics tend to be slightly more stable than results based on ELA. Again we note that to the extent that errors are uncounted for (e.g., sampling and equating error [Phillips, Doorey, Forgione, & Monfils, 2011]), this likely adds to instability.

We also examined whether model results are influenced by school intake characteristics (e.g., percent ELL and FRL). We consider these results to be important as this provides some evidence related to bias and fairness. To the extent that uncontrollable school input characteristics influence results, the model results need to be carefully examined. The results indicate that there is variability in the influence of intake on model results. Most notably, models most closely related to status (e.g., GTS) tend to be most influenced by intake characteristics, and much of the stability observed with these models is based on the stability of school enrollment. Given that we excluded student background in all but one specification, models that incorporated multiple assessment results (SGP, Panel, and LM) were least related to student background. These results are consistent with expectations.

Next we examine whether models perform similarly for elementary and middle schools. Models appear to behave similarly in middle and elementary schools although they sometimes are less accurate and less stable for middle schools. This also may be a function of the ICC, which tends to be larger in middle schools.

Finally, we consider whether models behave similarly across states. The models will perform differently across states. This result is associated with several factors. One, states have different assessments and how models function given the characteristics of the assessment cannot be completely predicted a priori. One obvious difference among states is the scale. One major difference is whether a state has a vertical scale. This naturally impacts gain-based results more than covariate adjustment model or SGP model results. Panel models allow perhaps the most concrete inference and are an improvement over gain models, but scale is important for interpretation beyond school rankings. Scaling decisions can vary within a vertical or non-vertical scale and these certainly may impact results as well, as has been previously demonstrated. States differ on testing and accountability policies. For example, one state allows students to retest for AYP purposes. This seems to have the effect of stabilizing results. States have different types of schools. We created similar school classifications across states[25] and find that models work differently for different types of schools and that this effect varies by state.

Simply importing a model to use for school accountability is not a good policy option. It is important to know the nature of the data. For example, is there a vertical scale? If so, then models such as the CARE, CAFE, SGP, LM, TSG, and Panel can be applied. If there

[25] These classifications were based on student characteristics and assume that classifications rules were the same across the states, which they likely are not. This further contributes to between state variability in model behavior.

is no vertical scale CARE, SGP, and Panel models might be the best options. However, variations in how scales are developed and how grade level assessments are equated from year to year are important in that unaccounted for equating error creates more ambiguity and potentially adds variability to results among state results. What is the ICC, that is, how much variability in growth or change can be attributed to schools? For example, the policy of allowing students multiple attempts on an assessment in each grade likely stabilizes results, but likely also reduces the variability between schools in growth — making it more difficult to attribute variation in performance to schools based on growth. Also, how do models react to potential ceiling or floor effects? A major concern is the types of schools a state has — these school characteristics and size impact how well models provide results. Again, this requires some preliminary examinations to ensure the results provide for tenable inferences about schools. Also, decisions need to be made with respect to performance bands — goals and the types of errors in school classifications policymakers want to avoid. Decisions related to stability and how to increase stability (e.g., averaging model results over time and using a model that includes multiple years) need to consider the tradeoff between having stable results and a system that is insensitive to true changes in performance.

Ultimately, we would expect the models within categories identified by the CCSSO growth model typology to perform similarly. However, as we noted, the data in Table 5 show that models with the same inferential intent may be estimated differently and this results in within-category variability. Policymakers must first decide what they believe constitutes school performance and how schools ought to be held accountable. Given a notion of how growth should be conceived (and what role it may take in an accountability system) models can be chosen. However, this will likely need to be an iterative process that should be partially dependent on the empirical results, as well as how the growth model works with other aspects of the accountability system.

For example, a school accountability model that aims to use both growth and status needs to carefully check the properties of each and how they work in conjunction to identify school performance. A model that uses both growth and status may unintentionally disadvantage schools that have either very high or very low performance; or in some cases the growth pieces may actually counterbalance the status element – resulting with many schools being deemed average, few schools being exemplary or failing[26,] and importantly, creating a system through which it is virtually impossible to move out of the average range.

[26] Or, for example, when using an A-F system that is being widely adopted, most schools will be a C, with few A's or F's.

REFERENCES

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

Allison, P. (2009) Fixed Effects Regression Models, Series: Quantitative Applications in the Social Sciences, Sage, Newbury Park.

Baler, E., Goldschmidt, P., Martinez, F., & Swigert, S. (2002). In search of school quality and accountability: Moving beyond the California Academic Performance Index (API). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST): U.S. Department of Education Office of Educational Research and Improvement.

Ballou, D. (2009) Test Scaling and Value-Added Measurement, *Education Finance and Policy* 4(4), 351-383.

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. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Briggs, D., & Weeks, J. (2011). The persistence of school-level value added. *Journal of Educational and Behavioral Statistics, 36*(5), 616-637.

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

Bryk, A.S., & Weisberg, H.I. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin, 84*(5), 950-962.

