5-2015

Comparing Schools: From Value Added to Sound Policy

Jeffery Richmond Dean
University of Arkansas, Fayetteville

Comparing Schools: From Value Added to Sound Policy

Comparing Schools: From Value Added to Sound Policy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Education Policy

by

Jeffery Richmond Dean
University of Arkansas
Bachelor of Arts in Philosophy and Physics, 2007

May 2015
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

____
Professor Gary Ritter
Dissertation Director

____
Professor Patrick Wolf
Committee Member

____
Dr. Denise Airola
Committee Member

____
Professor Robert Maranto
Committee Member

Abstract

Over the last twenty years, value added measures (VAMs) have proliferated in education research and policy. Whether applied to teachers, schools, or districts, VAMs have attempted to measure the contribution made by a unit of interest toward observed student outcomes, typically test scores in literacy and math. At the same time, a small number of states have developed methods to formally compare schools on those outcomes, and such methods may be used and intended in ways that qualify them as a kind of VAM. My primary interest is to evaluate the properties of a similar schools model I develop in comparison to three other VAMs. Using statewide student data and math test scores from from 2009 to 2014, I develop a similarity index for comparing schools based on observable student characteristics. Using the rank ordering of schools on this index, I then compare each school's mean math scores to the 15 schools immediately below it and the 15 schools immediately above it. Schools' rankings against their comparison groups are then considered as a VAM and compared to three other school effectiveness models: Student Growth Percentiles (SGP), Student Value Added (SVA), and Mean Prior Z (MPZ). The models are compared based on four properties: fairness, stability, validity, and transparency. I find that the Similar Schools Comparisons (SSC) model is more stable than SGP and SVA, but similar to MPZ. On fairness, defined as the strength of relationship between model results and schools' student demographics, SSC is fairer than the other three models, though all three show a weak overall relationship to demographics. On validity, defined as concurrence between the models, SSC aligns most closely with MPZ and has a modest relationship with SGP and SVA. On transparency for the public and educators, SSC is potentially valuable as it evaluates and compares schools in a highly visible way (ranking them against a known list of similar schools). Yet insofar as it relies on multiple regression to calculate the similarity index, SSC lacks transparency and requires specialized statistical knowledge. SSC is promising for exhibiting stability, fairness, and transparency, but further investigation is needed to

determine its validity and proper interpretation in comparison to other VAMs.

Table of Contents

List of Tables and Figures

Operational Definitions

Value Added Measure (VAM)

A value added measure, or VAM, is considered to be any measure which meets the definition specified by Lissitz (2012): "[a] statistical system that attempt[s] to estimate the effect of some independent variable or factor (teacher, school, etc) upon some dependent variable (state test performance for example)…VAM is a formal quantitative system that is defined by its intention." (p. 5)

Similar Schools Comparisons (SSC)

Similar Schools Comparisons is a model I develop in two stages. First, I develop a method for establishing similarity between schools, and rank schools from highest to lowest based on the achievement level one would expect given school characteristics. This is referred to as the index of similarity. Second, I rank each school against the schools that are most similar to it in terms of the index of similarity. These rankings are considered to be a VAM.

Index of Similarity

In SSCs, this is the measure that results from the regression used to estimate expected achievement levels for schools given their characteristics. It is the basis for establishing comparisons among schools.

Similar Schools Ranks

This term refers to the ranking of each school against its comparison group of 30 other schools. Schools are ranked from -15 to +15, with -15 meaning that a school was below all schools in its comparison group and +15 denoting a school that outperforms its entire comparison group. A school that ranks at the middle of its comparison group is ranked 0.

Student Growth Percentiles (SGP)

Student growth percentiles (Betebenner 2008) are a method for comparing gains in student learning within a subject to the gains observed for students at or near a student's prior test score.

Student Value Added (SVA)

This is a method I develop for estimating schools' contributions to student learning growth by conditioning current test scores on multiple prior test scores in math and literacy as well as a student random effect, the combination of which is intended as a sufficient proxy for student characteristics.

Mean Prior Z (MPZ)

Mean Prior Z is a model I develop which describes students' current achievement in comparison to the mean of all prior available test scores in the same subject before entering his or her current school. At the school level, current mean z-scores are compared to the school-level mean of all students' prior mean standardized scores within subject.

Validity

In this dissertation, validity is considered only as the *concurrent* alignment between models which may or may not have differing interpretations at the school level.

Reliability

In this dissertation, reliability is synonymous with stability, defined as the strength of relationship between school VAM estimates from one year to the next. The strength of this relationship is evaluated in multiple ways.

Fairness

In this dissertation, I define fairness as the measurable strength of relationship between school VAM estimates and schools' observable student characteristics. A model is fairer when it has a weaker relationship to student characteristics, particularly those relating to educational disadvantage.

Transparency

I define transparency as the degree to which a VAM is understandable, replicable, or usable by researchers, educators, or the public. Transparency may differ depending on the population one considers.

Chapter 1: Introduction and Overview

Over the last twenty years, value-added measures (henceforth VAMs) for schools and teachers have proliferated, both in development and use. Perhaps most prominently, researchers have used VAMs to attempt to quantify the contributions made by individual teachers to students' learning outcomes (McCaffrey et al. 2003; Hanushek et al. 2004). At the same time, VAMs have been developed to estimate the contribution made by schools toward student outcomes. These attempts have so far received less publicity than models focusing on teachers, and have received relatively less emphasis in policymaking. Teacher VAMs have also recently been given a measure of predictive validity, defined as differences in adult outcomes such as college attendance, college selectivity, income, wealth profile of one's neighborhood, and retirement savings (Chetty et al. 2011; Chetty et al. 2013) as a function of estimated teacher effectiveness. Examining teacher effectiveness in grades 4-8, Chetty and colleagues famously estimated that replacing a teacher in the bottom 5% of effectiveness with an average teacher boosted the present value of the lifetime earnings of students in a typical classroom by roughly $250,000. While this estimate may be sensitive to a variety of factors including the VAM used, it nevertheless lends support to interpretations of VAMs as measures which are predictive of substantial school impacts on a range of desirable social and individual outcomes. In comparison to the teacher VAMs considered in these studies, school VAMs have additional features which may help to give them traction in policy settings. Most importantly, larger sample sizes for schools than for teachers mitigate volatility, which if very high can undermine confidence in the measures themselves.

Related to the development of school VAMs, a few states (most notably California) have developed measures and rankings which attempt to make explicit comparisons between schools. These sorts of comparisons should be considered a type of VAM, since the intent behind their development and use is, like other VAMs, to estimate school effectiveness by accounting for non-school factors influencing school outcomes, particularly student demographics and prior achievement. Where other VAMs account for such factors through multiple regression at both the student and school level, similar school comparisons (SSCs) are made on the basis of regression models which yield relatively straightforward school-level expectations on student outcomes. To compare schools to those identified as similar, expectations are rank-ordered, and this ranking is used to compare a school to other schools with expectations ranked immediately above and below it. These sorts of comparisons, while subject to many of the same concerns as VAMs generally, nevertheless may yield information that is potentially more tangible to educators and the public. Where a VAM may simply yield a score that characterizes the difference between a school's expected and actual performance on a given measure such as proficiency rates or standardized scores—positive is good, negative is bad, null is typical—the same evaluative message that is conveyed in a school's coefficient or residual in a VAM is conveyed in the case of SSCs by ranking a school against a known, publicly visible group of comparison schools.

Potential Benefits of SSCs

If VAMs are to be meaningful to the public, they must be accepted largely based on the trust placed in the body or individuals developing such measures. As controls for non-school impacts on student achievement become more sophisticated and comprehensive, they improve VAMs' validity and fairness while at the same time reducing their transparency. Even if such measures are internally valid, it is extremely difficult for non-specialists to assess the meaningfulness of the measures produced by a value-added model. And in any case, a school being labeled "high value-added" still instinctively begs the question, "compared to what?" This question-begging points toward the intuitive appeal of SSCs. SSCs take the comparisons implicit in VAM scores and render them explicit by situating a school's outcome in comparison to the outcomes of a set of known schools. Thus, this measure is preferable to other VAMs in that it is more accessible to the public. The question I address in this study is: to what extent does this more accessible method yield the same results as more sophisticated VAM methods?

Even if there is a disjoint between public perceptions of what is comparable and what an SSC model shows to be comparable, the fact that such considerations take place would be evidence that comparability itself is being considered by the public, rather than measures such as the level of or changes in school proficiency rates, which are both less fair when comparing schools of differing demographics and more ambiguous due to widely varying levels of achievement denoted by 'proficiency' (Loveless 2012).[1] Such a phenomenon might engage the public with the question of school effectiveness, which otherwise is neglected or obscured in existing forms of school accountability.

By design, SSCs are thus more tangible, perhaps even more transparent, than most VAMs. If an SSC is shown to be as fair and valid as another VAM, yet more transparent, then the SSC should be preferable for use in policy and the public provision of information on school quality. In making this claim, I borrow from the analysis presented in Polikoff (2013) concerning the evaluation of state accountability systems for schools. Polikoff lays out four criteria by which to evaluate school accountability models. *Construct validity*, a concept borrowed from measurement questions typically pertaining to assessments themselves, asks whether the indicators used in an accountability model are meaningfully aligned to what they convey. *Reliability* concerns the degree to which schools are consistently placed in the same performance categories over time. Models are more or less *fair* depending on the degree to which performance classifications for schools are influenced by

[1] Examining the relationship between NAEP test scores and proficiency percentages at the state level, Loveless found correlations on 2009 literacy and math assessments in grades 4 and 8 below p=0.1.

demographics, which should be accounted for when attempting to describe the effectiveness of schools with a given set of inputs. *Transparency* describes 'the level to which the performance goal-setting process is clearly documented and the performance measures are clearly understandable.' Even beyond its particular use in the context of goal setting, transparency is an important criterion in evaluating school accountability measures. In Polikoff, the definition and operationalization of each of these criteria depends upon standards of practice issued in recent years by the American Psychological Association (APA), the American Education Research Association (AERA), and the National Council on Measurement in Education (NCME). I likewise borrow from these standards to compare VAMs. Beyond these four categories, I also consider the *currency* of models, which is the degree to which agents and stakeholders are likely to internalize and act upon the results of VAMs or SSCs.

As shown in Neal (2010), a system in which every agent is equally incentivized by goals which are achievable but challenging is likeliest to maximize aggregate outcomes. In the case of schools, a focus on effectiveness and improvement at the frontiers of what is feasible for each school is, among all possible test-based school accountability mechanisms, likeliest to yield optimal overall outcomes. This may be accomplished by defining goals in relative terms, such as percentiles or comparison groups, rather than absolute goals, such as 100% proficiency rates on student tests under No Child Left Behind. Similar schools comparisons are an example of such a system. The core of Neal's claim is that incentivizing marginal improvement for all schools (and within schools, incentivizing marginal improvement for all students) is optimal because no schools are so far above or below targets that marginal improvement is unlikely to result in a better rating. In most state accountability models, not all schools have targets which lie immediately above schools' current achievement levels. Schools with targets that greatly exceed current achievement will not exert optimal effort, because the likeliest outcome of such effort is an outcome that is marginally better, but not high enough to be rewarded in such a target-based system. At the other extreme, schools with achievement levels lying comfortably above their targets need not worry about falling below them, so these schools also face no incentive for marginal improvement. While new accountability models adopted under ESEA Flexibility since 2011 have ameliorated this dynamic, it still persists in many cases. Lastly, accountability systems that focus on proficiency further create the threat of inefficiency within schools, duplicating the school-level dynamic described here at the student level. This is perhaps best known as the infamous "bubble kid" or "educational triage" problem, in which a disproportionate emphasis is placed upon students near and immediately below proficiency cut points (Booher-Jennings 2005). In comparison to these systems, setting goals for schools that are relative to a set of similar schools is likelier to avoid the extremes of goals that are either effortless or not feasible.

While the insights of Neal (2010) have negative implications both for No Child Left Behind (NCLB) and ESEA Flexibility models adopted by states over the last dozen years, the value of marginality is worth considering as well in the policy context of SSCs and VAMs. Whether implicitly or explicitly, both compare schools serving similar student populations and are thus likely to help uncover effectiveness at multiple levels and along multiple dimensions. This presents the possibility for schools to learn from those that are shown to be highly effective with comparable populations of students (Eddy-Spicer 2014). Best practices for socioeconomically advantaged student populations may be (and likely is) quite different from best practices among schools serving disadvantaged students. Educators and school improvement specialists understand this, but some accountability systems do a poor job of uncovering relative success or underperformance at all levels of privilege and disadvantage. A model that does so can facilitate improvements in practice as well as provide meaningful information to the public.

The potential policy use of such measures extends beyond publicizing schools' similar school rankings, in the hopes of informing parents and communities. Recent developments in England align particularly well with SSCs. Since the current government came to power in 2010, a policy of "horizontal accountability" has been pursued, in which schools that are judged to be underperforming are paired with highly effective schools serving similar student populations (Eddy-Spicer 2014). It is worth noting that such judgments in the English system are made based ultimately upon school inspections carried out by expert educators, rather than based strictly on test scores; inspection results include reference both to objective outcomes and performance measures as well as the subjective judgment of evaluators. Nevertheless, the guidance given to evaluators encourages them to consider schools' results in the context of the student populations they serve, and the comparisons made possible under that system might likewise be possible even under an accountability framework utilizing only student test scores to generate comparisons. SSCs thus have the potential to inform and impact practice, as well as to fulfill the need to better inform educators and the public.

In the case that other VAMs become trusted and meaningful, they do so by creating a new dimension of understanding among educators and stakeholders. People are likely to understand and act upon such measures only after a matter of time, once the results of VAMs become intuitively understood and used as a point of comparison among schools. SSCs, on the other hand, allow people to compare their schools to a known, *publicly visible* set of comparison schools. As with other VAMs, guidance is needed in interpretation, for example, in order to determine how far above or below the median a school would need to perform in order to be truly considered over- or under-performing. Yet the existence and visibility of such a list could allow people to anchor SSC ratings to existing perceptions of schools. To the degree that SSC orderings align with those perceptions, they leverage them in conveying information on schools' relative effectiveness.

Lastly, both VAMs and SSCs provide evidence of school effectiveness on terms that are inherently relative. The contribution made by a school toward outcomes accounting for a set of inputs can never be stated absolutely, but only in comparison to other schools in the model. This means that, within any given model, whether a VAM or SSC, comparisons are zero sum. While such a property may intuitively seem to be a drawback, having zero-sum comparisons between schools can serve as a safeguard against merely political maneuvers used to over- or under-state the aggregate quality of a set of schools (Neal 2010; Loveless 2012). Inherently, if VAMs are zero-sum, then SSCs render this feature more explicit, without themselves being more or less relative than VAMs.

The potential benefits of SSCs in comparison to VAMs as well as existing accountability mechanisms are numerous: greater transparency and tangibility, fairer incentives for schools with respect to student background, and robustness to gaming and shifts in achievement criteria. To examine whether these potential benefits can be realized in an existing policy and testing framework, I organize my dissertation as follows:

In Chapter 2, I review the current state of the literature regarding the meaning, usefulness, and validity of VAMs and SSCs in education policy in the United States, with most attention focusing on school VAMs. I pay particular attention to the way in which VAMs are interpreted to and by the public. I also examine the degree to which school VAMs withstand many of the criticisms frequently brought against VAMs used for teacher evaluation. I consider general modeling questions regarding the relative merits of parametric and nonparametric methods in value-added modeling. Following this treatment, I examine the properties VAMs must satisfy to permit causal interpretations of their school or teacher estimates. Following this, I examine the relative merits of cross-sectional, 'snapshot' VAMs in comparison to growth-oriented VAMs at the school level.

In addition to considering the general concerns just described, in Chapter 2 I also include an examination of the particular VAMs selected for comparison in Chapters 4-5. I address the specification, interpretation, and shortcomings of each of four models: Similar Schools Comparisons, Student Growth Percentiles, Student Value Added, and Mean Prior Z.

In Chapter 3, I propose a model for Similar Schools Comparisons using student test scores in a southern state. Due to technical limitations pertaining to literacy scores, I use only math test scores to develop and evaluate SSCs. I further confine my analysis to grades 6-8. The SSC model employs a set of cross-sectional parametric controls for student demographics at the school level. For comparison, I develop and estimate three other VAMs. I estimate Student Growth Percentiles (SGP), which describe a student's scale score growth from one year to the next in relation to students with similar prior scores; at the school level, the median student SGP is used. I also explicitly develop the model I refer to as Student Value Added (SVA), aligning it more closely with the state of the art in VAMs used for teacher evaluation. This model uses student random effects as well as a typical set of student and school controls for non-school factors influencing achievement. Finally, the Mean Prior Z (MPZ) model uses the average of students' within-subject standardized scores prior to entering their current school as a measure of student characteristics, rather than using demographic data as in the SSC model. These three models are specified as additional models against which to assess the results of Similar Schools Comparisons.

Tentatively, the Similar Schools Comparisons model is considered a VAM, based not on validation, which is one of the central questions of this dissertation, but on its intent. In defining it as a VAM, I follow the definition proposed by Lissitz et al. (2012):

> "VAM[s] are statistical systems that attempt to estimate the effect of some independent variable or factor (teacher, school, etc.) upon some dependent variable (state test performance for example). […] Unlike multiple regression, which is a class of particular models that everyone who had a minimum of statistics background would recognize, there is no single or even limited class of models that qualify as a VAM model. In other words, VAM is a formal quantitative system that is defined by its intention." (p. 5)

Additionally, I consider the ideal number of schools in a comparison group for SSC; if there are too few, then comparisons may become prohibitively noisy. Yet having a very large number of comparison schools can complicate inference at the tails of school performance levels, as well as rendering each comparison group less intuitively understandable. It is possible to expect that communities and school leaders could gauge the fairness of 20-30 comparison schools, but one should not expect a similar level of understanding when faced with a comparison group of 100 schools. I also consider the sensitivity of these models to the type of assessment used through a comparison of school orderings using the state's criterion-based assessment and the Iowa Test of Basic Skills (ITBS). Lastly, I consider the reliability of these models using between one and four years of student test scores. I compare the models' results to measures currently used in the state.

Finally in Chapter 3, I introduce analytic methods used to compare the four VAMs, including SSCs, using the framework presented in Polikoff (2013): construct validity, reliability, fairness, and transparency.

In Chapter 4, I consider the quantitative requirements necessary to make similar schools comparisons useful for school accountability in state policy, through the criteria of fairness, reliability (or stability), and validity.[2] To consider fairness, I correlate school estimates under each of the VAMs with school characteristics, as well as regressing estimates upon the same set of characteristics. To consider stability, I calculate year-to-year correlations of school estimates in each of the four VAMs across the four years for which school VAM estimates are available. I also sort these estimates into quintiles and compare prior-year results against current-year distributions. Finally, I examine validity in two major ways. The first way is to simply compare school estimates by quintile between the different VAMs, and assess the strength of relationship between each of the

[2] As used here, the term "validity" refers to concurrent validity, broadly defined within psychology as the alignment of an instrument or measure of interest with another instrument for which criterion validity has already been established, when the two measures are estimated on the same population at the same time.

models. The second way is to examine whether schools on which the models substantially disagree have distinctly different characteristics from models on which the models show approximate agreement.

In Chapter 5, I summarize the overall cautions and limitations pertaining to SSCs in relation both to other school VAMs and to non-VAM measures used in school accountability. All measures used in school accountability are evaluated against a complex set of criteria, many of which trade off against each other. In addition to the properties borrowed from Polikoff's analysis, I further consider the potential for SSCs to inform practice and properly incentivize school leadership. I also discuss possible extensions to the dissertation in terms of examining transparency in greater detail, as well as greater attention to how the SSC model might be improved and best interpreted. It is important to consider the advisability of the use of SSCs and VAMs in policy separately from the current state of the science; the fact that a measure is currently too biased or noisy should not inhibit its further development for future use. The potential benefit of the use of SSCs, subject to appropriate requirements, is significant.

Chapter 2: Review of Literature

The research literature informing and motivating an examination of similar schools methods is drawn from a variety of sources, only one of which is the sizable literature on value-added measures in education. The impetus for the questions that follow also draws upon broader issues regarding the purposes and requirements of school accountability, as well as insights from economics which suggest the use of relative comparisons among individuals and institutions as optimal for maximizing desired outcomes. The following sections review the existing literature regarding the following questions:

(1) What statistical and organizational advantages do school value-added models (VAMs) have relative to teacher VAMs?

(2) What prior studies exist which compare VAMs in the manner proposed in this dissertation?

(3) What are the relative merits of parametric and nonparametric methods in the development and use of VAMs?

(4) What conditions are necessary to permit a causal interpretation of VAMs?

(5) What value do contemporaneous, "snapshot" VAMs potentially have in comparison to VAMs which rely primarily on changes in student achievement?

(6) What examples of similar schools comparisons (SSCs) have existed in education policy up to the present, and what has been the basis for establishing similarity?

(7) Should the facilitation of marginal comparisons between schools be a desirable property of VAMs?

In addition to these questions, I explicitly consider key questions underlying the SSCs against which

I compare the properties of VAMs in Chapters 4 and 5.

Advantage of School VAMs over Teacher VAMs

School VAMs have several potential advantages over the limitations and criticisms which are frequently brought against teacher VAMs (Harris 2009). By estimating and potentially incentivizing effectiveness at the school level, school VAMs avoid the threat of encouraging perverse competition between teachers within a school. Rather than zero-sum competitions within a school, every teacher reflected under a school VAM benefits (or suffers) from the excellence (or underperformance) of peer teachers. School leaders likewise face no incentive to assign students to teachers on unobservable traits to boost or lower individual teachers' estimates of effectiveness, which is a criticism frequently brought against teacher VAMs (Rothstein 2009). Many decisions impacting students are made at the school level, so focusing on the school rather than the teachers also avoids holding teachers responsible for organizational factors over which they have little or no control (Harris 2009). Schools also typically have anywhere from ten to fifty times as many students as do individual teachers in any given year, which greatly reduces the lack of precision and reliability in teacher VAMs. Previous research (Schochet & Chiang 2012) has estimated a significant reduction in the rate of Type I and II misclassification errors when shifting from teacher VAMs to school VAMs. Schochet & Chiang used simulated data with OLS and empirical Bayes estimators, which allowed them to examine the frequency with which teacher or school parameters (which, in the case of a simulation, are true by definition) were accurately estimated by their value-added estimators. They found that the error rate for schools (10%) was roughly half that seen for teachers (20%). If true in other settings, this can allow for a greater share of schools to be identified as significantly effective or ineffective, or reduce the error rate, holding the share of such schools constant (e.g. top and bottom 20% identified as highly effective or ineffective).

Harris also points out that school VAMs incorporate teacher effects in cases where teacher effects exist but cannot be observed. Due to turnover among teachers, only those with several years of experience may be reliably identified as effective; multiple years of data are typically recommended to improve reliability of teacher VAMs to an acceptable level. School VAMs implicitly incorporate the contribution of every teacher in every year and tested grade toward student outcomes, so that a one-year or two-year novice teacher who goes unidentified individually nevertheless contributes to a school's effectiveness calculation. To balance these advantages, Harris considers the threat of what he calls the "free rider problem" in school VAMs: that teachers would face a weaker incentive themselves and simply hide behind more effective teachers, diminishing aggregate effort and outcomes. He asserts that this can be ameliorated by properly incentivizing school leadership to focus on school-level effectiveness. School VAMs may provide this incentive through public visibility, which can inform local perceptions of school quality and drive local responses, or through tying them to explicit incentives such as financial rewards or greater operational autonomy from the state. Within the school, teachers then would be linked to this school-level incentive through evaluation by school leaders, peer feedback, or teamwork in which groups accept collective responsibility for desired outcomes.

In considering the incentives faced by school leadership, principal effectiveness is subsumed within, and also distinct from, school effectiveness. A nascent literature on principal value-added since 2009 has investigated this linkage empirically (Branch et al. 2013; Chiang et al. 2012) by examining the effect leadership changes have on school performance. Chiang et al. (2012) found that variations in principal effectiveness only explained 14% of the variation in school effectiveness, suggesting that school effectiveness is a poor proxy for principal effectiveness. While it may be acceptable to integrate school effectiveness estimates into principal evaluation in a purely diagnostic way, responsibility for school effectiveness is, by the best current evidence, broadly distributed among leadership, staff, and the broader community. Given this weak relationship, the best use of school effectiveness in regard to leadership incentives may be to inform and equip school boards and district leadership in their decision-making roles with regard to school leadership by considering effectiveness measures alongside local practice and judgment.

Comparing VAMs

Goldhaber et al. (2012) examined the agreement between six common teacher VAMs using student data from North Carolina. Although the focus of this dissertation is school VAMs, rather than teacher VAMs, Goldhaber's comparison is relevant to the comparison of models I develop and present in Chapters 3 and 4. Most importantly, Goldhaber and colleagues found that models which controlled for student background characteristics and lagged test scores were very highly correlated with teachers' median student growth percentiles (SGP). This was observed, in the words of the authors, "despite the fact that the two methods for estimating teacher effectiveness are, at least conceptually, quite different." (p. 4) They also found that the inclusion of a student fixed effect, rather than a simple lag or a random effect, lowered the correlation of VAMs with SGP.

Atteberry (2011) examined six relatively simple value-added models using student data and test scores from high schools in California. This study showed how sensitive school results and rankings were to the varying definitions underlying the broad question of school effectiveness. Further, she showed that the challenge of value-added modeling is unique for high schools in that students are not typically tested in the same domain or subject in consecutive years, thereby complicating inferences about learning growth. High schools also have multiple outcomes, such as state-mandated end-of-course (EOC) examinations, AP tests, minimum competency exams, and ultimately, graduation; whether and how to account for these differing outcomes is a different question than the major questions attending VAMs when students are tested in multiple consecutive grades.

Lissitz (2012) examined 11 student growth VAMs using student data from a large district in Maryland from 2008 to 2010. While all 11 models necessarily included prior test scores, they did not condition on student characteristics. Three of the models were normative, in that they compared student growth to other students starting at or near the same prior score; other models implicitly compared a student's gain to other students regardless of prior level, and therefore required a stricter set of assessment properties to be valid. At the school level, he found that growth VAMs which examine transitions across well-defined performance levels are more highly correlated with initial student achievement levels in math and reading than is the case for regression-based VAMs, which in his analysis included student growth percentiles (Betebenner 2008). He also found that correlations between school rankings on different growth VAMs decreased at higher grade levels, implying that the choice of model is more consequential for middle and junior high schools than for elementary schools.

Goldschmidt et al. (2012) examined the behavior of nine different growth models across four states. They classified their models into four categories: gain models, regression models, value added models, and normative models. They used a stricter definition of value added, so that even though any of the models could be intended to evaluate school effectiveness, all but one (the Layered Model) was not designed to strictly isolate schools' contributions to student learning. The regression models they used (Covariate Adjusted Fixed Effects and Covariate Adjusted Random Effects) controlled for student characteristics. Finally, they also included student growth percentiles for comparison, describing it as a normative model that is designed to describe student growth in comparison to students at or near the same prior test score, while not meriting a causal interpretation on the part of schools. Goldschmidt asserted that, among all the models considered, there was no single "best" model since each addresses different questions about schools.

Importantly, they found that there were differences between states in how each model worked, due most likely to contextual factors such as scaling and testing procedures, as well as student and school characteristics that are unique to each state. In other words, the assumption that VAMs will perform similarly across states is not warranted in their study.

Table 1 summarizes the scope and nature of the school VAM comparison studies just described. All models examined in all studies used student data including prior test scores, but varied in how they accounted for student characteristics as well as whether changes in student achievement were compared only to students at similar prior levels (normative models) or to students at all levels, as is the case for all non-normative models.

Table 1. Models Included in Selected VAM Comparison Studies

Study	Data Source	Number of Models	Models with Student Fixed or Random Effects	Models with Prior Test Scores	Models with Student Demographics	Normative Models
Goldhaber 2012	North Carolina	6	1	6	3	1
Goldschmidt 2012	DE, HI, NC, WI	9	2	9	2	1
Lissitz 2012	Maryland (district)	11	0	11	0	3
Atteberry 2011	California (four districts)	6	6	6	5	0

Stuit et al. (2014) also carried out an analysis similar to the one proposed in this dissertation, but rather than focusing on differences between VAMs, the authors focused on the sensitivity of VAMs to different assessments with the same students in Indiana. The most useful aspect of their study for my dissertation is the analytical techniques they used for comparing VAMs, which included correlational analyses (Pearson coefficients) as well as comparisons between VAM-assessment combinations by quintile. Finally, Stuit et al. considered statistical significance as an important threshold for comparison, since 95% confidence of effectiveness on one model should very rarely, if ever, correspond with ineffectiveness on another model regardless of quintile distributions or correlations.

Parametric vs Nonparametric Value-added Modeling

Using parametric methods such as ordinary least squares with fixed or random student effects has in recent years become the preferred method for estimating school and teacher effectiveness (Ehlert et al. 2013), as these models most closely meet the conditions described below. Ehlert and colleagues examine three different value-added models: student growth percentiles, which control very precisely for prior test scores but make no control for fixed student characteristics; a value-added model which introduces student observable characteristics and random effects in the model, reducing the relationship of estimates with classroom or school characteristics; and a third model, the so-called 'proportional' model, which fully eliminates the confound between effectiveness estimates and measured student characteristics. These methods all rely upon multiple prior years of student test scores, though strictly speaking, student growth percentiles are considered a form of quantile regression which is considered nonparametric with respect to variation in the independent variable (Betebenner 2008, 2011).

Wright (2010) treats the possible advantages of nonparametric methods over traditional parametric models involving some form of ordinary least squares (OLS) regression. Parametric regression models involving OLS assume that the underlying data are normally distributed, that the effects of regressors upon a dependent variable are linear, and that error variance is constant (homoscedasticity). When these requirements are met, or at least closely approximated, parametric models are more efficient and informative than nonparametric methods, and are thus preferable. Nonparametric models make fewer assumptions and are thus useful in cases where OLS assumptions are violated. Nonparametric methods become especially useful when the relationship between dependent and independent variables is nonlinear, especially in a way that is not easily captured by polynomial terms or is even multi-dimensional. Additionally, nonparametric methods occasionally can have the advantage of transparency, in that some of them may not require an understanding of OLS to interpret their validity and results. However, very simple nonparametric models may not yield straightforward estimates of error, perhaps requiring estimation methods such as bootstrapping. Bootstrapping is a statistical method in which a large number of random samples are drawn from existing data *with replacement* (i.e. a particular observation in existing data may be sampled once, more than once, or not at all in a given bootstrap sample), so that each random sample is unique despite each being drawn from the same base sample; these samples can then be compared against each other on desired properties such as mean and standard deviation to make error estimates on those measures where error could not be estimated using conventional techniques. This aspect of nonparametrics can inhibit transparency, both through its inherent opacity to individuals without statistical training, as well as having a different interpretation than standard errors in OLS. Thus in the context of considering parametric and nonparametric value-added methodologies, one must weigh the occasionally rival concerns of efficiency, transparency, and flexibility, as well as the threat of bias or false inference when OLS is used despite severe violations of its assumptions.

Wright (2010) treats the use of nonparametrics in the case of student growth modeling, considering both the standardized gain model (Reback 2008) and student growth percentiles (Betebenner 2009). Yet the principles he considers with regard to those models could be applied in examining non-growth measures as well.

Causal Interpretation

In order for school VAMs to be given a causal interpretation, a strict set of conditions must be met, or at least closely approximated in a way that can be ameliorated through the appropriate use and interpretation of VAMs (Reardon & Raudenbush 2009). Reardon and Raudenbush (2009) lay out several conditions required for the establishment of causality. First, it must be 'theoretically meaningful to define the potential outcome for each student' if that student were to attend each of the schools in a model. Second, students must possess one and only one potential outcome in a school. Third, and especially problematic, is the requirement that 'the units of the test score are on an interval scale of social interest.' Ballou (2009) shows that an interval interpretation of scales is typically unjustified, and that efforts to make interval interpretations can even be meaningless. Fourth, the causal effects of schools must be separable from and invariant to student background. This is especially a problem in the case of growth measures that do not condition on student characteristics, but only prior scores (Betebenner 2009; Ehlert 2013). While such measures are neither intended nor designed to have a causal interpretation, they are occasionally misused in efforts to establish causality, particularly with teachers. Fifth is the requirement of 'strongly ignorable treatment assignment', which simply requires that in the case where assignment of students into schools covaries with student background, all observable pre-assignment student characteristics are controlled for in a value-added model. Sixth and finally, Reardon and Raudenbush (2009) find that a model must specify potential outcomes for students who are *not* in a given school, known as the 'functional form' assumption.

These conditions require a strict set of controls and point toward the need for school value-added to be estimated using student growth on an interval scale, with a prior score for each student, as well as conditioning expected growth on student background. When these are met together, they merit a causal interpretation.

The choice and nature of assessment in developing VAMs is particularly salient given the third assumption put forth by Reardon and Raudenbush, that of a scale having interval properties. Additionally, assessments may differ in the degree to which they can accurately distinguish and measure student performance at the tails. Stuit et al. (2014) compared the results of value-added modeling for teachers in Indiana using the state's criterion-referenced assessment (ISTEP) in comparison to the well-known Measures of Academic Progress (MAP). They used three methods to understand differences in estimates of effectiveness using the two assessments: the correlation of effectiveness estimates for individual teachers between the two assessments, comparisons of quintile rankings of teachers across assessment, and finally asking whether teachers identified as significantly below mean effectiveness with 95% confidence on one assessment were ever found to be significantly above mean effectiveness on the other assessment. They found that, while the two models yielded modest disagreement, there were no teachers identified significantly at opposite ends of effectiveness across assessments. If this is the case with teachers, it is hypothesized to be even truer with schools, which due to a greater number of students suffer from less noise and intertemporal variability in their effectiveness.

Despite the many concerns attending the sensitivities and causal interpretation of school VAMs, there remains a place for measures of school performance that cannot be given a causal interpretation yet nevertheless help situate school performance in comparison to other schools on the basis of student characteristics. Perhaps the best instance of this is Betebenner's (2009) argument for the use of student growth percentiles, which he advocates as purely diagnostic, but which also have the advantage of being easily interpretable as well as robust to problems arising from assessments not being truly interval-scaled (Briggs & Weeks 2009). In certain respects an ordering of schools on demographic expectations could likewise be interpretable and robust to assessment weaknesses without meriting a causal interpretation. This analogy breaks down when considering that student characteristics are the control in the latter, whereas the control in the former is strictly prior test scores regardless of student characteristics.

The following sections present two models which can be used to evaluate school effectiveness but which vary in the degree to which a causal interpretation can be drawn from their results. They are presented here for comparison against similar schools methods, which are developed in Chapter 3.

Student Growth Percentiles

Student growth percentiles (SGPs), as first developed (Betebenner 2008), represent an attempt to provide a *description* of student growth that is robust relative to students with similar prior achievement. As a description, they are not intended to provide a causal interpretation, but may nevertheless help as a diagnostic tool in pursuing an understanding of causality at the teacher or school level. SGPs were developed, in the words of Betebenner, in response to the tendency of most value-added models to "[skew] discussions about growth models toward causal claims at the expense of description." As with similar schools methods, SGPs are intended as a useful and understandable first step toward causal inference that, ideally, is adaptable to the weaknesses and assumption violations often present in student assessments.

Student growth percentiles represent gains in student learning for a student in comparison to other students with prior test scores in the same range and with similar trajectories. Unlike the Student Value Added model discussed below, the methodology developed for estimating growth percentiles uses as many prior test scores in the same subject as are available (Betebenner 2008; Betebenner 2011). SGPs are based first on estimating conditional density given a student's test score in the most recent year prior to the period of analysis, using all prior test scores for conditioning. A student's growth percentile is thus the percentile at which the student scores in the latest year within the conditional density (i.e. other test scores closest to a student's prior scores). In simpler language, a student's growth is considered only in comparison to other students at or near the same point previously. Thus a very low-achieving student is compared only to a precise set of other low-achieving students, and the probability of such a student having a high growth percentile is similar to that for high-scoring students.

Betebenner uses quantile regression to estimate students' conditional densities. This differs from traditional ordinary least squares in several ways. First, a similar functional form is estimated independently for several different quantiles. The quantile, which represents a small portion of the variation in the independent variable at time t, is defined by separating prior test scores into percentiles and estimating growth within each of them. Further, the functional form used to relate the independent variable (prior test score) to the dependent variable (latest test score), rather than being linear, is instead a cubic base spline function. Splines are piecewise polynomial functions, in which the boundaries between pieces are known as knots. At these boundaries, in the case of cubic polynomials, the first and second derivatives must be equal so that the function is smoothed. It is perhaps most helpful to think of splines as a more flexible, robust version for establishing best fit, as opposed to linear estimations in OLS, by allowing nonlinearity and relaxing the dependence of parameterization on the conditional mean.

Coefficients in ordinary least squares are interpreted at the conditional mean assuming linearity, but quantile regression relaxes this assumption, thereby allowing that the relationship between dependent and independent variables be nonlinear or curvilinear. Quantile regression does this by estimating parameters on several conditional quantiles of the dependent variable. The use of more than one prior test score for each student helps situate the most recent prior score. For example, a most recent prior score that is well above a student's individual trajectory may indicate a higher likelihood of regression back toward his or her trajectory, regardless of true growth. When only a single prior test score is available, it is difficult to estimate within-student random error and thereby distinguish observed scores from true scores, but multiple scores allow positive or negative outliers to be identified and, in doing so, help account for the influence of random error in growth estimation. Quantile regression using multiple priors in this manner helps account for this challenge.

Finally, the best-fit functions for each conditional density (quantile) are estimated using median regression, which allows for heteroscedasticity in the data and is far less sensitive to outliers than is OLS; rather than summing the squared deviations from the best-fit line, median regression concerns only whether an additional value is above or below the line. Taken together, these three conditions yield a model that is highly adaptable to the flaws and idiosyncrasies of different assessments.