Bryk, A.S., Thum, Y.M., Easton, J.Q., & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education*, *2*, 103-142.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, *4,* 158-233.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Castellano, K., & Ho, A. (In Press). *A Practitioner's Guide to Growth Models*. Wash., DC: Council of Chief State School Officers.

Choi, K. (2007). Teacher effect change model: Latent variable regression 5-level hierarchical model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Choi, K., Goldschmidt, P., & Yamashiro, K. (2005) Exploring models of school performance: From theory to practice. *Yearbook of the National Society for the Study of Education*, *104*(2), 119-146.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2005). *Comparing like with like: The role of student and school characteristics in value-added models.* Paper presented at AERA, Montreal, Canada.

Counsel of Chief State School Officers (2011) Achievement Growth and Accountability: What to Look For – What to Look Out For, Washington DC.

Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Goldschmidt, Choi, K.C., Martinez, F., & Novak, J. (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment, *School Effectiveness and School Improvement*, 21(3), 337-357.

Goldschmidt, P., F. Tseng, & D. Goldhaber (2010). An Assessment of the Implications of Various Statistical Approaches to Measuring Teacher Effects, White Paper for the Bill and Melinda Gates Foundation, Seattle, WA.

Goldschmidt, P., Choi, K., Boscardin, C., Yamashiro, K., Martinez, J.F., & Auty, W. (2006). *Extending common methodological approaches to school accountability and evaluation: Latent class and longitudinal models*. Presented at AERA, San Francisco, CA

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). *Policymakers guide to growth models for school accountability: How do accountability models differ?* Washington, DC: Council of Chief State School Officers.

Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A, 159*, 384-443.

Hambleton, R. K., & Swaminathan, H. (1987). *Item Response Theory: Principles and applications*, Boston, MA: Kluwer.

Hanushek, E. A., & Raymond, M. E. (2003). Lessons about the Design of State Accountability Systems. In Peterson, P. E. & Martin R. W. (Eds.), *No Child Left Behind? The Politics and Practice of Accountability* (pp. 126-151). Washington, DC: The Brookings Institution.

Hanushek, E.A., Rivkin, S., & Taylor, L. (1996). Aggregation and the estimated effects of school. *The Review of Economics and Statistics*, *78*(4), 611-627.

Hanushek, Eric A. (1986, September). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 49(3), 1141-1177.

Hao L. and Naiman D. Q.(2007), Quantile Regression, Sage Publications, Thousand Oaks.

Herman, J., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koedel, C., & Betts, J. (2007). Re-Examining the Role of Teacher Quality in the Educational Production Function, National Center on Performance Incentives, Working Paper 2007-03, Vanderbilt Peabody College, TN.

Koenker, R, & Bassett. G. **(**1978)“Regression Quantiles.” *Econometrica.* January,46:1, pp. 33–50.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Linn, R.L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29-36.

Lockwood, J.R., & McCaffrey, D.F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, *1*, 223-252.

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

Marion, S.F. (2010). Constructing a validity argument for alternate assessments based on modified achievement standards. In M. Perie (Ed.), *Teaching and assessing low-achieving students with disabilities* (pp. 247-255). Baltimore, MD: Paul H. Brookes.

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*(1), 67-101.

McCaffrey, D., Sass, T., Lockwood, J. R., & Mihaly, K. (2009). The inter-temporal variability of teacher effects estimates. *Education Finance and Policy*, *4*(4), 572-606.

Meyer, R.H. (1997). Value-added indicators of school performance. In E.A. Hanushek & D.W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197-223). Washington, DC: National Academic Press.

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

Novak, J., & Fuller, B. (2003). *Penalizing diverse schools? Similar test scores, but different students, bring federal sanction*. PACE Policy Brief. Berkeley, CA: Policy Analysis for California Education.

Phillips G., Doorey, N.A., Forgione, P.D., & Monfils L. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Washington, DC: Council of Chief State School Officers

R Development Core Team. (2009). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria. Available from http://www.R-project.org.

Raudenbush, S. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology, 52*, 501-525.

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* (1), 121-129.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S.W., Cheong, Y.F., & Fotiu, R. (1995). Synthesizing cross-national classroom effects data: Alternative models and methods. In M. Binkley, K. Rust, & Winglee, M. (Eds.), *Methodological issues in comparative international studies: The case of reading literacy*, Chapter 9, (pp 243-275). Washington, DC: National Center for Educational Statistics.

Raudenbush, S. & Willms, D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307-335.

Rogosa, D. *(*1995). Myths and methods: Myths about longitudinal research plus supplemental questions. In John Gottman (Ed.), *The Analysis of Change*, (pp 3-39). Mahwah, NJ: Lawrence Erlbaum.

Rogosa, D. (2002). Irrelevance of reliability coefficients to accountability systems: Statistical disconnect in Kane-Staiger *Volatility in School Test Scores*, Stanford University. Retrieved January 23, 2012, from http://npe.educationnews.org/Review/Resources/StatisticalDisconnect.pdf.