Student Value Added

Student Value Added (SVA) attempts to quantify schools' contribution to student growth in test scores after conditioning on multiple prior test scores in multiple subjects. Anderson and Hsiao (1981) and Arellano and Bond (1991) have shown, in settings outside value-added modeling, that the use of multiple lags on distinct but related measures can effectively capture the same information as that contained in individuals' observable background characteristics. The most well-known model of this kind in education can be found in Aaronson et al. (2007). Aaronson and colleagues estimated teacher effects in Chicago by regressing students' current test scores on lagged test scores as well as individual, teacher, school, and year dummy variables. Aaronson's model intentionally excluded demographics, not only because prior test scores captured demographic effects, but also because test scores were, in their study, more readily available than demographic data, which if missing in non-random ways could have biased their model. Whether in education or any other use of individual-level panel data, one can generalize that to the degree demographics influence growth trajectories, the use of multiple lagged dependent variables effectively captures that influence. In the case of students included in this model, reading test scores represent the off-subject counterfactual when math is the subject of interest; if reading scores were the subject of interest, then math scores would represent the off-subject counterfactual.

The estimation thus employed at the student level for SVA is as follows:

$$Y_{it} = \beta_1 Y_{i,t-1} + \beta_2 Y_{i,t-2} + \alpha_1 X_{i,t-1} + \alpha_2 X_{i,t-2} + (u_{0i} + \varepsilon_{it})$$

Where Y_{it} is the math score of student I at time t, $Y_{i,t-1}$ and $Y_{i,t-2}$ are the math scores of student I at time t-1 and t-2 respectively, and $X_{i,t-1}$ and $X_{i,t-2}$ are the reading scores of student I at time t-1 and t-2. The individual-level random effect u_{0i} is absorbed in the error term with ε_{it}. A set of school indicator variables are further included at the second level in the model for the second level of estimation. The coefficient estimates on each of these schools is thus interpreted as the contribution made by the school toward explaining additional variation in outcomes Y_{it} at the first level. The decision of which school to treat as the reference point for school estimates does not influence the ordering of school effects, but may affect inference about which schools are significantly effective or ineffective. This is because all estimates on school dummies are made relative to the reference school, so that p-values on schools' coefficients can be interpreted as whether a school's estimated effectiveness is significantly different from that of the reference school. If the reference school is a particularly ineffective school, then most schools will be significantly effective by comparison. For this reason, I rescale school coefficients so that a school of median effectiveness has a coefficient of zero, known in the literature as "centering" (Mihaly 2009).

Comparing Similar Schools

In developing and analyzing the method by which I estimate schools to be similar to one another (the similarity index, discussed below and in Chapter 3), I intend the model as merely one among many possible bases on which to determine similarity. All that is necessary for the use of similar schools comparisons is that, for any given value-added model, school-level expected values can be extracted from the estimation technique and ordered. These expected values represent what outcome one would expect for a particular school given the inputs and controls specified in the model. On the basis of these expectations, one could then select a set of comparison schools and frame the outcome for one's school of interest against its comparison institutions. This generalized approach could be applied beyond the specific method used in this dissertation, given the modest requirement of school-level expectations.

In this dissertation, similar schools comparisons are generated using a method similar to that employed by the California Department of Education from 2000 to 2013 (California Department of Education, 2000, 2013). In California, all schools in the analysis were rank-ordered on expected outcomes. Expected outcomes were estimated by regressing school-level test scores on student's ethnicity, socioeconomic status, English language status, and mobility, as well as school-specific inputs such as the percentage of teachers who were fully credentialed, average class size, and whether schools offered year-round educational programs. Each school was then assigned a unique comparison group of 100 schools. Rather than establishing a set of fixed bands, this method ensures that each school gets the fairest comparison possible, in which the reference school's expected outcome is always the median expectation among its comparison schools. These 100 schools are the 50 schools immediately above and the 50 schools immediately below the reference school in the rank ordering. The reference school is, of course, excluded from its own comparison group. Once the comparison group is established, all schools are ranked based on observed outcomes, as opposed to expectations. The reference school is then placed alongside the outcome-ordered comparison group and scored normatively based on these comparisons.

The question of specification in these comparisons is somewhat subjective. How many schools should be included in the comparison group? What rank must a school obtain within its group to be considered significantly over- or under-performing? If the purpose of similar schools comparisons is partly to create a set of comparable institutions that is understandable to non-specialists, then very large groups should be discouraged. The advantage of a large group is likely to be greater stability in comparison groups and effectiveness estimates, but this advantage trades off against transparency. It is harder to remember 50 or 100 schools than it is to remember 30. Another problem with large groups is the ability to make comparisons at the tails. For a school to have a fair comparison to a group of size n, it must have n/2 schools below it in expectation and n/2 schools above it in expectation. If a school's expectation falls in the top or bottom n/2 of all schools in the sample—that is, if it has one of the very highest or lowest expected achievement levels in the entire state—then it is impossible to give it a truly fair comparison group of size n. As an example, it is impossible by definition for the school with the lowest expected math achievement in the state to be the median of its own comparison group. This problem happens at both extremes of expected achievement, and larger comparison groups make the problem worse by placing these extreme schools further from their group's median.[3]

Lastly, in comparing each school to its set of similar schools, what ranks should be considered as cut points for considering schools to be significantly (in-)effective? In statistical

[3] This problem is lessened as the total number of schools increases, holding constant the size of comparison groups. If one sets comparison group size at 30, then having 1000 schools total means that only 3% of schools are affected by tail problems, whereas this percentage would increase as the total number of schools declined.

inference, the null hypothesis is assumed to be true and the burden of proof lies on disproving the null. As a model becomes increasingly noisy, disproving the null becomes more difficult; differences from the mean must increase far more in order to attain a given degree of confidence. If there were no noise, then 100% of schools would be judged effective or ineffective. Given the presence of noise, it may be desirable for 40% of schools to correspond to the null (implying percentile cut points at the 30th and 70th percentiles), or perhaps even 60%. These correspond closely to current practice in teacher evaluation, in which roughly the top and bottom 30% of teachers receive a normative judgment on student learning growth while the 40% of teachers in the middle on student growth estimates are simply judged to have typical effectiveness (Diaz-Bilello & Briggs 2014). Lastly, one should take care in categorizing schools based on similar schools comparisons; terms such as "below average", "typical", and "above average" carry a more descriptive, agnostic interpretation against more causal terms such as "effective".[4]

The behavior of expectations and outcomes at the tails present a number of questions that deserve attention. First, it is probable that the absolute difference between consecutive expected values will be greater at the tails. *Prima facie*, this suggests that it would be harder for a school to move up or down in relation to its comparison group, making the chances of a school at the tails significantly outperforming or lagging its peer schools lower than would be the case at the sample mean. Yet if these schools also exhibit greater year-to-year variability in their outcomes, then ordinal changes may not be as difficult as they appear. Related to this potential is the concern that greater variability at the tails would also affect group stability, by frequently changing out the schools near

[4] There are any number of possibilities for how to frame differences between three categories. One could describe schools in the middle as "average" or "typical". Schools at the top and bottom could be labeled something like "exceeds peers" or "trails peers" to lend a looser interpretation to similar schools comparisons than is conveyed by more precise and normative terms such as "effectiveness".

the margin of each reference school's comparison group. Both these possibilities are examined in Chapters 4 and 5.

Determining Similarity Between Schools

My primary interest in this dissertation is whether similar schools comparisons can be meaningfully used to establish estimates of school effectiveness. The most important question underlying such comparisons is the basis on which similarity is determined. Generally, school characteristics estimations attempt to establish similarity between schools based on the outcomes expected given aggregate student demographics and school inputs. Although many factors may influence outcomes, controlling for these observable factors is nevertheless an important first step. Among variables that influence student outcomes, such as family/student background, community factors, and school inputs, individuals' background and family characteristics (including socioeconomic well-being and parents' education) likely explain the largest share of variation (Hoxby 2001), although there is substantial scholarly disagreement on exactly how much variation can be accounted for by non-school factors.

A well-known example of this kind of model was used by California from 2000 until 2013 (California Department of Education 2000). The School Characteristics Index, as it was known, included controls for pupil socioeconomic status, parents' highest level of education obtained, race/ethnicity, English language learner status, and mobility. In addition, the model controlled for a number of school factors: the percent of teachers fully credentialed, the percent holding emergency credentials, average class size per grade level, and whether schools operated multi-track year-round education programs.

Three criticisms can be made of this model, insofar as it is judged as an attempt to control for factors entirely or mostly beyond the control of the school. First, it possibly includes insufficient controls for student demographics. It does not consider the possibility of variation based on sex or special education status, two variables that are commonly available in state enrollment data. Whether a student is given a special education designation is partly affected by school decisions regarding the advisability of interventions required by an individualized education plan (IEP), and so is not measured with the same consistency across schools as is the case for more easily observable characteristics such as sex and race/ethnicity (Wolf et al. 2012; Winters 2013, 2014). Uneven identification of special education students may downwardly bias effectiveness estimates for schools that underidentify these students, and conversely lead to an upward bias for effectiveness when schools overidentify students as requiring special education. In the work of Wolf and Winters cited above, inconsistencies in the financial incentives for identifying students as needing special education were identified across the private and public sectors. Because my analysis is confined to traditional public schools within a single state, which face uniform identification incentives due to sharing the same funding system, this bias is likely weaker. Nevertheless, special education students have markedly lower achievement levels, and accounting for these characteristics through *some* means rather than none should help to further explain variance in school results as a function of student characteristics. Unobservable bias on this variable can in any case be addressed through setting stricter thresholds for identifying significantly effective or ineffective schools, thereby reducing the likelihood of Type I error.

Second, California's model includes measures which are perhaps expedient from a policy perspective but unlikely to help differentiate among schools based on their characteristics, particularly after controlling for demographic and socioeconomic factors. Including data on teacher credentialing and the type of programs offered at the school may help send an important signal to schools and districts on what their priorities should be, but it is unclear whether these factors should be expected to make a unique contribution toward establishing school-level expected values elsewhere.

Third, one must consider whether the inclusion of school resource inputs is advisable in estimating school effectiveness. This goes to the heart of the question of what school effectiveness means. Is effectiveness what the school accomplishes with the kids *and resources* it receives? Or should the definition exclude those resources, instead asking only what the school accomplishes given its student population, and no other factors? In Chapter 3, I follow the definition put forward in the latter question, intentionally ignoring school resource inputs. This approach, beyond the immediate purposes of this dissertation, might help shed light on the degree to which school resources mediate the relationship between school outcomes and student characteristics. Further, such underlying impacts are likely modest due to state funding equalization, which smooths spending within the state examined below. By excluding resource inputs, I intentionally arrive at a definition of similarity which considers only student demographics, as well as school size.

The expectations which result from the similarity index I develop are thus intended to represent an expected level of achievement given only student characteristics and school size. Naively, one could interpret the residual in such a model—the difference between actual and expected achievement—as a school's effectiveness given its demographics, although such an interpretation could be subject to criticism. As shown in Chapter 3, I instead use expectations to create unique comparison groups for each school. The combination of demographic-based expectations with such comparisons is not coincidental: my similarity index, unlike any tool available under the other VAMs considered here, yields a range of expected outcomes that are meaningfully different from one another, while also exhibiting relative stability over time, insofar as the demographics on which the index is based tend to change more slowly than do other school factors and outcomes. For the sake of public usefulness and acceptability, both these factors are important; if similar schools comparisons were random or highly volatile they would not be capable of earning the trust of educators and the public. This is the primary reason that my index of similarity, in comparison to a (noisier) growth-oriented model, may be useful as a basis for determining similarities on which comparisons can be made.

Contemporaneous VAMs

Attempts to compare schools or teachers apart from estimates of student learning growth may be discounted or ignored in value-added modeling because the idea of a school's or teacher's contribution to student learning (the intent of any VAM) intuitively leads to attempts to account for student achievement prior to a student being exposed to the intervention, whether understood as the school or the teacher. In the case of teachers, which a student typically has for one year, the prior is straightforward, namely, a student's most recent test score before the current school year. However, in the matter of schools, the prior year's test score may itself have taken place in the school under analysis. While this control may be desirable if the interpretation sought is a school's contribution to learning in the most recent year, doing this obscures the question of what a student's achievement would look like had he or she attended elsewhere over the full period since entering the school.[5] Further, schools serving grades at or below the first grade in which students are tested may have no counterfactual at all. That is, no baseline test score exists prior to the student entering the school. For a school serving students from the lowest tested grade, a VAM which relies on prior test scores for estimation will necessarily exclude the grade, restricting estimation to non-baseline grades. Yet baseline scores partly represent the school's contribution to student learning in the year of

[5] This description assumes that states test students in the spring. If fall testing is the norm, then the tests given in the lowest tested grade in a school may qualify as a true baseline.

instruction leading up to the first test. In summary, growth models in which the prior year is included among the model's regressors may, depending on the grade structure of the school and the question one is asking, partly control for the intervention itself. If the period of interest is only the most recent year, then growth models are as applicable in the case of schools as with teachers and one would not be controlling for the intervention. However, if one is interested in the cumulative contribution of a school to student achievement, then including the prior year's achievement on the right-hand side will control for part of the intervention toward which the question is directed, namely, the school's contribution to student learning in the period(s) prior to t-1.

One estimation technique that could get around the threat of partly controlling for the intervention at school level is to examine only students who switch schools due to mobility or choice (Harris 2009), but the degree to which these estimates may be representative of a school's typical effectiveness for all students is questionable. In addition to overall representativeness, there may be selection bias in student switching, in which case students switch into or out of schools for systematically different reasons given the characteristics of a school and the schools with which it is exchanging students.

In the most general sense then, two possibilities exist for estimating school effectiveness: modeling changes in student achievement, or levels in student achievement. The most common problem with using levels of student achievement, rather than yearly changes as discussed above, is that they do not account for two very common problems in statistical inference: the existence of unobserved student fixed effects which influence school outcomes but also are beyond the control of schools, and the selection of students into schools on unobserved but quality-relevant factors such as student and parental motivation. As shown in Harris (2009, p. 103), the chief concern is that such student-level unobservables might translate into systematic bias at the school level, which would occur if student and family sorting into schools on unobservables was clustered (i.e.

nonrandom). Selection on unobservables, an instance of omitted variable bias, motivates the use of both fixed effects and student growth (methodologically, an instance of differencing) to accurately estimate schools' and teachers' contribution to student learning, although each method still may be slightly biased in opposite directions (Angrist 2009). The relationship between these sorts of models and the results obtained from various growth-oriented models is perhaps best illustrated in Goldhaber et al. (2013), which compares five types of VAMs controlling for prior scores and student characteristics in different ways. Though they find high correlations between different VAM specifications at the teacher level, they wisely point out that model choice can still generate meaningfully different rankings of teacher effectiveness, advising that any such sensitivities and differences are simply made very clear in the development and evaluation process. Yet by always including at least one prior score among their regressors, all such growth VAMs at least partly account for unobservables which purely cross-sectional, level-oriented VAMs cannot address.

Conversely, modeling effectiveness based on levels in student achievement has two advantages. First, insofar as effectiveness estimates are interpreted as the difference between schools' (or students') expected achievement based on fixed and observable characteristics and their actual achievement, they account for cumulative growth from the point at which students began receiving the intervention—whether a particular school, or the full history of schooling received by a given population. Second, and closely related to the consideration of cumulative growth rather than annual growth, one should expect that the variability of contemporaneous level-based estimates should be less than estimates based on yearly student growth given regression to the mean (Harris 2009, p. 95). Above-trend growth for a student in one year increases the likelihood of below-trend growth in the following year; annual growth estimates derive some of their inherent noisiness from this phenomenon. Cumulative growth is more stable due to these errors cancelling each other out to some degree over multiple years, and it is from this likely phenomenon that level-based models should be expected to derive their stability relative to growth models.

Whether one should prefer level- or growth-oriented VAMs depends somewhat on the degree to which one addresses their relative strengths and weaknesses. That is, if a level-oriented VAM can account for individual fixed effects beyond what demographics or school characteristics would allow, then its chief weakness would be addressed and it would be preferable. Alternatively, growth-oriented VAMs would become preferable in cases where the intervention is not endogenous, i.e. only measured in the dependent variable, or cases in which their annual variability is reduced through smoothing or the use of multiple prior test scores.

Marginal Incentives

Orderings of schools along a single range of expectations have the possible advantage of promoting marginal improvement for schools equally across the full range of performance. Schools respond to incentives, particularly when those incentives are highly visible, easily understood, and sharply discontinuous. Reback (2008) found that schools in Texas in the 1990s responded to incentives to pass students on minimum competency examinations. Minimum competency exams, like the more common proficiency-focused exams that have proliferated since the early 2000s, incentivize schools to improve the performance of students who are on or just below the margin of receiving a passing or proficient score on the test. Reback's analysis found that students in classrooms with many students near the passing cut point benefited even when those students weren't themselves on the margin, suggesting perhaps that greater attention or more effective instructors are allocated toward marginal students on a classroom basis. Booher-Jennings (2005) documented the now well-known phenomenon of "bubble kids" or "educational triage" anecdotally familiar to educators since No Child Left Behind. She found that students outside the margin of proficiency did not benefit from the shift to focusing on proficiency. Even though these tests have included at least four performance levels, and thus at least three cut points, the only cut points that appeared to spur observable student progress were those tied to schools' overall accountability ratings, namely, the boundary between students being deemed proficient and not proficient. Because the cut points above and below proficiency *by definition* didn't count toward proficiency rates, they didn't induce significant student gains.

Rothstein (2008) discussed the political manipulations to which proficiency thresholds and definitions are subject. Anchored to the NAEP to allow cross-state comparisons, the differences between the difficulty of attaining 'proficiency' in one state often differs greatly from 'proficiency' in another state. The fact of this manipulation brings into question the coherence of the definition of proficiency (Loveless 2012) and thus the rationale behind placing a great deal of emphasis upon a single achievement level. Admittedly, any evidence of nonlinearities in the marginal value of higher test scores near proficiency cut points could justify such an emphasis given a desire to maximize aggregate outcomes, but no such evidence exists.

Very simple, fixed goals such as minimum competency or proficiency can be problematic in two ways. First, they neglect the importance of student achievement at levels far from the threshold. This is a problem if the underlying assessment is intended to accurately describe and incentivize higher performance for all students. Tests focused around a single cut point tend to have low standard errors near the cut point, but much higher standard errors of measurement elsewhere. Tests may also be subject to floor or ceiling effects. These both confound the accuracy and validity of student test scores at most points in the distribution. Second, schools face no explicit incentive to improve test scores for high- and low-achieving students outside the cut point margin.

The phenomena described by Reback (2008) and Booher-Jennings (2005) are not necessarily bad things. It is difficult to know what student learning growth would look like absent a fixed goal like proficiency. But Neal (2010) provides a rationale by which one should expect that such incentives are unlikely to maximize aggregate student outcomes in comparison to other possible policy mechanisms.