Rogosa, D.R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726-774.

Rogosa, D.R., & Willett, J.B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *20*, 335-343.

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.

Seltzer, M., Frank, K., & Bryk, A. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to the choice of metric. *Educational Evaluation and Policy Analysis*, *16*, 41-49.

SPSS. (2001). *The SPSS TwoStep cluster component*. White paper. Retrieved January 23, 2012, from http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf.

Tekwe, C., Carter, R., Ma, C-X., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. (2004). An empirical comparison of statistical models for value added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *20*(1), 11-36.

Thum, Y.M. (2003). Measuring progress toward a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods and Research*, *32*(2), 153-207.

Wheeler, D.J., & Lyday, R.W. (1989). *Evaluating the measurement process*. 2d ed. Knoxville, TN: SPC Press.

Willett, J.B., Singer, J.D., & Martin, N.C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395-426.

Willms, D., & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, *26*(3), 209-232.

Wright, S.P. (2010). *An investigation of two nonparametric regression models for value-added assessment in education*. Retrieved January 23, 2012, from http://www.sas.com/resources/whitepaper/wp_16975.pdf.

Wright, S. P. (2008). *Estimating educational effects using analysis of covariance with measurement error*. Paper presented at CREATE/NEI Conference, Wilmington, NC, October 2008. Retrieved December, 24, 2009, from http://www.createconference.org/documents/archive/2008/2008wright.pdf.

Yen, W.M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299-325.

APPENDIX 1: DATA QUALITY

The data quality approach for this study was comprised of five components. Each component was sequentially dependent and required minor variations to address the four data sets (DE, HI, NC, and WI) used in the student data file structure. The components were to

- normalize the raw files using common data elements;
- establish longitudinal structures via the unique student identifier (USI), unique school identifier (USchID), and unique district identifier (UDI);
- develop stored structured query language (SQL) procedures that created subject-specific, multi-year tables;
- screen data ranges within each element; and,
- produce descriptive data for each state.

The aforementioned components were subsumed into a standardized, production cycle that ensured data from each state were prepared in the same manner. This approach eliminated the possibility of introducing unwanted variance within the targeted outputs. Variance introduced by the production cycle increases the likelihood of confounding any observed difference in the dependent (output) variables. Further, although the data quality procedures strive toward a zero defect (Wheeler, 1989 product, the likelihood of obtaining such a high standard in educational data is unlikely at this time.

The production cycle began with the unprocessed data being received from each state education agency (SEA) (WI, DE, and HI) then subsequently converted from its original file transfer structure into a delimited format. Data element descriptions were reviewed and aligned into year-specific tables. A standardized nomenclature and codification was developed and applied to each data element. Further elements were also repositioned within each table so as to be consistent from table to table and SEA to SEA. Extraneous variables were limited from the tables. Normalization also included the enforcement of the USI representing a single student with valid performance data. Duplicated students within schools create erroneous n-counts for the school and subpopulations. Duplicated student records that attribute the same performance to multiple entities can significantly influence overall results for units of analysis with small n-counts. To resolve duplicated records from the operational tables, the following decision logic was applied:

Process Rule 1 IF a duplicated record had no valid performance values in RDGPL and/or MTHPL, AND the school ID was equal, THEN only one record was selected.

Process Rule 2 IF a duplicated record had no valid performance values in RDGPL and/or MTHPL, AND the school ID was not equal, THEN select the record with FAY = 1.

Process Rule 3 IF a duplicated record had one set of valid performance values, THEN the code 9 (missing) record(s) was deleted.

Process Rule 4 IF a duplicated record had two or more sets of valid performance values in RDGPL and/or MTHPL, AND the school ID was equal, THEN only one record was selected [EQUAL RDGSS], ELSE select record with the highest RDGSS.

Process Rule 5 IF a duplicated record had two or more sets of valid performance values in RDGPL and/or MTHPL, AND the school ID was not equal, THEN select one record assigned to both schools only with FAY = 1, ELSE selected record with FAY = 1 was retained.

Process Rule 6 IF WI USI = 11, AND no SOR could be identified, THEN all records were deleted (but documented in an audit spreadsheet).

Process Rule 7 IF a duplicated record had two or more sets of valid performance values, AND the GRADE was not equal and other demographic data suggested two different student profiles, THEN USI for one student was replaced by the SYSTEM ID (but documented in an audit spreadsheet).

Process Rule 8 IF a duplicated record had two or more sets of valid performance values, AND the GRADE was not equal and other demographic data does not suggest two different student profiles, THEN the USI for one student was replaced by the SYSTEM ID (but documented in an audit spreadsheet).

The aforementioned process rules were applied to each data set prior to establishing the required longitudinal structures for the growth model analysis. Growth models (GMs) have at least one common characteristic inherent to their data requirements; at least two measures in time are needed. For some designs, such as projection-type GMs, multiple waves of data are needed that are linked by a primary key (student USI) and other foreign keys (district UDI and school USchID).

Given the population (N) for year (Y_x), several assumptions must be considered. First, the population is underrepresented by students that do not participate in the assessment, but were eligible to participate. Federal regulations (34 C.F.R. §200) stipulate a threshold of 95% participation be attained by schools and subpopulations within schools in order to demonstrate adequate yearly progress (AYP). The assumption that non-participants are randomly distributed in a given year and that no systematic exclusion, by policy or practices, can be detected within population must be evaluated. Second, the USI linking performance over time represents a single individual and that duplicated records are identified, removed, or recoded. Students with multiple, valid test results create minor errors in aggregated values (e.g., district and school); however, in aggregation units with small n-counts (e.g., Limited English Proficient [LEP], American Indian), the impact is magnified. Third, the population's attrition rate (inability to link records) increases systematically as a function of assessment grade restrictions, grade configurations, student grade retention, students exiting the system as either dropouts or transfer-outs, and data integrity over time. Finally, there is a magnitude of missing data points within data elements recoded as missing, left blank, or incorrectly coded. These within-element integrity issues may result in spurious outputs, particularly when they are specific to any given year. These "missing points" across the time continuum create unique challenges to the precision of regression and projection-based growth models.

Given the aforementioned considerations, this study constructed cohort groups across multiple combinations of years. Groups were established and "mapped" forward by developing right join relationships based on the USI. The d-base used structured query language to organize the data tables, screen contents for duplicate records, create multi-year tables, and report data. Because the SQL procedure was relatively simple, they were stored as one application. The use of this approach allows researchers the ability to quickly select targeted sections of the code and rerun data to accomplish ad hoc tasks such as identifying when a student left the cohort or detecting when the same student may have reentered at a later time.

The final data tables created from the stored SQL procedure were then manually examined by a quality assurance manager. This step required a visual inspection of each data element to ensure the recoding procedures were applied correctly. Further, any record missing a USI was replaced with a system-generated ID; however, because the actual student could not be identified across years, the record was preserved for only the given year. Several tables were rerun when coding errors were identified. The final tables were then exported into a format (.txt) that reduced the file size for transfer to the primary researcher.

The primary researcher was also provided with a descriptive report for each SEA data set. The analytics for the data tables explored the frequency distribution of key demographics such as gender, race, and socioeconomic status. This allowed the primary researcher a quick view of those independent variables being introduced into the growth models being tested. Also, potential dependent variables were also examined. The frequency distribution of performance levels for both reading and mathematics was produced along with a new "proficiency" variable. In this variable, the performance levels were coded (see RDGPROF and MTHPROF) into a dichotomous variable identifying students that are proficient (i.e., according to NCLB) or not. The distribution of proficiency was then cross-tabulated by selected independent variables to better understand the performance of subpopulations of students.

APPENDIX 2: DATA ELEMENTS FOR EACH STATE

Field Name	Category	Description	Type	Length
ID1	System	System generated ID used in tracking data migration	Text	0-6
USI_10	Identifiers	Anonymized unique student id number; can be matched across years: Year stamped extension	Text	6-7
UDI_10	Identifiers	Anonymized unique district number; can be matched across years: Year stamped extension	Text	2-4
USchID_10	Identifiers	Unique school identifier created via concatenation of UDI + SchID	Text	4-9
Gender_10	Demographics	Gender; F=Female, M=Male, 9=Unknown	Text	1
Grade_10	Demographics	Grade level; 3-8	Text	1
Race_10	Demographics	Race ethnicity code; 2=Asian/pacific islander, 3=Black, 4=Hispanic, 1=American Indian, 5=White, 9=Unknown	Text	1
SWD_10	Demographics	Indicator for disability status; 0=Student without disability, 1=Student with disability	Text	1
ED_10	Demographics	Indicator for economically disadvantaged status; 0=not economically disadvantaged, 1=economically disadvantaged	Text	1
LEP_10	Demographics	Indicator for limited English proficiency; 0=not limited English proficient, 1=limited	Text	1

RDGProf_10	Performance	Reading proficiency status: 0=Non-proficient; 1=Proficient; 9=NA/Missing	Numeric	1
MTHPL_10	Performance	Proficiency level in the mathematics; 1=Below Basic, 2=Basic, 3=Proficient, 4=Advanced; 9=NA/Missing	Numeric	1
MTHProf_10	Performance	Mathematics proficiency status: 0=Non-proficient; 1=Proficient; 9=NA/Missing	Numeric	1
RDGSS_10	Performance	Scale score on the reading; 9=NA/Missing	Numeric	3
MTHSS_10	Performance	Scale score on the mathematics; 9=NA/Missing	Numeric	3
RDGSEM_10	Performance	Standard error of scale score on the reading; <blank>=NA/Missing	Numeric	2
MTHSEM_10	Performance	Standard error of scale score on the mathematics; <blank>=NA/Missing	Numeric	2