Neal (2010) argues that accountability mechanisms focusing on what he calls 'efficiency' are likelier to maximize aggregate learning outcomes than mechanisms focused on explicitly defined absolute goals such as the original goal of 100% proficiency for all students in all schools put forth under No Child Left Behind. Specifically, he argues that every school should face an incentive for marginal improvement which is within their reach yet still requires collective effort. Annual Measurable Objectives (AMOs), the most common form of performance targets faced by schools, may fail this condition when schools' targets are either so far above current performance that they cannot feasibly be met, or so modest that schools can take meeting targets for granted. This phenomenon is similar to the problem of educational triage described above in that it derives from the setting of absolute criteria and targets, but differs in that the effect occurs across schools, rather than within in the case of triage and 'bubble kids'. Drawing upon the research of Holmstrom and Milgrom (1991) into multitask principal-agent problems in economics, Neal argues for policy mechanisms that facilitate normative comparisons between schools against which every school has a reasonable range of outcomes depending on effort and effectiveness. The development of similar schools methods is thus motivated by the desirable motivating effects of such comparisons in relation to more traditional methods of target-setting such as AMOs.

Summary

The literature reviewed above is drawn from a range of sources which converge to recommend closer investigation of similar schools comparisons as a promising direction of inquiry both for value-added modeling and for the use of such comparisons in educational evaluation and practice. There remains an urgent need to develop VAMs which are both valid and transparent, as most rigorous scholarship has focused so far on the former to the neglect of the latter; in the extreme case, the model developed from the work of Sanders (1994; 1996; 1998) is sufficiently opaque to prevent replication, which allows its use to be limited by the company that owns it. Similar schools methods constitute an effort in the opposite direction, prioritizing transparency and currency while preferably retaining validity. A value-added model which explicitly renders comparisons between schools can leverage and guide comparisons which communities and educators already frequently make in ways that largely remain merely instinctive, not to mention blunt.

The facilitation of comparisons aligns naturally with the use of visible and public knowledge in formally determining similarity, insofar as one hopes to align comparisons with perceptions which, right or wrong, are not easily contradicted by abstract measures. For this reason, so-called 'snapshot' (Harris 2009), level-oriented models may be preferable. Most of the reasons for their dismissal in teacher evaluation are mitigated when considering schools, and lingering concerns regarding the validity or fairness of such models should be considered alongside the expectation that at the same time they are likely to be preferable to growth-oriented models on the basis of stability and transparency.

Chapter 3: Data and Methods

Following the methods and concerns addressed in Chapter 2, in this chapter I explicitly develop a similar schools model alongside three other VAMs for comparison. To do this, I proceed as follows. After reviewing the major research questions of this dissertation, I use and describe student test scores and demographic data from a single southern state for the years 2009 to 2014. Next, I discuss the proper use of the state assessments for value-added modeling given their technical limitations. I briefly discuss the rationale behind the use of multiple years of student data within each model, as well as the usefulness of multiple years of school estimates. I then describe in detail the methods used to develop each of the four models presented in this paper: Similar Schools Comparisons (SSC), Mean Prior Z (MPZ), Student Growth Percentiles (SGP), and Student Value Added (SVA). Finally, I describe the analytic strategies I will use in Chapter 4 to compare the models on desired properties, ideally in a way that can be framed and communicated for public use.

Research Questions

The purpose of my quantitative study is to investigate whether a similar schools model, which I develop as a particular but not exhaustive example of similar schools comparisons, can substantially replicate the results of common methods employed in estimating schools' value-added, and if so, under which conditions. The following research questions provided a frame for the methods used in this study.

(1) What are the statistical properties (stability, validity, and statistical significance with respect to demographics) of the Similar Schools Comparisons (SSC) model in comparison to other representative measures which could be used to examine school effectiveness?

(2) What are the characteristics of schools for which the Similar Schools Comparison

(SSC) model produces results which disagree significantly with the results of the two growth-oriented value-added models (Student Growth Percentiles and Student Value Added), as well as Mean Prior Z (MPZ)?

(3) To what degree are models likely to be understandable and meaningful to the public, independently of their statistical properties?

Context

I use data obtained from an anonymous southern state on public school students from 2009 to 2014. These data were obtained from that state's education agency. The data included are for tested students in mathematics who also have available demographic data or prior test scores. The assessment instrument used is the state's Augmented Benchmark exam, which is administered annually to all public school students in grades 3-8 in literacy and mathematics, as well as in grades 5 and 7 for science. The Augmented Benchmark is part of the the state's primary accountability law, which has been in place in the state since 1999.

Participants

Participants for this study included students with a valid mathematics test score on the state-mandated assessment in grades 6-8 for the years 2011-2014. Test scores extending back to 2009 are used only for VAM estimation, and are not used to determine the sample. Table 2 presents statistics on the students and test scores used in my analysis based on the above criteria: public school students tested in math on the state Augmented Benchmark assessment in grades 6-8 from 2009 to 2014. A majority of students are low-income, with nearly 60% being eligible for free or reduced-price lunches. Twenty-one percent of students are African-American, while 9% identify as Hispanic. Sixty-six percent of students identify as non-Hispanic white, leaving roughly 4% of students (not shown) who identify as either Asian, Hawaii/Pacific Islander, or Native American. Just under 10% of students are designated as receiving special education services, while about 6% of students are identified as English language learners. Fifty-one percent of students are male.

Table 2. Characteristics of Tested Students in Grades 6-8, 2009-14

Demographic	Frequency	Percentage
Free/Reduced Lunch (gr 6-8, 2009-14)	356,002	59.60%
African American (gr 6-8, 2009-14)	126,106	21.10%
Hispanic (gr 6-8, 2009-14)	56,346	9.40%
White (gr 6-8, 2009-14)	394,102	66.00%
Male (gr 6-8, 2009-14)	303,511	50.80%
Special Education (gr 6-8, 2009-14)	57,556	9.60%
English Lang. Learner (6-8, 2009-14)	36,561	6.10%
Grade 6 (2009-14)	200,533	33.57%
Grade 7 (2009-14)	199,873	33.46%
Grade 8 (2009-14)	196,945	32.97%
Total Number of Math Scores (N)	**597,351**	**100.0%**

Only students who attend a school for the full academic year are included. In addition to aligning with the state's test-based accountability methods, this exclusion removes a potential confound in estimation. Schools with higher rates of student mobility, whether students moving in or out, face challenges not otherwise captured in student characteristics. Going back to the core question of what should be included here in the definition of school effectiveness, the frequency with which students enter or leave a school is mostly a concern over which schools have little control. While school-induced attrition or attraction are possible, particularly in settings with a high degree of choice, mobility for most students in most public schools is more frequently the result of family and economic circumstances. Excluding mobile students from modeling removes the most obvious confound (the student himself), but does not account for possible peer effects of high student mobility on non-mobile students within schools. These effects remain as a potential confound.

Table 3. Schools Included in Study, 2009-2014

Year	Frequency
2009	559
2010	564
2011	561
2012	560
2013	547
2014	543

Table 3 shows the number of schools from which the students in grades 6-8 (described in Table 1) are drawn in each year used for modeling. The number of schools in the state serving students in this grade range is fairly steady between 543 and 564 over the period.

Assessments: Choice and Proper Use

The purpose of my analysis is to compare the results and statistical properties of Similar Schools Comparisons to three comparison value-added models (VAMs) for estimating school effectiveness. Because the analysis depends on test scores, idiosyncrasies and artifacts in the scores themselves can obscure a fair comparison of the models. Therefore attention must be paid to the choice of grades and subjects to include in the analysis. For reasons explained below, my analysis focuses on math test scores on the state's Benchmark assessment in grades 6-8 over the years 2009-2014. Literacy test scores are employed only as a conditioning variable in one of the models (SVA), never as an outcome.

The assessments used in the analysis are vertically moderated, but not vertically equated (Lissitz & Huynh 2003). The former serves a different purpose from the latter, which is typically preferable for student growth modeling along a continuous range. A criterion-referenced assessment is vertically moderated when, for example, its performance cut points are linked to specific cut points within the same subject across grades to allow quasi-interval interpretations in those specific cases. However, this property does not imply that scale score changes in general represent equal intervals of learning across grades and assessments, which is what is denoted by vertical equating (Patz 2007). This can be taken into account in value-added modeling, typically by either standardizing scale scores within grade, year, and/or subject, as well as including grade-specific dummy variables as a control in value-added estimation. Perhaps more difficult to account for are differing demographics by grade, which when coupled with between-grade differences in student achievement can confound estimates of the impact of demographics on student outcomes. If a given demographic characteristic is more common in one grade than another, and performance levels differ systematically by grade, then the impact of that demographic will necessarily be confounded with grade impacts.[6] Additionally, some demographic variables are not measured consistently across grades; rates of students receiving free or reduced price lunches, for instance, are well known to decline as students get older, despite there being no obvious reason to believe that teenagers face lower levels of economic disadvantage than children in elementary school. This is most likely due to eligible students frequently refusing to accept the offer of cheaper lunches, perhaps for reasons relating to social stigma or a desire for self-sufficiency.

To disentangle possible grade-level effects from the impact of particular demographics, I confine my analysis to students in grades 6-8 in addition to standardizing scale scores and including grade dummy variables in estimations. Students in grades 6-8 are tested yearly in math and literacy. Since testing begins at grade three, all grade levels have prior test scores which permit growth

[6] There are two main ways this problem can be addressed. The first, employed here, is to standardize test scores by grade and subject, so that the mean score in each grade is zero by definition and criterion-based judgments such as proficiency are removed from the analysis. The second method would be to include grade-level dummy variables in any regression-based estimation of school effects.

modeling, so the only students who lack a prior score are those who were not enrolled and tested in public schools within the state in the prior school year.

Literacy test scores are rejected as a dependent variable in modeling due to the non-normality of test score distributions, which constitute a violation of the assumptions required by the VAMs relying on ordinary least squares (OLS) regression. Thus the analysis is confined to school effectiveness estimates for math only. Figures 1-6 below show the distribution of ACTAAP scale scores in math and literacy for grades six, seven, and eight in 2014. Literacy scores have a non-normal distribution, due to a sharp ceiling in scale scores which becomes more significant in each grade, with eighth grade scores showing the greatest number of students near the ceiling. Math scores, on the other hand, meet the assumption of normality required in some of the models used here. The local irregularities which appear in the middle of the distribution in these histograms are likely due to fixed bin sizes including different numbers of actual scale scores; not all integers in the range shown correspond to possible scale scores, so it is possible that some bins contain more possible score values than others.

Figure 1. Grade 6 Literacy

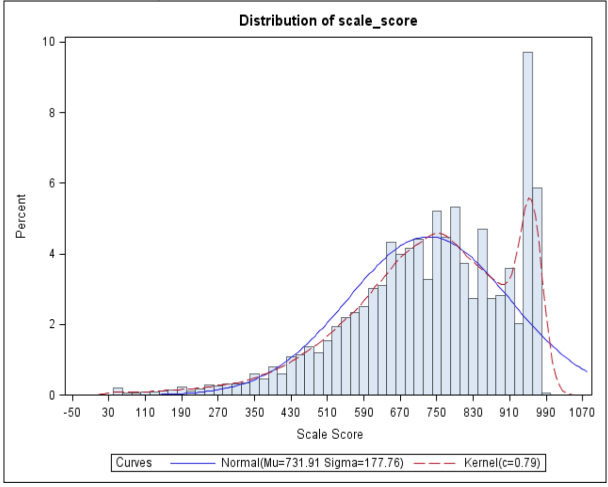


Figure 2. Grade 6 Math

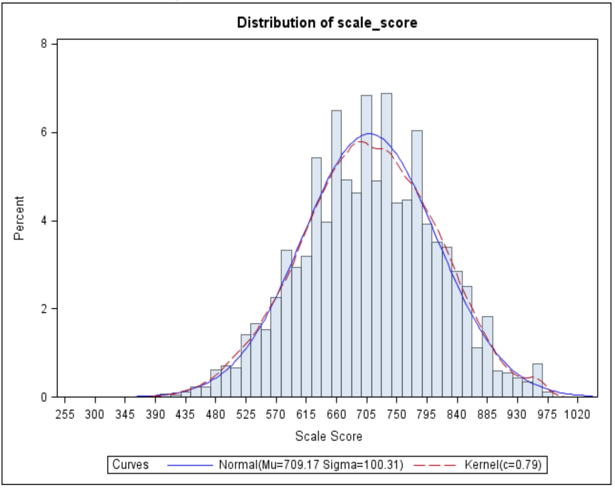


Figure 3. Grade 7 Literacy

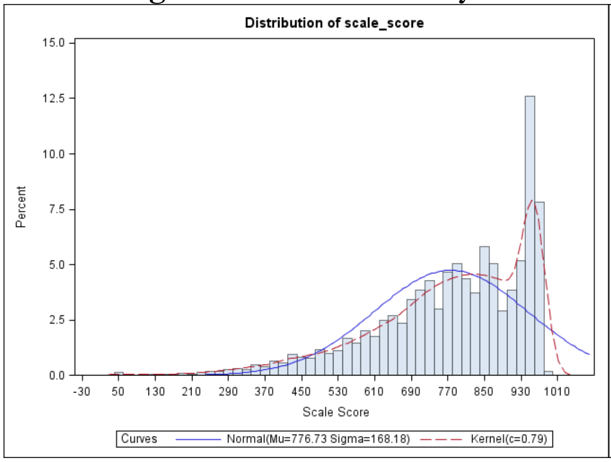


Figure 4. Grade 7 Math

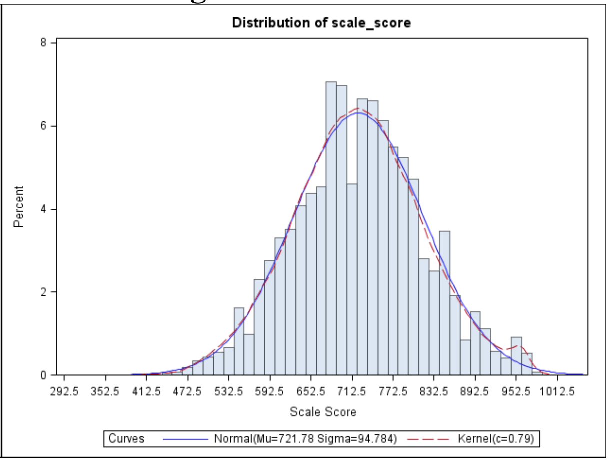


Figure 5. Grade 8 Literacy

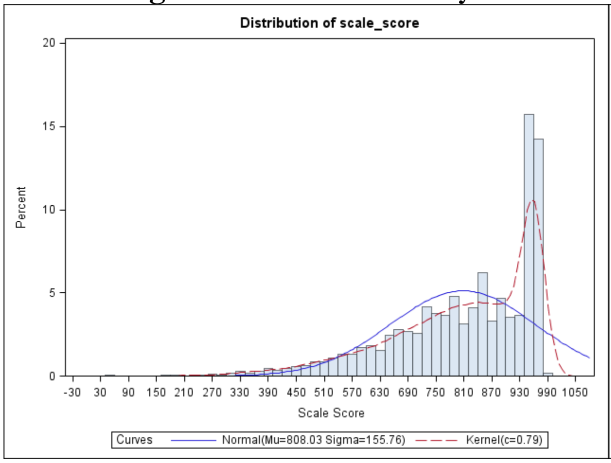
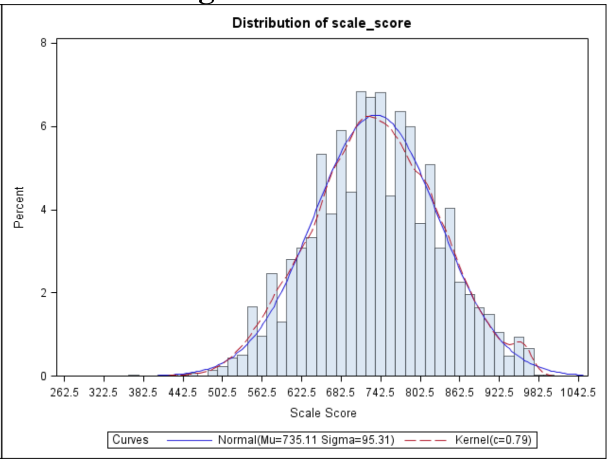


Figure 6. Grade 8 Math



Role of Multiple Years of Data in Value-Added Modeling

Multiple years of data are used in two ways. First, in models which allow one or more years of prior data for estimation, I allow for as many years of prior data as are possible given the model and the year for which an estimate is sought. The full panel of data, extending from 2009-14, allows for as many as five prior years of student data in the case of student growth percentiles (SGP) and Mean Prior Z (MPZ), the methodology of which permits as many prior within-subject test scores as are available. Student value-added (SVA) never uses more than two prior years of student test scores, and Similar Schools Comparisons (SSC) makes no use of prior test scores.

Second, models may be smoothed by averaging school estimates over more than one year, to provide greater stability to school estimates. While doing this is likely to improve the stability of school estimates over time, it nevertheless is peripheral to considering the relative stability of different school effectiveness models. If one measure is more stable than another when comparing changes in one-year estimates, then it is likely, though not certain, that the same relative stability will obtain when comparing changes in multi-year estimates. For this reason, the consideration of three-year averages is confined to the discussion of improvements and extensions in Chapter 5.

Thus the findings presented in Chapter 4 represent single-year school estimates under each of the four models defined below: Similar Schools Comparisons (SSC), Mean Prior Z (MPZ), Student Growth Percentiles (SGP), and Student Value Added (SVA). Results for the models are presented over the years 2011-2014, so that years are pooled in presenting estimates. Although the student data used for modeling in this dissertation extend back to 2009, no school estimates are made prior to 2011, thereby allowing for SGP and SVA to rely upon no fewer than two years of prior data for modeling. Consistent with this restriction for SGP and SVA, the SSC model, based on the school groupings derived from the similarity index, only employs data from 2011 forward since it does not consider prior test scores.

Similar Schools Comparisons

The SSC model has two components that are used to describe school performance relative to schools with similar characteristics. The first component of the Similar Schools Comparison model is an index of similarity, defined as a function of a school's student demographics in which the weight given to different demographic characteristics is proportional to their typical influence on school outcomes. The second component of the Similar Schools Comparisons model is a mechanism for comparing schools with similar expected outcomes on the similarity index.

I begin by describing the first component. This method compares schools to a fixed number of comparison schools as determined by nearby ranked values on the similarity index. Every school is assigned a unique comparison group consisting of the fifteen schools ranked immediately above it in terms of school demographic characteristics, and the fifteen schools ranked immediately below it. The school under consideration, which is denoted as the *reference* school, is then ranked against its 30 comparison schools on actual outcomes, as opposed to expected outcomes based on school demographics generated by the similarity index.

I model the similarity index as follows, using a method similar to that employed by the California Department of Education from 2000 to 2013 (California Department of Education, 2000, 2013). The index uses ordinary least squares with schools' mean standardized math scores as the dependent variable to estimate the relationship between math achievement and school demographic percentages as well as school size. Statistically significant coefficient estimates are then used with schools' demographic percentages to estimate expected outcomes for schools, which are the result of interest. Formally, the similarity index is relatively simple:

$$Y_{it} = \boldsymbol{\beta X}_{it} + \beta \gamma_{it} + \varepsilon_{it}$$

The dependent variable Y_{it} is standardized scale (*z*-) scores in math for students in grades six through eight. To stabilize estimates, I use three years of pooled data. I regress student z-scores in math on a vector $\boldsymbol{X}_{it}$ of school demographics including free/reduced lunch status, race/ethnicity, sex, special education status, and English language learner status. At the school level, each of these are represented as a simple percentage, and every variable but race/ethnicity is binary at the student level. I exclude schools' percentages of white students from the estimation, so that schools' share of these students are the reference group against which parameters are estimated for other race/ethnicity indicators.[7] Finally, school size γ_{it} is included as a regressor.

All schools in the analysis sample are rank-ordered on expected outcomes as determined by the similarity index. Once this ordering is obtained, each school is assigned a unique comparison group of 30 schools. Rather than establishing a set of fixed bands, this method ensures that each school is always the median expectation among its comparison schools.[8] These 30 schools are the schools immediately above and below the reference school in the rank ordering (15 of each). The reference school is, of course, excluded from its own comparison group.

[7] I also tested for interaction effects between these demographic variables at the student level, but found that the inclusion of demographic interactions did not substantially improve the ability of the model to explain variation in schools' mean math scores. For this reason interactions are not included in the final model.

[8] An exception to this occurs at the tails of the similarity index, in cases for which there are fewer than 15 schools above or below a reference school on the similarity index. These schools are still ranked against the 30 schools with closest similarity index rankings, but their comparison groups necessarily have unequal numbers of schools above or below the reference school on the similarity index.

Once the comparison group is established, all 30 schools are ranked based on *observed* outcomes. The reference school is then placed alongside the outcome-ranked comparison group. If the reference school outcome equals or exceeds the outcome of the sixth-highest school in the group, then it is considered to be exceeding its peers. At the other extreme, if the reference school's outcome equals or falls below the outcome of the sixth-lowest (twenty-fifth highest) school in the group, then it is considered to be trailing its peers. Cut points at the sixth and twenty-fifth schools in the rank-ordering of 30 school outcomes correspond to the 20th and 80th percentiles. If the reference school falls in the range between the sixth and twenty-fifth schools, then it is judged to have an outcome that is typical of its peers.

The 20th and 80th percentiles are used as conservative cut points for labeling schools as "exceeding" or "trailing" peers due to the presumption of the null hypothesis and the presence of statistical noise. These thresholds are more conservative than much of current practice in teacher evaluation, in which roughly the top and bottom 30% of teachers receive a normative judgment on student learning growth while the 40% of teachers in the middle on student growth estimates are simply judged to have typical effectiveness (Diaz-Bilello & Briggs 2014).

Student Growth Percentiles

As a comparison for the SSC model, I estimate student growth percentiles (Betebenner 2008, 2011; Wright 2010), which I then aggregate into school-level medians. Student growth percentiles (SGPs) compare a student's year-to-year test score changes within a subject to other students scoring at or near the same level in the prior year. These results are then translated to school-level measures by simply computing the median growth percentile for all students tested in a subject in a given year at each school.

The methodology developed for estimating growth percentiles uses as many prior test scores in the same subject as are available. Thus an eighth-grader who has tested in every grade in the same state has five prior scores, going back to third grade, and Betebenner's methodology, as implemented here, incorporates all five priors. SGPs are based firstly on estimating conditional density given a student's test score in the most recent year prior to the period of analysis, using all prior test scores for conditioning. A student's growth percentile is thus the percentile at which the student scores in the latest year within the conditional density (i.e. other test scores closest to a student's prior scores). In simpler language, a student's growth is considered only normatively in comparison to other students at the same point previously. Thus a very low-achieving student is compared only to a precise set of other low-achieving students, and the possibility of such a student having a high growth percentile is similar to that for high-scoring students.

The model specifications for SGP used here are those recommended as defaults by Betebenner, and discussed at length in the previous chapter. Taken together, these four features—estimation of quantiles, the use of cubic splines, the inclusion of multiple prior scores, and the use of median regression rather than least squares—yield a model that is intended to be descriptively useful given a minimal set of assumptions about assessment properties.

Student Value-Added

The third model I use is what I am calling student value-added (SVA), which is a parsimonious characterization of students' test score gains as a function of previous scores in both the subject of analysis as well as another subject as an additional control for student-level trend. This model implicitly accounts for student characteristics influencing expected growth through the inclusion of multiple prior years of data in multiple subjects, as well as the allowance for student-level random effects. It is distinct from SGP in three major ways. First, it imposes linearity on the effect of prior test scores on current achievement, which the nonparametric SGP relaxes. Second, it allows for student-level random effects, which SGPs do not consider. Third, it introduces off-subject test scores as an additional control in an attempt to capture individuals' characteristics as conveyed through test score trends. Because the analysis which follows focuses on math test scores, the off-subject test scores are drawn from literacy tests.

The impact of demographics is further accounted for by the use of student fixed or random effects. Random effects require a stricter set of conditions in order to be applied, but if those conditions are met, it can be shown that random effects estimation is more efficient than fixed effects, which use a greater number of degrees of freedom in the model. In the case of these data, random effects are used once it can be shown that the necessary conditions are met.

The estimation thus employed at the student level for SVA is as follows:

$$Y_{it} = \beta_1 Y_{i,t-1} + \beta_2 Y_{i,t-2} + \alpha_1 X_{i,t-1} + \alpha_2 X_{i,t-2} + (u_{0i} + \varepsilon_{it})$$

where Y_{it} is the math score of student I at time t, $Y_{i,t-1}$ and $Y_{i,t-2}$ are the math scores of student I at time t-1 and t-2 respectively, and $X_{i,t-1}$ and $X_{i,t-2}$ are the reading scores of student I at time t-1 and t-2. The individual-level random effect u_{0i} is absorbed in the error term with ε_{it}. A set of school indicator variables are further included at the second level in the model for the second level of estimation. The coefficient estimates on each of these schools is thus interpreted as the contribution made by the school toward explaining additional variation in outcomes Y_{it} at the first level.

Mean Prior Z

Mean Prior Z (MPZ) conditions current-year standardized math scores (z) on mean z-scores in math prior to students' entry into their current schools. In the case of students who do not move schools between grades 6-8, a student's mean includes math achievement in grades 3-5.[9] If a student moves, then his or her prior scores up to the time at which she moved are included. In short, the model takes as many math test scores as are available on the state assessment for each student prior to entry into the school attended at time *t*, and calculates the arithmetic mean of those scores. Once each student's mean is calculated, then school-level arithmetic means are calculated from all students with prior test scores. Schools are then ranked by mean prior z-score, and these rankings are treated identically to similarity index rankings in the Similar Schools Comparisons model; each school is compared to the 30 schools nearest it based on mean prior z-score, and then ranked against those schools based on actual z-scores at time *t*. MPZ thus replicates the comparison mechanism used in Similar Schools Comparisons, while differing from it by using only student-level prior test scores as a control rather than school-level demographics.

Relationships Between Models

The four models considered for analysis in this dissertation can be understood in relation to each other in the following way. SGP represents a pure growth approach to value-added modeling; at the other extreme, Similar Schools Comparisons, relying upon the similarity index, represent a snapshot (or cross-sectional) approach to value-added modeling, by not including any controls for prior achievement at the school or student level. SGP strictly ignores student background characteristics, examining only normative growth patterns; similar schools, through its reliance on

[9] The model uses as many prior scores as are available in grades 3-5; if a student has only one score in that grade range, then that single score is considered the prior mean.

the similarity index, is concerned *exclusively* with student characteristics to the degree those characteristics influence outcomes. Conceptually, the SVA represents a middle ground between SGP and the similarity index used in the SSC model in one sense, namely the importance it ascribes to student characteristics. It implicitly conditions expected growth on individuals' background characteristics (both observed and unobserved) through the use of multiple prior scores in multiple subjects, and then seeks to estimate schools' contribution toward explaining variance in outcomes not otherwise explained by individual background. Like SGP, it is growth-focused; like the SSC, it accommodates background differences, although it only does so implicitly, by capturing demographic influence through levels and trends in prior test scores. Thus the degree to which it appears to split differences between similar schools estimates and SGP in Chapter 4 is of great interest.

Analytic Strategy

I analyze the characteristics and results of the models outlined above using the following statistical analyses:

- Correlations of model estimates to school characteristics
- OLS regression of model estimates on school characteristics
- Within-model correlations of school estimates from year to year
- Classification consistency among models as determined by quintiles
- T-tests on differences in school characteristics in cases of high disagreement between models.

These analyses are done within the conceptual framework provided by Polikoff (2013) of desirable properties of school accountability systems.

First, to determine the relationship of each of the four models to demographics, I examine correlations of model results with school demographics, including race/ethnicity, socioeconomic status, sex, special education, and students designated as English language learners (ELL), as well as with unconditioned mean standardized math scores. The VAM results are then regressed on these same demographic variables, both to determine which variables show a statistically significant relationship with the models' results as well as to estimate and compare the amount of variance in each model that can be explained by the combination of all available demographics. Together, these analyses provide insight into the *fairness* of the models with respect to demographics.

After examining the relationship between the models and demographics, I consider the stability of the models over time. I do so in two ways. First, I examine correlations between school's current year and prior year estimates under each of the models, as well as on math z-scores. These correlations, while always representing consecutive years, will be pooled across multiple consecutive-year pairs between 2011 and 2014. Second, using the same set of pooled estimates over the same period, I examine the prior year's distribution of the top quintile of schools in each model in the current year. I repeat this analysis for the bottom quintile of schools.

Once the question of stability has been treated, I examine agreement between the models on the distribution of school effectiveness. I begin this analysis with a simple examination of correlation coefficients. In the case of comparisons not involving similar schools, the coefficient in question is the Pearson coefficient. Where similar schools are involved, the correlation used is the Spearman rank correlation coefficient. After examining correlations, I then turn to distributional comparisons between the models, in the manner used recently by Stuit et al. (2014), comparing quintiles on each of the models to each other. In other words, I examine the distribution of the top quintile of schools in one model, by quintile on another model, then do the same for the second quintile, and so forth. Each of the four VAMs is compared to the others in this way.

The degree of agreement between the four models can be understood in two ways. First, if the models all largely convey the same information on school effectiveness despite different methods, then one may prefer the models that are more transparent and easily understood. Second, to the degree that models disagree, one must inquire into the source of that disagreement and, as far as possible, prefer one or another in accordance with the properties most desired of a particular VAM.

Of particular interest in any distributional comparison is the frequency of cases of extreme disagreement, as well as cases of very close agreement. For this reason, I summarize the results implicit in the quintile comparison just discussed by examining the percent of all cases where two models place schools in the same quintile, as well as the percent of all cases where two models place schools in extreme quintiles (e.g. a school is top quintile in one model but bottom in the other, or vice versa). I define extreme disagreement as a quintile difference of at least three quintiles for the same school in two different models. Where possible, I also examine agreement between models by comparing schools based on their identification as significantly effective or ineffective. This is only possible for SVA and similar schools, as data limitations at the time of modeling did not allow for estimations of error around schools' median growth percentiles. While the cutoffs for significance in similar schools comparisons are conservative and subject to debate, they nevertheless represent plausible values at which such determinations might be made in policy.

Finally, I investigate sources of variation between models by comparing school characteristics across three categories: schools which fall at least three quintiles higher in model X than model Y, those at least three quintiles lower in model X, and schools for which the models yield disagreement of less than three quintiles. I conduct *t*-tests on the differences to test for statistical significance.

The methods and analytic strategies outlined above are intended as an approximation to questions which are vital to policymaking and school accountability: how to fairly evaluate schools, how to make such evaluations accessible and meaningful to non-specialists, and how to encourage all schools toward feasible improvements in achievement that align meaningfully with the stated goals of education and are not merely the result of gaming or compliance. Considering all these questions, the integrity with which available data can be used to facilitate fair and valid comparisons between similar schools is the overarching aim toward which each of these methods is directed. In the following chapter, I present the analysis of similar schools which should serve as a tentative answer to these questions.

Chapter 4: Results

Overview of Questions & Analytic Strategies

I am primarily interested in four questions with regard to properties of models aimed at evaluating school effectiveness based on student test scores. These four questions correspond to the four criteria borrowed from Polikoff (2013) as discussed in the previous chapter: validity, reliability, fairness, and transparency. In this chapter, I evaluate the first three of these criteria, leaving a discussion of transparency to Chapter 5. Fairness is considered first by examining the relationship of the models to observable school characteristics, namely student demographics and school size. Reliability is addressed by examining the stability of the models over time, in comparison to each other as well as unconditioned mean test scores at the school level. Validity is treated by examining the degree of agreement and disagreement between models. After determining the degree of agreement, I then investigate the nature of disagreement between models by examining the observable characteristics of schools for which the models produce very different estimates. Transparency is treated in Chapter 5 because it is not directly subject to quantitative analysis, which is the focus of the current chapter.

Below I present my analysis on four main models presented in Chapter 3: Similar Schools Comparisons (SSC), Student Growth Percentiles (SGP), Student Value Added (SVA), and Mean Prior Z (MPZ).

Fairness: Relationship of Models to Demographics

Table 4 provides correlations with observable student demographics for five different measures. The first four are the VAMs discussed in Chapter 3: Student Growth Percentiles (SGP), Student Value Added (SVA), Similar Schools Comparisons (SSC), and Mean Prior Z (MPZ). In addition, I compare the four VAMs to simple school-level standardized performance (*z*-scores) in math. Student growth percentiles range from 1 to 100, student value-added ranges mostly between -1.0 and +1.0, and similar schools ranks range between +15 (top rank) and -15 (bottom rank), in accordance with the fact that each school has 30 comparison schools.

Table 4. Correlations of Model Estimates with Student Demographics, 2011-14

Demographic	Student Growth Percentiles	Student Value Added	Similar Schools Comparisons	Mean Prior Z	Math Performance (z)
Male (%)	-0.0232	0.0098	0.0081	0.0180	-0.1598
Hispanic (%)	0.1107	0.0910	0.0233	-0.1240	-0.0014
African American (%)	-0.1277	-0.1409	-0.0173	0.1471	-0.5451
White (%)	0.0561	0.0806	0.0039	-0.0636	0.4920
Free/Red. Lunch (%)	-0.1360	-0.1402	-0.0494	0.1725	-0.5901
Special Educ. (%)	-0.0894	-0.0254	-0.0760	0.1488	-0.4225
Eng. Lang. Learner (%)	0.1164	0.0927	0.0237	-0.1298	-0.0254
Math Mean Prior Yr (z)	0.2661	0.2735	-0.4575	-0.3769	0.8349

As can be seen, the highest correlation between school-level student demographics and the school-level models presented in Table 4 is for current-year math performance levels, which do not adjust in any way for student characteristics typically associated with achievement. Among the four school value-added measures (VAMs), Similar Schools Comparisons (SSC) shows the weakest relationship to student characteristics; with the exception of special education (ρ=0.08), all demographic correlations for the SSC are below 0.05 in magnitude. SGP and SVA have very similar patterns of correlation with demographics, for which the strongest demographic correlations occur with schools' percentages of African-Americans (ρ=-0.13 and -0.14 respectively) and low-income students (ρ=-0.14). These negative correlations show that schools with low SGP and SVA results tend to have higher percentages of low-income and African-American students, though the relationship is weak enough to be negligible. SGP and SVA are positively correlated with schools having more Hispanics and English Language Learners. All four VAMs in Table 4 show weak correlations with student characteristics, though SSC correlations are the weakest. Of particular concern are SGP and SVA correlations with schools' percentage of low-income and African-American populations. All four VAMs are also modestly correlated with math performance, though SSC has an inverse relationship with prior test scores.

In addition to examining simple correlations of different VAMs with student characteristics, I regress the results of each model on the same set of school covariates used to generate the similarity index. The results of these regressions are shown below in Table 5. Mean standardized test scores are included as well for comparison.

Table 5. Impact of School Characteristics on Selected School Measures, 2011-14

Demographic	Student Growth Percentiles	Student Value Added	Similar Schools Comparisons	Mean Prior Z	School Performance (z)
Male (%)	0.01	0.02	-0.03	-0.02	-0.03*
Hispanic (%)	0.02	0.04	-0.03	-0.04	0.11*
Native American (%)	0.02	-0.01	0.02	-0.00	-0.01
Asian (%)	0.11**	0.10**	-0.02	-0.13**	0.08**
African American (%)	-0.05*	-0.07*	-0.03	0.06*	-0.38**
Pacific Islander (%)	0.01	0.00	-0.03	-0.03	-0.01
Two or More Races (%)	-0.01	-0.01	-0.01	-0.01	0.01
Free/Red. Lunch (%)	-0.11**	-0.13**	0.07*	0.12**	-0.31**
Special Educ. (%)	-0.06**	-0.00	0.08**	0.12**	-0.32**
Eng. Lang. Learner (%)	0.08	0.05	0.00	-0.06	-0.14**
School Size (per 1000)	-0.02	-0.05	0.06*	-0.00	0.06**
R^2	**0.06**	**0.05**	**0.01**	**0.09**	**0.56**

* - 95% confidence

** - 99% confidence

NOTE: All coefficients are standardized with respect to each regressor's variance, so magnitudes are comparable.

All four VAMs have a weak relationship with the full set of school characteristics as indicated by the coefficient of determination, though significant relationships exist with individual demographic regressors in the model. While patterns of significance could be impacted by multicollinearity between closely related percentages, this threat does not extend to R^2, which is interpretable as the percent of variance explained by the model regardless of whether regressors are collinear or independent of each other. An overall weak fit is desirable since each model attempts in different ways to account for non-school factors influencing student outcomes. These characteristics explain only 9% of variation in Mean Prior Z, 6% of variation for Student Growth Percentiles, and only 5% of variation in estimates of Student Value Added. The coefficient of determination for Similar Schools Comparisons is even lower, at only 1%. For reference, these same characteristics explain 56% of the variance in school performance as summarized by mean standardized scaled scores in math.

An overall weak fit does not mean that each model perfectly controls for non-school factors, and the existence of large, statistically significant coefficients on some characteristics for school VAMs presents an opportunity for further investigation. Because coefficients are standardized, their magnitudes are comparable between the models. SGP and SVA show a consistently significant relationship with schools' Asian percentages (positive), African-American percentages (negative), and the percentage of students receiving free or reduced-price lunches (negative). By comparison, similar schools ranks show no significant relationship to Asian or African-American percentages. Similar schools ranks' relationship to free/reduced lunch students is both more modest and in the opposite direction (positive) than is the case for SGP and SVA; higher levels of economic disadvantage appear to *boost* schools' results under SSC and MPZ. Curiously, similar schools ranks show a modestly positive but significant relationship to special education students, as well as school size. Although these relationships do attain statistical significance with at least 95% confidence, their overall impact on the models is nevertheless bounded by the low R^2 observed on each of the models, with the SSC model showing the lowest R^2 among the models.