APPENDIX 3: DETAILED RESULTS OF OVERALL MODEL IMPACT

Elementary School

Estimates of Fixed Effects[b],[c]

Parameter	Estimate	Std. Error	df	t	Sig.
Intercept	-0.20	0.86	84783.00	-0.23	0.82
[model=1.00]	-0.34	0.86	84783.00	-0.40	0.69
[model=2.00]	0.64	1.01	84783.00	0.63	0.53
[model=3.00]	-0.26	1.18	84783.00	-0.22	0.83
[model=4.00]	-0.86	1.24	84783.00	-0.69	0.49
[model=5.00]	-0.78	1.25	84783.00	-0.63	0.53
[model=6.00]	-0.54	1.31	84783.00	-0.41	0.68
[model=7.00]	0.29	2.08	84783.00	0.14	0.89
[model=8.00]	0.56	0.87	84783.00	0.64	0.52
[model=9.00]	0.08	0.86	84783.00	0.09	0.93
[model=10.00]	0.23	1.82	84783.00	0.12	0.90
[model=11.00]	0[a]	0.00	.	.	.
[state=1.00]	0.27	1.30	84783.00	0.21	0.84
[state=2.00]	0.37	1.40	84783.00	0.26	0.79
[state=3.00]	0.71	1.12	84783.00	0.63	0.53
[state=4.00]	0[a]	0.00	.	.	.
[model=1.00] * [state=1.00]	-1.38	1.30	84783.00	-1.06	0.29
[model=1.00] * [state=2.00]	-0.67	1.40	84783.00	-0.48	0.63
[model=1.00] * [state=3.00]	-0.54	1.13	84783.00	-0.48	0.63
[model=1.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=2.00] * [state=1.00]	-1.57	1.44	84783.00	-1.10	0.27
[model=2.00] * [state=2.00]	-3.06	1.54	84783.00	-1.98	0.05
[model=2.00] * [state=3.00]	-0.42	1.39	84783.00	-0.31	0.76
[model=2.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=3.00] * [state=1.00]	-0.19	1.65	84783.00	-0.12	0.91
[model=3.00] * [state=2.00]	-0.07	1.70	84783.00	-0.04	0.97
[model=3.00] * [state=3.00]	-0.39	1.44	84783.00	-0.27	0.79
[model=3.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=4.00] * [state=1.00]	-0.09	1.89	84783.00	-0.05	0.96
[model=4.00] * [state=2.00]	0.34	1.76	84783.00	0.19	0.85
[model=4.00] * [state=3.00]	0.53	1.49	84783.00	0.35	0.72

[model=4.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=5.00] * [state=1.00]	-0.09	1.95	84783.00	-0.05	0.96
[model=5.00] * [state=2.00]	0.43	1.78	84783.00	0.24	0.81
[model=5.00] * [state=3.00]	0.53	1.49	84783.00	0.36	0.72
[model=5.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=6.00] * [state=1.00]	0.10	1.57	84783.00	0.06	0.95
[model=6.00] * [state=2.00]	0.22	1.65	84783.00	0.14	0.89
[model=6.00] * [state=3.00]	0[a]	0.00	.	.	.
[model=7.00] * [state=1.00]	-0.35	3.49	84783.00	-0.10	0.92
[model=7.00] * [state=2.00]	-0.43	2.58	84783.00	-0.17	0.87
[model=7.00] * [state=3.00]	-0.54	5.25	84783.00	-0.10	0.92
[model=7.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=8.00] * [state=1.00]	-0.81	1.31	84783.00	-0.61	0.54
[model=8.00] * [state=2.00]	-0.64	1.41	84783.00	-0.46	0.65
[model=8.00] * [state=3.00]	-1.04	1.16	84783.00	-0.90	0.37
[model=8.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=9.00] * [state=1.00]	-0.74	1.30	84783.00	-0.57	0.57
[model=9.00] * [state=2.00]	0.01	1.40	84783.00	0.01	0.99
[model=9.00] * [state=3.00]	-0.86	1.13	84783.00	-0.76	0.45
[model=9.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=10.00] * [state=1.00]	-0.29	2.11	84783.00	-0.14	0.89
[model=10.00] * [state=2.00]	-0.71	2.14	84783.00	-0.33	0.74
[model=10.00] * [state=3.00]	-0.39	2.75	84783.00	-0.14	0.89
[model=10.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=11.00] * [state=1.00]	0[a]	0.00	.	.	.
[model=11.00] * [state=2.00]	0[a]	0.00	.	.	.
[model=11.00] * [state=3.00]	0[a]	0.00	.	.	.
[model=11.00] * [state=4.00]	0[a]	0.00	.	.	.
subject	0.18	0.77	84783.00	0.24	0.81
cohort	-0.06	0.76	84783.00	-0.08	0.94
[model=1.00] * subject	-0.75	0.77	84783.00	-0.98	0.32
[model=2.00] * subject	-1.29	0.89	84783.00	-1.46	0.14

[model=3.00] * subject	-0.12	1.00	84783.00	-0.12	0.90
[model=4.00] * subject	-0.16	1.05	84783.00	-0.15	0.88
[model=5.00] * subject	-0.13	1.04	84783.00	-0.13	0.90
[model=6.00] * subject	0.00	0.77	84783.00	0.00	1.00
[model=7.00] * subject	-0.26	1.93	84783.00	-0.14	0.89
[model=8.00] * subject	-0.44	0.77	84783.00	-0.57	0.57
[model=9.00] * subject	-0.21	0.77	84783.00	-0.28	0.78
[model=10.00] * subject	-0.16	1.43	84783.00	-0.11	0.91
[model=11.00] * subject	0[a]	0.00	.	.	.
[model=1.00] * cohort	1.03	0.76	84783.00	1.34	0.18
[model=2.00] * cohort	0.96	0.91	84783.00	1.06	0.29
[model=3.00] * cohort	0.14	1.06	84783.00	0.13	0.90
[model=4.00] * cohort	-0.24	1.12	84783.00	-0.21	0.83
[model=5.00] * cohort	-0.25	1.14	84783.00	-0.22	0.82
[model=6.00] * cohort	0.21	0.77	84783.00	0.27	0.79
[model=7.00] * cohort	0.10	2.03	84783.00	0.05	0.96
[model=8.00] * cohort	0.52	0.77	84783.00	0.67	0.50
[model=9.00] * cohort	0.42	0.76	84783.00	0.55	0.58
[model=10.00] * cohort	0.07	1.48	84783.00	0.05	0.96
[model=11.00] * cohort	0[a]	0.00	.	.	.
ST_Disadv1	-0.62	1.29	84783.00	-0.48	0.63
ST_Large	-0.21	2.49	84783.00	-0.08	0.93
ST_Disadv2	-0.72	0.98	84783.00	-0.74	0.46
ST_Mobile	-0.98	1.05	84783.00	-0.94	0.35
[model=1.00] * ST_Disadv1	0.92	1.29	84783.00	0.71	0.48
[model=2.00] * ST_Disadv1	0.53	1.59	84783.00	0.33	0.74
[model=3.00] * ST_Disadv1	0.71	2.08	84783.00	0.34	0.73
[model=4.00] * ST_Disadv1	0.90	2.22	84783.00	0.41	0.68
[model=5.00] * ST_Disadv1	0.77	2.17	84783.00	0.36	0.72
[model=6.00] * ST_Disadv1	1.11	1.30	84783.00	0.86	0.39
[model=7.00] * ST_Disadv1	0.46	4.75	84783.00	0.10	0.92
[model=8.00] * ST_Disadv1	0.99	1.30	84783.00	0.76	0.45
[model=9.00] * ST_Disadv1	1.35	1.29	84783.00	1.04	0.30
[model=10.00] * ST_Disadv1	0.34	3.11	84783.00	0.11	0.91
[model=11.00] * ST_Disadv1	0[a]	0.00	.	.	.
[model=1.00] * ST_Large	1.22	2.49	84783.00	0.49	0.62
[model=2.00] * ST_Large	0.04	3.01	84783.00	0.01	0.99
[model=3.00] * ST_Large	0.34	2.83	84783.00	0.12	0.90
[model=4.00] * ST_Large	0.44	2.93	84783.00	0.15	0.88

[model=5.00] * ST_Large	0.32	2.91	84783.00	0.11	0.91
[model=6.00] * ST_Large	-0.03	2.49	84783.00	-0.01	0.99
[model=7.00] * ST_Large	0.18	5.08	84783.00	0.04	0.97
[model=8.00] * ST_Large	-0.17	2.49	84783.00	-0.07	0.94
[model=9.00] * ST_Large	0.54	2.49	84783.00	0.22	0.83
[model=10.00] * ST_Large	0.06	3.80	84783.00	0.01	0.99
[model=11.00] * ST_Large	0[a]	0.00	.	.	.
[model=1.00] * ST_Disadv2	1.04	0.98	84783.00	1.06	0.29
[model=2.00] * ST_Disadv2	0.93	1.17	84783.00	0.79	0.43
[model=3.00] * ST_Disadv2	1.25	1.39	84783.00	0.90	0.37
[model=4.00] * ST_Disadv2	0.47	1.49	84783.00	0.32	0.75
[model=5.00] * ST_Disadv2	0.39	1.49	84783.00	0.26	0.79
[model=6.00] * ST_Disadv2	0.79	0.98	84783.00	0.80	0.42
[model=7.00] * ST_Disadv2	0.51	2.52	84783.00	0.20	0.84
[model=8.00] * ST_Disadv2	0.69	0.98	84783.00	0.70	0.48
[model=9.00] * ST_Disadv2	0.62	0.98	84783.00	0.63	0.53
[model=10.00] * ST_Disadv2	0.69	1.84	84783.00	0.37	0.71
[model=11.00] * ST_Disadv2	0[a]	0.00	.	.	.
[model=1.00] * ST_Mobile	-0.41	1.05	84783.00	-0.39	0.70
[model=2.00] * ST_Mobile	-1.06	1.20	84783.00	-0.88	0.38
[model=3.00] * ST_Mobile	0.91	1.31	84783.00	0.70	0.49
[model=4.00] * ST_Mobile	0.42	1.37	84783.00	0.31	0.76
[model=5.00] * ST_Mobile	0.24	1.39	84783.00	0.17	0.87
[model=6.00] * ST_Mobile	0.52	1.05	84783.00	0.50	0.62
[model=7.00] * ST_Mobile	0.56	2.37	84783.00	0.24	0.81
[model=8.00] * ST_Mobile	-0.14	1.05	84783.00	-0.13	0.89
[model=9.00] * ST_Mobile	-0.04	1.05	84783.00	-0.04	0.97
[model=10.00] * ST_Mobile	0.30	1.94	84783.00	0.15	0.88
[model=11.00] * ST_Mobile	0[a]	0.00	.	.	.