In sum, all models examined here show weak-to-modest relationships with student characteristics, whereas unconditioned levels of school test scores in math show a very strong relationship to student characteristics. Although similar schools ranks show the weakest relationship to student characteristics, all four of the models sufficiently account for student background characteristics to be admissible as value-added measures that exhibit a high degree of fairness.

While it is encouraging that all four models are fair with respect to demographics, the fact that all four are very fair does not help discriminate among them when considering whether one is more preferable than the others for school policy or public information. In addition to considering fairness, I next turn to a consideration of the models' stability over time, which is important if the measures are to be believed and accepted by educators and the public.

Stability of Models Over Time

One of the major questions that must be asked of any model is its stability over time. A volatile or noisy model is likely to be confusing or mistrusted in practice. While some variability is necessary, since true school effectiveness certainly changes at least modestly from year to year, wild jumps rather than smooth trends can rightly lead stakeholders to question the informational content of effectiveness measures. For this reason reliability or stability is a highly desirable property of VAMs. Table 6 shows the correlation of school VAMs (as well as school z-scores) with prior-year values. I include schools' math z-scores as a point of comparison; they are not to be understood as constituting a VAM alongside similar schools, SGP, and SVA. Coefficients are shown for each year in relation to its prior year. Coefficients for 2012, for instance, represent the correlation between 2011 and 2012 values for each measure.

Table 6. Correlation of School Outcomes with Prior Year Values

Year	Student Growth Percentiles	Student Value Added	Similar Schools Comparisons	Mean Prior Z	School Performance (z)
2012	0.43	0.43	0.64	0.57	0.83
2013	0.35	0.44	0.59	0.50	0.82
2014	0.51	0.55	0.64	0.64	0.83

Similar Schools Comparisons shows higher stability from year to year than the other three VAMs here, though MPZ is as stable as SSC between 2013 and 2014 (*p*=0.64). Similar schools coefficients range from 0.59 to 0.64, while those for SGPs and Student Value Added are below 0.45 in 2012 and 2013, rising to 0.51-0.55 in 2014. As one would expect, all school VAMs have lower year-to-year stability than standardized scale scores (0.82-0.83).

To gain a more practical understanding of stability, it is useful to examine distributional changes from year to year. Table 7 shows the prior-year distribution of all schools in the top quintile in a given year on each measure. Table 8 shows the same table for schools in the bottom quintile in a given year.

Table 7. Previous Year's Distribution of Current Year's Top Quintile, 2012-14

Quintile (t-1)	Student Growth Percentiles	Student Value Added	Similar Schools Comparisons	Mean Prior Z	School Performance (z)
Top 20%	43.2%	50.1%	58.2%	50.7%	67.8%
Higher 20%	23.0%	21.7%	22.8%	25.1%	21.3%
Middle 20%	14.0%	13.3%	10.1%	13.6%	6.3%
Lower 20%	11.5%	9.3%	6.3%	8.1%	3.1%
Bottom 20%	8.4%	5.5%	2.7%	2.5%	1.6%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 8. Previous Year's Distribution of Current Year's Bottom Quintile, 2012-14

Quintile (t-1)	Student Growth Percentiles	Student Value Added	Similar Schools Comparisons	Mean Prior Z	School Performance (z)
Top 20%	8.7%	5.8%	1.9%	5.0%	0.6%
Higher 20%	9.6%	11.8%	9.0%	7.9%	1.6%
Middle 20%	17.6%	15.2%	13.8%	13.2%	7.2%
Lower 20%	22.7%	23.0%	26.7%	23.2%	19.9%
Bottom 20%	41.5%	44.2%	48.6%	50.7%	70.7%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

The relative quintile distributions of schools on each measure vary similarly to model differences in year-to-year correlations: SGP, SVA, and MPZ show the greatest prior-year spread, while similar schools show a slightly narrower spread. In Table 7, 58% of top-quintile schools for similar schools rankings were also in the highest quintile in the previous year, compared to 50% for student value-added and 43% for SGPs.[10] As is the case in the correlational analysis, schools' mean z-scores in math show greater stability. The percentage of schools jumping from the bottom quintile to the top in one year is 2.7% for similar schools, not much higher than is the case for simple z-scores. The corresponding percentages for SGPs and student value-added are 8.4% and 5.5%, respectively.

Examining the prior spread of each year's bottom quintile as shown in Table 8 reveals a similar pattern. SSC shows less spread than the other VAMs but greater spread than school z-scores. Bottom-quintile stability for MPZ is two percentage points higher than SSC. Forty-nine percent of schools in the bottom 20% on similar schools rankings in a given year were also in the bottom quintile in the previous year, versus 51% on MPZ. Same-quintile stability for school z-scores is much higher, at 71%. As with the top-quintile analysis, fewer schools jumped from one end of the distribution to the other on SSC and MPZ than on the other VAMs. Only 1.9% of bottom-quintile schools were in the top quintile in the previous year. This compares favorably to SGPs (8.7%), SVA(5.8%), and MPZ (5.0%) but again, school z-scores are more stable, with only 0.6% of bottom-quintile schools coming from the previous year's top quintile.

[10] It is worth noting that the similarity index, upon which Similar Schools Comparisons are based, shows greater stability than any of the four VAMs as well as simple z-scores. This is to be expected, since it is designed to reflect student demographics and school inputs to the degree they impact student performance, and such inputs are less variable from year to year than are test scores. Because it only informs one of the VAMs, rather than being a VAM in its own right, it is not included in these comparisons.

As an additional measure of stability, Table 9 shows the level of within-model agreement in consecutive years as measured by Cohen's kappa coefficient. This coefficient takes into account the distribution that would be observed if the measure of interest were randomly distributed in each year, so that a kappa coefficient of 0 represents the amount of agreement that one would expect to observe purely by chance, while a coefficient of 1 represents perfect agreement.

Table 9. Cohen's Kappa, School Estimates for Current and Prior Year

Year	Student Growth Percentiles	Student Value Added	Similar Schools Comparisons	Mean Prior Z	School Performance (z)
2012	0.17	0.17	0.27	0.24	0.41
2013	0.15	0.17	0.27	0.19	0.42
2014	0.18	0.18	0.25	0.21	0.42

For properly interpreting the strength of agreement indicated by the coefficient, Landis and Koch (1977) recommend the following:

Poor: Below 0.0
Slight: 0.00-0.20
Fair: 0.21-0.40
Moderate: 0.41-0.60
Substantial: 0.61-0.80
Almost Perfect: 0.81-1.00

Given these interpretations, the stability of Similar Schools Comparisons (SSC) is fair (0.25-0.27), Mean Prior Z is marginally fair (0.19-0.24), and the stability of Student Growth Percentiles (0.15-0.18) and Student Value Added (0.17-0.18) is slight. All are less stable than unconditioned school performance in math (0.41-0.42), which exhibits moderate stability.

Among the four VAMs considered in this section, Similar Schools Comparisons exhibit the highest degree of stability over time. Their stability lies above that of the growth-oriented SVA and SGP models, and below simple unconditioned math test scores. Mean Prior Z is nearly as stable.

Curiously, the top quintile of schools as judged by Similar Schools Comparisons appears slightly more stable than the bottom quintile year-to-year, but at both extremes they are more stable than the same groupings under SGP and SVA.

Components Underlying Relative VAM Stability

The comparison of the four VAMs to school z-scores with respect to stability requires careful consideration of the sources of student achievement. Suppose that each student has a set of fixed individual characteristics which are separable from their year-to-year performance on tests; these characteristics are partially but not fully captured by observable demographics. At the school level, the aggregation of these characteristics, likewise partly measurable through demographics, likely shows less variability from year to year than test scores themselves. Student z-scores, then, contain both a fixed student characteristic and a residual, defined as the difference between one's individual mean and his current-year performance in a given subject. This conceptualization occurs at the school level as well. A measure such as the similarity index attempts to account for these characteristics. The use of average prior student test scores fulfills an analogous function in the case of the MPZ model as well.

However, by controlling for the most stable component of test scores, namely the fixed student characteristic, one should expect the variability in the residuals, upon which a VAM is based, to be greater than the variability of actual test scores. Thus any VAM which attempts to account for fixed student characteristics inherits a degree of variability that likely exceeds the variability observed in simple test scores themselves, and the degree of stability observed in schools' yearly z-scores may be unattainable for any VAM without compromising its validity. Given this, the modest degree of stability observed for similar schools ranks may fall closer to the upper bound of stability for VAMs than a naïve comparison with z-scores would otherwise suggest. On the other hand, a certain degree of VAM stability may be due to the failure to fully control for students' fixed characteristics, in which case a VAM based on inadequate controls would be stable for the wrong reasons, being a composite of unobserved individual fixed effects and true school effectiveness. The MPZ model represents an example VAM that accounts for prior student fixed effects to a degree unattainable in the case of Similar Schools Comparisons, and its stability is very similar to SSC stability as indicated in the above tables.

The necessity of inherent variability in residual-focused VAMs suggests that a very high degree of stability may in fact undermine validity, by controlling away actual changes in school effectiveness over time. This means that the role of stability in determining one's preferred model should be secondary to that of validity, so that stability might adjudicate only between models that are approximately valid.

In the following two sections I undertake the question of validity by examining patterns of agreement and disagreement between the four models.

Agreement Between Models

In addition to reliability, the degree of agreement between Similar Schools Comparisons and the other three VAMs is a central question of this dissertation. This question is taken up to consider the validity of the SSC in comparison to the other models. Validity can be understood broadly in two ways. First, does the model tell us what it is supposed to tell us about schools based on what is valued? Second, if not, what is the proper interpretation for a model? While there are innumerable approaches to pursuing these questions, an examination of agreement between the models can at least shed light on whether Similar Schools Comparisons, the ultimate model of interest in this dissertation, closely align with the results of the growth-oriented SGP and SVA models. If they do, then the SSC model can be interpreted as a simpler and more compelling presentation of the same information contained in growth VAMs. If they do not, then it would be wrong to conclude that SSC estimates are invalid; rather, one should examine the school factors underlying models' disagreements and on that basis determine whether SSC results provide information that is useful in considering school effectiveness yet also independent of, or even orthogonal to, the other three VAMs.

Table 10 shows correlations between the four VAMs as well as mean standardized scale scores (z-scores). Similar Schools Comparisons correlate modestly with the other three VAMs, at 0.51 with SGP and 0.53 with SVA. Its correlation with MPZ is somewhat higher, at 0.75. SGP is very highly correlated with student value-added estimates, at 0.87. This is noteworthy since the two models, despite both being growth-oriented, use quite different methods to estimate schools' contributions to student growth. SSC estimates are more highly correlated with achievement levels (z-scores), at 0.64, than are SGPs and student value-added, which each correlate with z-scores at 0.47.

Table 10. Correlations Between Model Estimates, 2011-14

	Student Growth Percentiles	Student Value Added	*Similar Schools Comparisons*	Mean Prior Z	School Performance (z)
Student Growth Percentiles	1.0000				
Student Value Added	0.8717	1.0000			
Similar Schools Comparisons	***0.5064***	***0.5311***	***1.0000***		
Mean Prior Z	0.7980	0.7504	***0.5621***	1.0000	
School Performance (z)	0.4731	0.4667	***0.6444***	0.7405	1.0000

Table 11 compares SSC ranks (columns) with school estimates of SVA (rows) by quintile, for the years 2011-14. If the two models agreed perfectly, then the diagonal cells would each show 100%. The further schools are from the diagonal, the greater the disagreement between the two models for that particular school. In every column, the greatest percentage of schools fall in the cell that corresponds to the same quintile on Student Value Added. Agreement between the two is greatest for the top quintile (rightmost column) and bottom quintile (leftmost column), while the middle quintiles show greater amounts of disagreement, as can be seen in the middle column. The middle 20% of schools on similar schools ranks is spread nearly evenly among the five quintiles on school value-added; if these schools were uniformly distributed, then each cell would show 20%, while the observed distribution shows no cell with less than 15% or more than 25%. One possible reason for this is that higher rates of quintile agreement at the tails may be due to floor and ceiling effects, in which the conversion of school estimates to ordinal rankings masks significant differences in the interval properties of the models at the tails.

Table 11. Quintile Comparison, Similar Schools Comparisons with Student Value Added, 2011-14

	SS Bottom 20%	SS Lower 20%	SS Middle 20%	SS Higher 20%	SS Top 20%
SVA Bottom 20%	***56.6%***	25.4%	15.1%	10.6%	4.3%
SVA Lower 20%	24.1%	***26.6%***	20.1%	18.1%	9.2%
SVA Middle 20%	11.5%	23.1%	***24.9%***	21.1%	15.8%
SVA Higher 20%	5.4%	16.7%	22.5%	***25.2%***	27.6%
SVA Top 20%	2.4%	8.2%	17.5%	25.0%	***43.0%***
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 12. Quintile Comparison, Similar Schools Comparisons with Mean Prior Z, 2011-14

	SS Bottom 20%	SS Lower 20%	SS Middle 20%	SS Higher 20%	SS Top 20%
MPZ Bottom 20%	***46.5%***	25.6%	15.3%	8.2%	3.7%
MPZ Lower 20%	26.7%	***28.9%***	20.8%	15.7%	7.9%
MPZ Middle 20%	16.5%	22.1%	***24.0%***	18.7%	15.3%
MPZ Higher 20%	6.6%	16.0%	24.2%	***28.1%***	22.0%
MPZ Top 20%	3.8%	7.5%	15.7%	29.3%	***51.1%***
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 12 compares Similar Schools Comparisons (columns) with Mean Prior Z by quintile, for the years 2011-14. As with Table 11, agreement is indicated by the diagonal cells, which represent schools in the same quintile on both models. Agreement between the two models is greatest for the top and bottom quintiles, while the greatest disagreement occurs for the middle quintiles. Schools falling in the second-highest quintile on Similar Schools Comparisons are slightly more likely to fall in the highest quintile on MPZ (29.3%) than in the equivalent MPZ quintile (28.1%). Tables 13 and 14 also compare quintiles between the models, and levels of agreement are similar. While Table 14 compares MPZ with SGP, I omit a comparison of MPZ with SVA since the results are very nearly identical to the MPZ-SGP comparison; SVA and SGP themselves align extremely closely, as shown in Table 15. Overall, levels of agreement are very similar between the three tables (see Table 16).

Table 13. Quintile Comparison of Similar Schools Comparisons with SGP, 2011-14

	SS Bottom 20%	SS Lower 20%	SS Middle 20%	SS Higher 20%	SS Top 20%
SGP Bottom 20%	***53.0%***	26.4%	15.2%	12.3%	7.0%
SGP Lower 20%	24.5%	***26.9%***	19.2%	17.4%	10.0%
SGP Middle 20%	11.8%	21.7%	***27.2%***	18.4%	16.1%
SGP Higher 20%	7.4%	15.6%	22.5%	***26.4%***	25.1%
SGP Top 20%	3.2%	9.5%	15.8%	25.4%	***41.8%***
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 14. Quintile Comparison of Mean Prior Z with SGP, 2011-14

	MPZ Bottom 20%	MPZ Lower 20%	MPZ Middle 20%	MPZ Higher 20%	MPZ Top 20%
SGP Bottom 20%	***64.1%***	21.1%	6.3%	1.1%	1.2%
SGP Lower 20%	25.1%	***44.4%***	22.6%	6.0%	0.7%
SGP Middle 20%	7.6%	22.8%	***38.7%***	23.4%	3.7%
SGP Higher 20%	2.2%	9.5%	24.9%	***41.1%***	20.1%
SGP Top 20%	1.1%	2.2%	7.5%	28.3%	***74.3%***
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

To further investigate the high correlation observed between SVA and SGP in Table 8, Table 15 compares Student Value Added (SVA) estimates with SGPs. The two models do show very high levels of agreement by quintile here, particularly in comparison to Tables 11-13. In no quintile are agreement levels lower than 49%, and the top and bottom quintiles have agreement levels at or above 75%. Notice as well that strong disagreements are virtually nonexistent: no schools are in the bottom quintile for SVA estimates but the top quintile for SGP, nor vice versa. This is not surprising, since SGP and SVA are more methodologically similar to each other than either is in comparison to SSC and MPZ, which are themselves similar to each other.

Table 15. Quintile Comparison of Student Value Added with SGP, 2011-14

	SVA Bottom 20%	SVA Lower 20%	SVA Middle 20%	SVA Higher 20%	SVA Top 20%
SGP Bottom 20%	***83.6%***	20.5%	1.6%	0.5%	0.0%
SGP Lower 20%	14.4%	***58.5%***	18.8%	3.9%	0.9%
SGP Middle 20%	1.6%	19.2%	***48.9%***	19.1%	5.7%
SGP Higher 20%	0.5%	1.8%	29.0%	***51.6%***	17.7%
SGP Top 20%	0.0%	0.0%	1.8%	24.9%	***75.7%***
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 16 combines the five *between-model* comparisons to describe overall levels of agreement, as well as strong disagreement. Strong disagreement is defined here as school estimates occurring in the bottom quintile on one measure but the top quintile on another for estimates in the same year; these cases are the focus of the section immediately following. SSC estimates have low levels of strong disagreement with the other two models. About 1.4% of schools with results on the SSC model are in the top or bottom quintile yet the opposite quintile on Student Value Added (SVA). The rate of strong disagreement when comparing SSCs with student growth percentiles is similarly low, at 2.1%. Even when allowing disagreements to span 3 quintiles (1 → 4; 2 → 5; 1 → 5), rates of disagreement are 8.2% and 9.9% respectively. As one would expect given the very strong relationship between SVA and SGP, there are no schools that fall at the top or bottom on one of those measures but at the opposite end on the other, and 3-quintile disagreement is only 0.4%. The frequency of agreement shown here is the percent of all schools with estimates in both models that fall in the same quintile on each. For SSC, about 36% of schools fall in the same quintile on both SSCs and SVA, and quintile agreement is 35% when comparing SSCs to SGP.

Table 16. Quintile Disagreement Between Models, 2011-14

	SSC vs SVA	SSC vs SGP	SSC vs MPZ	MPZ vs SGP	SGP vs SVA
Frequency of 1-5 Disagreement	1.4%	2.1%	1.6%	0.5%	0.0%
Frequency of Disagreement >= 3 Quintiles	8.2%	9.9%	7.7%	1.7%	0.4%
Frequency of Disagreement	64.3%	64.8%	63.8%	47.3%	36.2%

In addition to analyzing results by quintile, it is valuable to compare disagreements between models when the results of those models are judged statistically significant with respect to each of their methods. This is useful because models may be more or less noisy, and a quintile analysis does not discriminate with respect to noise or precision. Table 17 represents a slightly different specification of agreement between school effectiveness models. Rather than simply analyzing agreement by quintile, Table 17 compares SVA and similar schools by significance.[1] SGP, at the time this analysis was conducted, did not allow for the use of confidence intervals around school estimates due to limitations in the availability of state data on test score error estimates. If available, this would allow one to determine whether a school's median SGP is significantly different from the state median of 50. For this reason, SGP is not included in the analysis presented in Table 17.

Statistical significance is available for SVA estimates since the model yields standard errors which can be used to construct confidence intervals; a 95% confidence interval is used here to distinguish estimates from the null hypothesis that a school's value-added was actually zero. As can be seen in the right column of Table 17, about one-quarter of all SVA estimates are significantly positive, another one-quarter are significantly negative, and the remaining half of estimates are indistinguishable from the null hypothesis. For similar schools ranks, significance is imputed at the 20th and 80th percentiles as described in Chapter 3.

If similar schools significance bore no relationship to SVA significance, then all columns would look like the right column, with about half of similar schools in a given column having SVA estimates that cannot be distinguished from the null. Among schools near the top of their comparison group (Exceeds Peers), 54% have significantly positive SVA estimates despite only 24% of all SVA estimates being significantly positive. The strength of agreement is weaker at the other end, with only 46% of schools judged as trailing their peers also having a significantly negative SVA estimate. In fact, a greater share of these schools have null SVA estimates (50%). Among all schools in the sample with both SVA and similar schools estimates, only 1.7% are significantly positive in one model but significantly negative in the other (percentage not shown). This represents a problem in practice, but may be small enough to permit solutions that do not threaten the basic validity or functional form of either model.

Table 17. Significance Agreement Between Similar Schools Comparisons (SSC) and Student Value Added (SVA), 2011-14

	SSC Exceeds Peers	SSC Typical	SSC Trails Peers	Percent of all SVA Estimates
SVA Positive, 95% Confidence	53.7%	21.4%	3.2%	23.5%
SVA Null, 95% Confidence	40.9%	53.8%	50.3%	50.6%
SVA Negative, 95% Confidence	5.5%	24.7%	46.5%	26.0%

The presence of small but nontrivial levels of disagreement between similar schools ranks and growth-oriented VAM estimates under SGP and SVA presents the need to seek out the factors underlying disagreement between the models in order to make a fair determination of the validity of the SSC in comparison to SGP and SVA. This analysis is taken up in the following section.

Characteristics of High-Disagreement Schools

An examination of the characteristics of schools which receive widely different estimates on the VAMs presented here helps shed light on the degree to which the models truly provide different information on school effectiveness. Disagreement need not be taken at face value; it is possible that high-disagreement cases are due to statistical quirks or noise, rather than stable and meaningful disagreement. An example of the former would be a very small school that shows high variability in its estimates from year to year, or a school achieving at the very top or bottom of schools in the state, where VAMs may tend to be less trustworthy. An example of the latter—that is, stable and meaningful disagreement—would be a school that greatly exceeds its demographic expectation, but which shows low student growth from year to year, or conversely tends to fall below its demographic expectation but which shows high student growth from one year to the next.

To examine this question, I continue the use of quintiles to compare the results of the four VAMs. Specifically, I identify three different groups of schools. The first group of schools are those for which SSC estimates fall in the top two quintiles but which have SVA estimates at least three quintiles below their SSC estimates; that is, the SSC method produces results that are significantly different, not just marginally different. The second group of schools represents the inverse: schools which fall in the bottom two quintiles on SSC estimates but are at least three quintiles higher on SVA. The third group of schools represents the bulk of schools in my analysis, for which the two models have lower, but not necessarily insignificant, levels of disagreement. A simple way to think of the high-disagreement groups is that each represents a set of schools that are, by quintiles, always substantially above the average on one model but below the average on the other model. This grouping scheme is illustrated in Figure 7. The analysis described here, comparing SSC results to SVA results, is performed identically to also compare similar schools results to SGP.

Figure 7. Analysis Groups for Quintile Disagreement Between Models

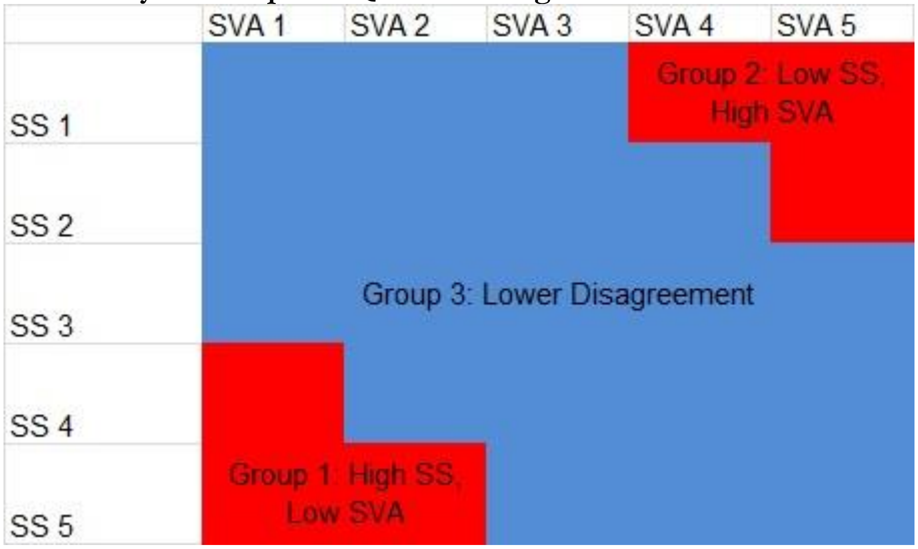

SVA 1
SVA 2
SVA 3
SVA 4
SVA 5
SS 1
SS 2
SS 3
SS 4
SS 5
Group 2: Low SS, High SVA
Group 3: Lower Disagreement
Group 1: High SS, Low SVA

Table 18 represents the comparison visualized above by comparing school characteristics for the three groups shown in Figure 7.

Table 18. Characteristics of High Disagreement Schools Between Similar Schools Comparisons and SVA

School Measure	Group 2 (Low SSC, High SVA)	Group 3 (Similar Results)	Difference	Group 1 (High SSC, Low SVA)	Group 3 (Similar Results)	Difference
Students Tested	146.4	186.2	39.83	126.1	186.2	60.125**
Math Score (z)	-0.334	-0.044	0.290*	0.037	-0.044	-0.082*
Similarity Index (z)	-0.094	-0.045	0.049	-0.188	-0.045	0.143**
Free/Reduced Lunch (%)	65.6%	66.3%	0.7%	73.7%	66.3%	-7.4%**
African American (%)	21.5%	18.0%	-3.5%	35.4%	18.0%	-17.4%**
Hispanic (%)	9.7%	7.5%	-2.3%	4.1%	7.5%	3.4%**
White (%)	65.0%	71.3%	6.3%	58.4%	71.3%	12.9%**
Special Education (%)	13.7%	11.4%	-2.2%	10.2%	11.4%	1.2%
Limited English Proficient (%)	6.0%	4.5%	-1.4%	1.5%	4.5%	3.1%**
N	**71**	**2031**		**109**	**2031**	

* - 95% confidence
** - 99% confidence

The most obvious pattern in Table 18 is the preponderance of significant differences on the right side of the table, which compares high SS – low SVA schools to those for which the models produce roughly similar results. Group 1 represents schools which exceed their expectation based on demographics in comparison to peer schools, but which exhibit low student growth as judged by SVA. These schools are much smaller than most schools in the sample. They are also more disadvantaged as judged by the variables used in the similarity index, which aligns with demographic differences: more of their students receive free or reduced-price lunches, they serve significantly more African-Americans, and they are less white. They show modestly higher math scores than the schools exhibiting lower levels of disagreement.

On the other side of Table 18, there are fewer significant differences in school characteristics when comparing schools with low outcomes on similar schools but high outcomes on SVA. Only mean math scores are significantly different, with Group 2 schools having lower scores than Group 3 schools.

Tables 19 and 20 present the same analysis, comparing characteristics of schools showing high disagreement between similar schools and SGP. These results follow a pattern very similar to that seen in Table 15, which is not surprising given the very high correlation between SVA and SGP.

Table 19. Characteristics of High Disagreement Schools Between Similar Schools Comparisons and SGP

School Measure	**Group 2 (Low SS, High SGP)**	**Group 3 (Similar Results)**	**Difference**	**Group 1 (High SS, Low SGP)**	**Group 3 (Similar Results)**	**Difference**
Students Tested	162.5	188.0	25.5	103.2	188	84.867**
Mean Math z-score	-0.281	-0.048	0.233*	0.08	-0.048	-0.128**
Similarity Index (z)	-0.06	-0.047	0.013	-0.162	-0.047	0.115**
Free/Reduced Lunch (%)	66.1%	66.2%	0.1%	73.2%	66.2%	-7.0%**
African American (%)	20.4%	18.2%	-2.2%	29.9%	18.2%	-11.8%**
Hispanic (%)	8.6%	7.5%	-1.1%	4.8%	7.5%	2.7%**
White (%)	66.5%	71.1%	4.6%	63.3%	71.1%	7.8%**
Special Education (%)	12.2%	11.5%	-0.7%	10.3%	11.5%	1.2%
Limited English Proficient (%)	5.1%	4.6%	-0.5%	2.0%	4.6%	2.5%**
N	**88**	**1992**		**131**	**1992**	

* - 95% confidence

** - 99% confidence

As with Table 18, the most obvious pattern in both Tables 19 (SSC vs SGP) and 20 (SSC vs MPZ) is that school characteristics are significantly different on the right side of the table (Group 2 vs Group 3), while showing smaller and less significant differences on the left side of the table (Group 1 vs Group 3). The direction and magnitude of the differences in Tables 19 and 20, as well as the pattern of statistical significance, are nearly identical to those observed in Table 18: Group 1 schools are smaller, higher achieving, more disadvantaged, poorer, and serve a greater share of African-Americans than Group 3 schools. On the left side of the table, Group 2 schools have significantly lower math scores than Group 3 schools, but other than this, no significant differences are observed.

Table 20. Characteristics of High Disagreement Schools Between Similar Schools Comparisons and Mean Prior Z

School Measure	Group 2 (Low SS, High MPZ)	Group 3 (Similar Results)	Difference	Group 1 (High SS, Low MPZ)	Group 3 (Similar Results)	Difference
Students Tested	109.7	183.7	74.038	63.38	183.7	120.362*
Mean Math z-score	0.187	-0.0470	-0.234**	-0.693	-0.0470	0.647**
Similarity Index (z)	0.0170	-0.0500	-0.0670	-0.668	-0.0500	0.618**
Free/Reduced Lunch (%)	66.0%	66.5%	0.00500	80.1%	66.5%	-0.136*
African American (%)	5.8%	18.9%	13.1*	48.2%	18.9%	-0.293**
Hispanic (%)	6.9%	7.4%	0.5	3.9%	7.4%	0.0350
White (%)	84.9%	70.4%	14.4*	44.9%	70.4%	0.255**
Special Education (%)	11.2%	11.3%	0.1	37.0%	11.3%	-25.7**
Limited English Proficient (%)	2.8%	4.5%	1.7	2.1%	4.5%	0.0240
N	**24**	**2171**		**13**	**2171**	

* - 95% confidence

** - 99% confidence

The most obvious conclusion which can be drawn from this is that, even in cases of high disagreement between models, the similar schools method developed here (SSC) does what it is intended to do. When SGP, SVA, or MPZ is high but similar schools estimates are low, schools are roughly similar in student composition but schools with low similar schools ranks have lower test scores. When one of the comparison VAMs is low but SSC results are high, schools are serving substantially disadvantaged populations but are actually exceeding state averages on their students' math performance.

This finding is encouraging as a validation of the similar schools methodology used here, but it necessarily leaves unanswered the question of the degree to which unobservable student characteristics and within-school peer effects mediate differences in the results of the two models. This question is examined in greater detail in Chapter 5.

Chapter 5: Discussion of Results and Further Study

The major finding of the analysis presented in the previous chapter is that the Similar Schools Comparison (SSC) model, a simple, non-growth, cross-sectional value-added model for schools, is modestly correlated (0.55) with VAMs focused on student-level changes in test scores (SGP, SVA) as well as with the Mean Prior Z (MPZ) model. The very high correlation (0.87) between the Student Growth Percentile (SGP) and Student Value Added (SVA) models is indicative of their mutual focus on student learning growth without an explicit consideration of demographics. Secondary to this finding, the Similar Schools Comparison model is more stable over time than are the three comparison VAMs. This is most likely due to the greater level of noise inherent in any growth model compared to cross-sectional analyses which only consider the level of achievement of schools and students at a single point in time. Interestingly, the MPZ model, which is capable of measuring growth over multiple years at the student level, shows a level of variability that is still greater than SSC but less than SGP and SVA.

Deciding among different school VAMs in a policy setting is complex. The original question is quite simple: what is the best way to measure a school's contribution to student learning? Yet in choosing among possible models, one must confront multiple, occasionally competing priorities, in addition to the pure question of which model most accurately captures school effectiveness. Borrowing from the analysis presented in Polikoff (2013), I compare the model results of the analysis presented in the previous chapter with respect to four priorities: validity, stability, fairness, and transparency.

Validity

The validity of a VAM is the degree to which it accurately measures what it is intended to measure, which is typically considered to be some form of schools' contribution to student learning, controlling for non-school factors impacting school outcomes (Polikoff 2013). There is, however, some latitude within this broad definition. Consider the differences between SSC and SGP. At least in the short-term, Similar Schools Comparisons do not pretend to capture schools' contribution to student learning growth from year to year. Nor do student growth percentiles pretend to account for the effect of demographics on student achievement. While both represent broad attempts to estimate schools' contributions to student learning, they do so in very different ways. Aside from the technicalities involved in estimation, both are intuitively understandable as different approaches to the same question. To the degree that they are communicated in a manner that aligns with their methods, both may be equally valid with respect to their own intent. The question thus becomes which intent is preferable as a tool for understanding school effectiveness more broadly.

The differences described above hold in the short-term, but beyond one year they must converge. It is worth noting that both models point in the same direction, namely, toward the contribution to learning that schools make over multiple years beyond what one would expect given student characteristics. A school which exceeds its demographic expectation does so because, over a period of years, one of two things has happened: either students experienced exceptional growth in the school they were in previously, and the current school has maintained them above expectation, or students came in at expectation and have achieved exceptional growth while in the school. In either case, exceptional growth, presumably measurable at some point in time through a growth-oriented VAM, has led to a current situation in which students' performance exceeds their demographic expectation.

VAM-related efforts in recent years have focused more on the side of learning growth rather than contemporaneous comparisons, but it is not obvious that year-to-year growth should be the preferred basis for estimating school effectiveness. If one believes that student growth is the ultimate basis on which to judge school effectiveness, then the relatively modest correlation (0.55) between SSC and the two most growth-oriented VAMs (SGP and SVA) may be considered as an indictment of similar schools ranks. Yet this assumption should not be granted without serious forethought, particularly when considering schools rather than teachers. In the case of teachers, the period of 'treatment' is typically one year, so a VAM can focus on a simple year-to-year change: the level at which a student was achieving in the year prior to having a given teacher, and the gain that student achieved under the teacher.[11] Even when multiple prior test scores are used, the focus of estimation is still a single year. Yet in the case of schools, a student experiences multiple years in the same environment. When regression to the mean is present, a VAM which focuses on short-term growth will penalize a school for past success with students. For every student, there is thus a need for a counterfactual which does not depend on a school's prior contribution to student learning when the same school is the unit of analysis. Student growth percentiles, as well as the SVA model, only rarely are capable of providing this counterfactual, and it is partly for this reason that school estimates under these models exhibit less stability than similar schools ranks. Given this, the MPZ model represents an instructive comparison for SSC, by still being growth oriented (current scores minus average of scores prior to entry into current school), yet considering growth over the full period of treatment, namely, the time since entering the school. Because MPZ correlates with SSC at a similar level to SGP and SVA (0.56 vs 0.53 and 0.51 respectively), it appears unlikely that measuring student growth across multiple years substantially strengthens the relationship between SSC and the other VAMs.

[11] This is not to say that VAMs account only for crude changes in individual test scores under a teacher from year to year. Multiple prior years can be included as controls, as well as information on student characteristics. Further, one can include information on classroom composition, without which a couple troublesome students can pull down classroom effectiveness in a way not easily captured by models. Still, all these factors typically serve to frame and condition *annual* changes in student performance to estimate teacher effectiveness, while the period required to estimate school effectiveness should be equal to the amount of time students have been receiving education in a given school.

Thus far I have argued that the Similar Schools Comparisons model provides information that is different from growth-oriented VAMs by design, so that Similar Schools should not be judged on the same set of criteria. However, granting that similar schools comparisons should be judged against the definition outlined above, in which similarity is judged by controlling for the impact of student characteristics on test scores, still leaves a major threat to internal validity. By relying on demographic controls only, the Similar Schools Comparisons model fails to capture fixed student characteristics that are unobserved in demographic data but which influence outcomes. The MPZ model attempts to provide such a comparison by estimating a fixed effect for each student by averaging all available within-subject test scores prior to a student's entry into his or her current school. If the SSC model effectively captured this, then one would expect a significantly stronger relationship between SSC and MPZ than between SSC and the other two VAMs. This, however, is not the case and bears further study. One possible confounding factor in the comparison of SSC and MPZ is that the regressor of interest in MPZ is based on simple averages of students' prior test scores, so it does not allow for peer effects, which are, on the other hand, implicitly absorbed in the school-level estimations used in the SSC model.

In short, the validity of the Similar Schools Comparison model in comparison to the other three VAMs is only answerable in this analysis by assuming the antecedent or begging the question, namely, that some version of student growth is the preferable measure of school effectiveness. If this is not granted, and the different models are communicated accurately, then the ultimate determination of validity remains open, and cannot be answered since true effectiveness remains elusive.

Stability

In educational assessment, reliability is the degree to which an instrument yields stable results on a given population over time. The term is conceptually useful and well-known but is used interchangeably here with *stability*. The use of the term reliability implies that changes over time are due entirely or mostly to statistical noise; that assumption is relaxed when using stability to describe the frequency and magnitude of changes over time. The stability of school estimates over time is a concern for two reasons. First, a model which produces results that classify many schools as highly effective one year but ineffective the next is unlikely to be believed by the public, regardless of its statistical merits. Second, there are *a priori* reasons to believe that school effectiveness, while certainly time-varying, does not vary greatly from one year to the next. School inputs, both observable and unobservable, are fairly stable from year to year: the teaching workforce, school leadership, funding levels, community involvement, and curriculum are all matters which tend to change either slowly or infrequently. Insofar as effectiveness is a function of these inputs, one should expect it to likewise change slowly, or at least to expect dramatic changes in effectiveness to occur infrequently.