a. This parameter is set to zero because it is redundant.
b. Dependent Variable: School_Q.
c. Residual is weighted by FX_SEsqr.

Middle School

Estimates of Fixed Effects[b],[c]

Parameter	Estimate	Std. Error	df	t	Sig.
Intercept	0.01	0.67	54102.00	0.02	0.98
[model=3.00]	-0.27	0.81	54102.00	-0.34	0.74
[model=4.00]	0.10	0.84	54102.00	0.12	0.90
[model=5.00]	0.08	0.85	54102.00	0.09	0.93
[model=6.00]	0.20	1.20	54102.00	0.17	0.87
[model=7.00]	0.15	0.97	54102.00	0.15	0.88
[model=8.00]	0.29	0.67	54102.00	0.43	0.67
[model=9.00]	0.10	0.67	54102.00	0.15	0.88
[model=10.00]	0[a]	0.00	.	.	.
[state=1.00]	-0.06	0.86	54102.00	-0.07	0.94
[state=2.00]	-0.37	0.98	54102.00	-0.38	0.71
[state=3.00]	0.04	1.12	54102.00	0.03	0.97
[state=4.00]	0[a]	0.00	.	.	.
[model=3.00] * [state=1.00]	0.76	1.15	54102.00	0.66	0.51
[model=3.00] * [state=2.00]	0.76	1.24	54102.00	0.61	0.54
[model=3.00] * [state=3.00]	0.75	1.28	54102.00	0.59	0.56
[model=3.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=4.00] * [state=1.00]	0.63	1.28	54102.00	0.50	0.62
[model=4.00] * [state=2.00]	-0.53	1.26	54102.00	-0.42	0.68
[model=4.00] * [state=3.00]	0.56	1.31	54102.00	0.43	0.67
[model=4.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=5.00] * [state=1.00]	0.56	1.31	54102.00	0.43	0.67
[model=5.00] * [state=2.00]	-0.56	1.28	54102.00	-0.44	0.66
[model=5.00] * [state=3.00]	0.59	1.31	54102.00	0.45	0.65
[model=5.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=6.00] * [state=1.00]	-0.04	1.28	54102.00	-0.03	0.97
[model=6.00] * [state=2.00]	0.38	1.35	54102.00	0.28	0.78
[model=6.00] * [state=3.00]	0[a]	0.00	.	.	.
[model=7.00] * [state=1.00]	0.12	1.80	54102.00	0.07	0.95
[model=7.00] * [state=2.00]	0.30	1.74	54102.00	0.18	0.86
[model=7.00] * [state=3.00]	0.02	1.72	54102.00	0.01	0.99
[model=7.00] *	0[a]	0.00	.	.	.

[state=4.00]					
[model=8.00] * [state=1.00]	-0.09	0.87	54102.00	-0.10	0.92
[model=8.00] * [state=2.00]	0.48	0.98	54102.00	0.49	0.63
[model=8.00] * [state=3.00]	0.03	1.13	54102.00	0.03	0.98
[model=8.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=9.00] * [state=1.00]	-0.21	0.86	54102.00	-0.24	0.81
[model=9.00] * [state=2.00]	0.00	0.98	54102.00	0.00	1.00
[model=9.00] * [state=3.00]	-0.30	1.12	54102.00	-0.27	0.79
[model=9.00] * [state=4.00]	0[a]	0.00	.	.	.
[model=10.00] * [state=1.00]	0[a]	0.00	.	.	.
[model=10.00] * [state=2.00]	0[a]	0.00	.	.	.
[model=10.00] * [state=3.00]	0[a]	0.00	.	.	.
[model=10.00] * [state=4.00]	0[a]	0.00	.	.	.
Subject	0.06	0.62	54102.00	0.09	0.93
Cohort	0.08	0.63	54102.00	0.14	0.89
[model=3.00] * subject	-0.24	0.75	54102.00	-0.33	0.74
[model=4.00] * subject	-0.24	0.78	54102.00	-0.31	0.76
[model=5.00] * subject	-0.18	0.79	54102.00	-0.23	0.82
[model=6.00] * subject	0.03	0.62	54102.00	0.05	0.96
[model=7.00] * subject	-0.06	0.94	54102.00	-0.07	0.95
[model=8.00] * subject	0.01	0.62	54102.00	0.02	0.98
[model=9.00] * subject	-0.17	0.62	54102.00	-0.28	0.78
[model=10.00] * subject	0[a]	0.00	.	.	.
[model=3.00] * cohort	0.03	0.76	54102.00	0.04	0.97
[model=4.00] * cohort	-0.17	0.79	54102.00	-0.21	0.83
[model=5.00] * cohort	-0.21	0.80	54102.00	-0.26	0.79
[model=6.00] * cohort	-0.10	0.63	54102.00	-0.17	0.87
[model=7.00] * cohort	-0.08	0.94	54102.00	-0.09	0.93
[model=8.00] * cohort	-0.17	0.63	54102.00	-0.27	0.79
[model=9.00] * cohort	-0.01	0.63	54102.00	-0.02	0.99
[model=10.00] * cohort	0[a]	0.00	.	.	.
ST_Disadv1	-0.08	1.04	54102.00	-0.08	0.94
ST_Large	-0.38	1.54	54102.00	-0.25	0.81
ST_Disadv2	-0.14	0.75	54102.00	-0.19	0.85
ST_Mobile	-0.64	1.15	54102.00	-0.55	0.58
[model=3.00] * ST_Disadv1	0.99	1.30	54102.00	0.76	0.45
[model=4.00] * ST_Disadv1	-1.16	1.34	54102.00	-0.86	0.39

[model=5.00] * ST_Disadv1	-1.05	1.36	54102.00	-0.77	0.44
[model=6.00] * ST_Disadv1	-0.16	1.05	54102.00	-0.15	0.88
[model=7.00] * ST_Disadv1	-0.25	1.56	54102.00	-0.16	0.87
[model=8.00] * ST_Disadv1	-0.13	1.05	54102.00	-0.13	0.90
[model=9.00] * ST_Disadv1	-0.20	1.04	54102.00	-0.19	0.85
[model=10.00] * ST_Disadv1	0[a]	0.00	.	.	.
[model=3.00] * ST_Large	0.18	1.90	54102.00	0.09	0.92
[model=4.00] * ST_Large	-0.06	1.95	54102.00	-0.03	0.97
[model=5.00] * ST_Large	-0.09	1.95	54102.00	-0.05	0.96
[model=6.00] * ST_Large	0.00	1.54	54102.00	0.00	1.00
[model=7.00] * ST_Large	0.18	2.54	54102.00	0.07	0.94
[model=8.00] * ST_Large	0.05	1.54	54102.00	0.03	0.97
[model=9.00] * ST_Large	0.63	1.54	54102.00	0.41	0.68
[model=10.00] * ST_Large	0[a]	0.00	.	.	.
[model=3.00] * ST_Disadv2	0.26	0.93	54102.00	0.28	0.78
[model=4.00] * ST_Disadv2	-0.94	0.96	54102.00	-0.98	0.33
[model=5.00] * ST_Disadv2	-0.92	0.98	54102.00	-0.94	0.35
[model=6.00] * ST_Disadv2	-0.40	0.75	54102.00	-0.53	0.59
[model=7.00] * ST_Disadv2	-0.35	1.13	54102.00	-0.30	0.76
[model=8.00] * ST_Disadv2	-0.42	0.75	54102.00	-0.55	0.58
[model=9.00] * ST_Disadv2	-0.07	0.75	54102.00	-0.09	0.93
[model=10.00] * ST_Disadv2	0[a]	0.00	.	.	.
[model=3.00] * ST_Mobile	-0.11	1.28	54102.00	-0.08	0.93
[model=4.00] * ST_Mobile	-0.11	1.33	54102.00	-0.08	0.94
[model=5.00] * ST_Mobile	0.07	1.34	54102.00	0.05	0.96
[model=6.00] * ST_Mobile	-0.15	1.16	54102.00	-0.13	0.90
[model=7.00] * ST_Mobile	0.29	1.82	54102.00	0.16	0.87
[model=8.00] * ST_Mobile	-0.77	1.16	54102.00	-0.66	0.51
[model=9.00] * ST_Mobile	0.83	1.16	54102.00	0.72	0.47
[model=10.00] * ST_Mobile	0[a]	0.00	.	.	.

a. This parameter is set to zero because it is redundant.

b. Dependent Variable: School_Q.

c. Residual is weighted by FX_SEsqr.