Because stability strictly concerns the frequency and magnitude of changes over time, it is less subject to divergent interpretations than is validity. For this reason, the Similar Schools Comparison model can be compared directly to SGP and SVA with regard to stability in a way that cannot be done when considering validity. As one would expect when comparing a non-growth measure (SSC) to growth measures (SGP, SVA, and MPZ), Similar Schools exhibits greater stability from year to year than do SGP and SVA, as well as MPZ. The year-to-year correlation for similar schools ranks ranges between 0.59 and 0.64, while the same values for SGP and SVA range broadly between 0.35 and 0.55, depending on the year and model. MPZ rank coefficients are in between those two ranges in all years, at 0.50 to 0.57, though MPZ shows the same level of stability as SSC from 2013 to 2014. All school VAMs analyzed in Chapter 4 showed lower stability than did mean standardized scale scores, which were correlated between 0.82 and 0.83 in the period studied. The correlation observed in scale scores can be understood as likely exceeding the upper bound on the stability of any VAM that considers school outcomes over the same period. Not only can a VAM not be more stable than the outcomes of which it is a manifestation; if, as discussed in Chapter 4, scale scores are partly the result of individual fixed characteristics, then controlling for those characteristics will yield residuals that are considerably noisier themselves than are simple test scores. Thus the 0.82 correlation between yearly math z-scores is likely too high as an upper bound estimate of VAM stability, thereby placing the 0.59-0.64 correlation observed for similar schools closer to maximum stability than a naïve comparison might suggest. As such, although the stability of similar schools ranks is only moderate in absolute terms, it likely lies at the upper end of stability for VAMs, which is limited by the variability inherent in conditioned school outcomes.

Fairness

Fairness is defined here as the degree to which estimates of school effectiveness are independent of school demographics and non-school inputs. As shown in Chapter 4, all four VAMs presented in this dissertation appear to be much fairer with respect to demographics than any unconditioned treatment of school outcomes (see Ch. 4, Fig. 2). While observable demographics explain 56% of the variance in schools' mean z-scores in math, they explain less than 10% of the variance in results for each of the four VAMs (SSC, SGP, SVA, and MPZ). Nevertheless, among the four VAMs, Similar Schools Comparison shows the weakest observed relationship with student demographics, which only explain 1% of variance in the model. This implies that the SSC as a whole is unbiased with regard to school demographics, although the method may imperfectly control for specific factors despite an overall R^2 of 0.01: statistically significant coefficients were observed on free/reduced lunch percentage, special education percentage, and school size.

The correlation of SGP, SVA, and MPZ with school demographics is slight but unmistakable. In the case of SGP, this relationship is in line with the analysis carried out by Ehlert et al. (2013), which showed a similar relationship between demographics and VAM estimates when not conditioning on student characteristics. While all models represent a substantial improvement in fairness over unconditioned outcomes, the Similar Schools Comparison model appears fairest among the four models.

Transparency

As pointed out in previous chapters, among all possible VAMs there is likely to be a tradeoff between sophistication and transparency, or the degree to which a model is understood and acted upon by non-specialists (Goldhaber 2012). If a VAM is to be valued by the public and accepted by educators, it is preferable that non-specialists be able to at least conceptually explain the model to one another and value it accordingly. Additionally, the use of visible and publicly available information on schools lends transparency because of its immediate accessibility. Finally, models may be considered transparent to the degree they relate to measures which educators and communities already understand and value—in other words, measures that possess currency.

Quantitative comparisons alone cannot distinguish among the models regarding transparency, as is the case for validity (somewhat), stability, and fairness. Rather, transparency must be considered by reference to probable perceptions and capabilities of policymakers, educators, and the general public. Among the four models, it is arguable that the student value-added (SVA) model is the least transparent, although it represents the closest analogue to VAMs currently used in education policy and has the strongest methodological purchase in causal interpretation. Its lack of transparency is due to the fact that its methodology does not lend itself to easy explanation despite having a parsimonious specification. The very features of SVA that allow its parsimony are also what make it opaque to non-specialists: the use of individual fixed or random effects, and the use of multiple prior years of data as regressors. Despite both having the intent of accounting for individual inputs to learning that are outside the control of schools, fixed effects are more abstract than demographic controls which account for publicly visible information on students. In a similar way, while the use of multiple prior test scores as controls can more reliably capture true student growth, an understanding of how they work relies upon an understanding of multiple regression, knowledge of which is less common than the arithmetic required to calculate simple year-to-year changes. Moreover, SVA computations rely on data – student level test scores – that are not available in a public database. Lastly, the SVA produces a result that is essentially a coefficient on a school dummy variable, less meaningful in and of itself than a 1-99 median (SGP) or a -15 to +15 rank (SSC).

Student growth percentiles (SGP) likewise involve estimation techniques that are difficult to understand: quantile regression, piecewise cubic estimation, and the use of multiple prior test scores. However, SGPs can be intuitively explained and understood as representing a comparison of each student's growth to other students achieving at nearly the same prior level. This is frequently communicated visually, and an understanding of the estimation techniques underlying SGPs is not necessary to understand the results of the model. Further, school estimates can be calculated by anyone in possession of student-level growth percentiles, by simply rank-ordering all student SGPs and taking the median, or middle, growth percentile in the within-school ordering of student SGPs. These estimates are also very understandable: a median of 50 is typical, while a median below 40 or above 60 is low or high, respectively. As such, the SGP model is more transparent than SVA estimates. Once again, however, SGP computations rely on data – student level test scores – that are not available in a public database.

Mean Prior Z (MPZ), in comparison to SGP and SVA, is more computationally simple, but still relies on access to student-level data. The degree to which these factors impact transparency depends on the preferred definition of the term. If transparency means that a model is replicable by diligent individuals armed with public data, then MPZ is not transparent. If, on the other hand, one is concerned only with the degree to which a model can be intuitively understood, but not necessarily replicated by anyone, then MPZ is relatively transparent, relying as it does only on a simple average of students' prior test scores before entering their current school.

The Mean Prior Z (MPZ) model may be considered transparent due to its computational simplicity. Yet among the criteria previously stated—simplicity, accessibility, and currency—MPZ falls short of Similar Schools Comparisons on the latter two concerns. The MPZ model relies on the use of student-level data, which are not publicly accessible. Regarding currency, it is difficult to argue that average prior test scores are known, valued, and meaningfully acted upon, due to their lack of visibility. As such, MPZ is likely less transparent than Similar Schools Comparisons, though perhaps more transparent than SVA due to relative simplicity.

It is arguable that the Similar Schools Comparison model, based in this case upon the similarity index, are the most transparent among the four VAMs considered here. Although the calculation of the similarity index requires the use of ordinary least squares, it is nonetheless a relatively parsimonious model with a well-known and publicly visible set of regressors. Further, its motivating concept is easily understood: generate school-level expectations for achievement based upon student characteristics. Anyone armed with a rank-ordering of school expectations and schools' actual results can determine both a school's comparison group and its rank within that group; doing so would require something as simple as a spreadsheet. Moreover, one of the prime motivations behind similar schools comparisons is to leverage communities' existing knowledge about other schools in helping to situate and interpret the results of each school. To the degree that the comparison groups resulting from the model align with those perceptions, they receive a further boost to transparency. Finally, SSC computations rely on data – school level test scores – that are available in a public database.

Although all four models require some measure of faith on the part of non-specialists, SGP and SSC are presentable in an intuitive way and at least parts of the process leading to final school estimates can be undertaken and verified by educators and the public.

Summary of Criteria Comparisons

Table 21, shown below, draws together the results of the analysis just presented, using Polikoff's analysis (2013) as a guiding framework. The characterizations presented below are easily summarized as low, moderate, or high with the exception of validity, which requires reference to the definitions under which each of the models fall. This framework lends itself to which model to use given the relative importance one may attach to each of the criteria. Finally, these results should not be strictly interpreted as applying to all possible variations on each of the models shown, but instead apply only to the particular models developed and presented in this dissertation. Considered as a whole, the advantages of the SSC model in reliability, fairness, and transparency in comparison to the other VAMs recommend its potential usefulness in policy despite the threat to validity posed by the lack of a control for unobserved individual characteristics.

Table 21. Summary of VAMs with Respect to Selected Criteria

	Validity	**Reliability**	**Fairness**	**Transparency**
Student Growth Percentiles	Unconditioned annual growth: high	Moderate	High	Moderate
Student Value Added	Conditioned annual growth: high	Moderate	High	Low
Mean Prior Z	Conditioned cumulative growth: moderate	High	Moderate	Moderate
Similar Schools Comparisons	Student characteristics: moderate	High	High	High

Accountability As Flashlight or Hammer?

The concerns described in the above discussion regarding validity and disagreement between models may gain greater salience when models are used to explicitly incentivize schools, whether through rewards or sanctions. For instance, rewarding a school which performs well in the SSC model but fails to show significant growth as measured by SGP, or even shows below-average growth, may lead to resentment or confusion on the part of schools which excel on the latter given the achievement levels of incoming students. At the other extreme, sanctioning low-performing schools on the SSC with a loss of autonomy or of non-categorical funding may be punitive based on incomplete information about unobservable student and family characteristics for which the SSC model cannot account. In such a case, a direct consequence could be felt which is based on misleading or misinterpreted information.

While efforts to develop VAMs for principals are notoriously fraught by the difficulty of disentangling principal effects from other school-level effects (Chiang, Lipscomb, and Gill 2012), this does not imply that schools as a whole should not face VAM-related incentives, nor does it imply that principals should be fully insulated from school test results. States which reward schools based on changes in proficiency rates, for instance, might be much likelier to see a return on investment using VAMs, including SSCs. For instance, doing so could help properly frame teacher evaluation; if SSCs consistently incentivize leadership to improve, then the subjective component of teacher evaluations over which principals have responsibility would itself be accountable to a school-level result, potentially reducing the threat of favoritism or flippancy in such evaluations.

Even in the case where incentives are relaxed, it is conceivable that the use of any of the VAMs considered here, including the SSC model, may help to uncover true positives at the school level when properly interpreted, without facing the threat of false negatives being applied to schools which appear ineffective under a given model. Such an approach aims to uncover examples of relative effectiveness with similar populations of students so that other schools may learn from peer institutions which are showing unusual success with similar sets of advantages and challenges. Viewed in this way, shortcomings or caveats in validity may be less impactful than would be the case under traditional sanctions such as a loss of institutional autonomy.

Usefulness of Multi-Year Smoothing

One potentially helpful way to reduce rates of extreme disagreement between similar schools rankings and SGP (or SVA) would be to smooth school estimates by including multiple years of data. For example, using three years of data for both similar schools rankings as well as SGP and SVA would reduce the frequency of 1-5 disagreements (Ch. 4, Table 16) to the degree that such disagreements are an artifact arising from statistical noise. In any case, three-year averaging would reduce the contribution of noise toward such disagreement and thereby provide an estimate of true disagreement rates. The degree to which disagreement may be diminished when comparing model estimates over multiple years represents a promising line of further inquiry from the current dissertation.

Aligning Comparison Groups with Perceptions

One of the critical design questions that this dissertation has briefly mentioned but has not treated analytically is the alignment of comparison groups under the similarity index with the probable perceptions of educators and communities. If one of the major benefits of similar schools comparisons is to leverage stakeholders' tacit knowledge about other schools, then it would be wise for similar schools comparisons to account for factors that mediate that knowledge. To name two examples, communities are likelier to compare themselves to nearby schools, as well as schools in similar settings (rural, suburban, or urban). The degree to which these factors could be included in a model without diminishing its explanatory power over school outcomes is a matter of further investigation. California, the most well-known example of similar schools comparisons until the state stopped their use in 2013, explicitly considered this question in its technical working groups leading up to the adoption of its similar schools measure in 2000 (CST Working Group, 2000). Several measures are available for incorporating mediators of comparability alongside demographics. In addition to ordinary least squares, California policymakers considered the use of alternative weighting and selection techniques to arrive at a unique comparison group of a given size for each school.

Improving the Similarity Index

As discussed above, the greatest limitation of similar schools ranks, insofar as they depend upon a measure such as the similarity index, which attempts to control for fixed student characteristics, is that observable demographics cannot fully control for those fixed characteristics which constitute non-school inputs to learning. This represents a threat to validity. There are two possible solutions to this.

The first solution would be to optimize the use of observable demographics in the construction of a similarity index. This could be done in several ways. It would be wise to consider whether quantile regression would improve the explanatory power of certain demographic factors. This would be advisable if it could be shown that percentage changes in a given demographic had a nonlinear impact on expected outcomes. For example, if a school's FRL rate increasing from 70% to 80% had a different impact on expected outcome than the same rate increasing from 30% to 40%, then quantile regression would be justified. While previous research focusing on peer effects and the impact of 'concentrations of poverty' have argued that student poverty has institutional effects as well as affecting poor students individually (Coleman 1966; Kennedy 1986), this research does not constitute strong evidence of nonlinearity in the FRL-test scores relationship at the school level. Nevertheless, for such an important variable, the relationship is worth examining.

Another improvement to the similarity index would be to test for demographic interactions. For example, male-female learning differences may be greater within some ethnicities than others. Collectively, such a difference should be considered as a factor lying beyond the control of any single school, regardless of aggregate policy goals focused on the closing of achievement gaps. Such a difference would not be captured through the linear combination of gender and race/ethnicity coefficients in the similarity index and would need to be accounted for through the inclusion of an interaction term in the estimation method ultimately used to determine the similarity index.

A third improvement to the similarity index would be to test whether geographic information on schools helped explain variation in outcomes above and beyond what can be explained using student demographics. Geographic factors that could be considered include urbanicity, region or county within a state, and broader community characteristics such as income, household characteristics, and adult education levels.

The second solution to the problem initially posed in this section is demonstrated in the design of the Mean Prior Z model, which dispenses entirely with observable demographics and instead seeks out a purely achievement-based control for student achievement prior to entering a given school. Such an approach accounts for a greater share of student fixed effects than do observable demographics. However, any peer effects would be ignored in such a model, whereas they would be captured in a model which relies on estimations at the school level, which by definition include both pure individual effects and peer effects in aggregate achievement levels

Perhaps the best rationale for the exclusive use of prior test scores would be its considerable political benefit, by not explicitly situating school-level expectations in reference to observable traits, which some might construe as yet another example of what George W. Bush famously called the 'soft bigotry of low expectations' (Bush 2000). The challenge with this approach is that for many schools, information on ability or academic achievement is not collected prior to students entering the school, while for others, the existence of a single prior test score would provide a noisier or more biased estimate of a student's true ability than would the availability of a full set of observable demographics. To be fully useful, such models would nevertheless need some way to incorporate peer effects if the intervention of interest is the school as a whole rather than isolated individual students.

Rankings vs. Residuals

The combination of similar schools comparisons with the similarity index in the analysis just presented is a reasonable combination, but the two are not inseparable. Similar schools ranks are understandable as a framing and ordering mechanism that is adaptable to any value-added model that generates a complete set of both school-level expectations and school-level outcomes. In this regard, SSCs are implicitly an alternative to the use of residuals or coefficients on indicator variables as the outcomes of interest in value-added modeling. Although beyond the scope of this dissertation, one could extend the analysis in Chapter 4 by attempting to establish school-level expectations on SGP and SVA that take account of student characteristics implicitly; although demographics are not explicitly included in either model, prior test scores nevertheless are determined by student-level fixed characteristics, some of which are demographic in nature. Establishing school-level expectations in SGP and SVA could be done either by aggregating student expectations at the school level, or by regressing school outcomes on observable demographics to account for the impact of demographics *post hoc*. Nevertheless, the growth orientation of these models is likely to lead to school expectations that are unstable across years. This instability, if observed, would result most likely in unstable comparison groups, as well as unstable school ranks within each comparison group. In other words, it is possible to take nearly any outcome and frame it in a similar schools comparison, given an expectation or weighting on which to establish a basis for similarity.

The potential for similar schools comparisons as a fair, meaningful mechanism for modeling school effectiveness and informing school accountability is not exhausted by the empirical analysis presented in this dissertation. While the model presented shows modest alignment with more common measures used to evaluate schools, the differences in interpretation between the models leaves open the question of validity. Although the similar schools model presented here is very desirable with regard to stability and fairness, the inability to determine true validity necessarily leaves a caveat to the overall desirability of the model for policy purposes. Further, refinements to the method used to determine similarity could help adapt similar schools comparisons to the priorities of states and policymakers. Similar schools comparisons nevertheless represent a promising tool for states and cities, and in proportion to their promise these methods remain underdeveloped.

Bibliography

Aaronson, Daniel, Lisa Barrow, and William Sander. "Teachers and student achievement in the Chicago public high schools." *Journal of Labor Economics* 25, no. 1 (2007): 95-135.

Anderson, Theodore Wilbur, and Cheng Hsiao. "Estimation of dynamic models with error components." *Journal of the American statistical Association* 76, no. 375 (1981): 598-606.

Anderson, Theodore Wilbur, and Cheng Hsiao. "Formulation and estimation of dynamic models using panel data." *Journal of econometrics* 18, no. 1 (1982): 47-82.

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Arellano, Manuel, and Stephen Bond. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *The review of economic studies* 58, no. 2 (1991): 277-297.

Atteberry, Allison. "Defining School Value-Added: Do Schools that Appear Strong on One Measure Appear Strong on Another?" *Society for Research on Educational Effectiveness* (2011).

Ballou, Dale. "Test scaling and value-added measurement." *Education Finance & Policy* 4, no. 4 (2009): 351-383.

Betebenner, D. "Norm-and criterion-referenced student growth." *Educational Measurement: Issues and Practice* 28, no. 4 (2009): 42-51.

Betebenner, Damian W., and Robert L. Linn. "Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability." *Retrieved June* 1 (2010): 2010.

Betebenner, Damian W. "A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment." (2011): 439-450.

Booher-Jennings, Jennifer. "Below the bubble:"Educational triage" and the Texas Accountability System." *American educational research journal* 42, no. 2 (2005): 231-268.

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. *Estimating the effect of leaders on public sector productivity: The case of school principals*. No. w17803. National Bureau of Economic Research, 2012.

Briggs, Derek C., and Jonathan P. Weeks. "The sensitivity of value-added modeling to the creation of a vertical score scale." *Education* 4, no. 4 (2009): 384-414.

Bush, George W. "Speech to the 91st Annual Convention of the NAACP." July 10, 2000. Available http://www.washingtonpost.com/wp-srv/onpolitics/elections/bushtext071000.htm

California Department of Education, Office of Policy and Evaluation. "Construction of California's School Characteristics Index and Similar Schools Ranks." PSAA Technical Report 00-1, April 2000. Available http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0400.pdf

California Department of Education, Analysis, Measurement, and Accountability Reporting Division. "Report to the Legislature: Alternative Methods in Place of Decile Rank for the Academic Performance Index." October 2013. Available http://www.cde.ca.gov/ta/ac/ap/documents/rlaltmthdsdecilrankapi.pdf

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. No. w17699. National Bureau of Economic Research, 2011.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. *Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood*. No. w19424. National Bureau of Economic Research, 2013.

Chiang, Hanley, Stephen Lipscomb, and Brian Gill. *Is School Value-Added Indicative of Principal Quality?*. No. 7587. Mathematica Policy Research, 2012.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert York. "Equality of educational opportunity." *Washington, DC* (1966): 1066-5684.

Diaz-Bilello, Elena K., and Derek C. Briggs. "Using Student Growth Percentiles for Educator Evaluations at the Teacher Level: Key Issues and Technical Considerations." (2014).

Eddy-Spicer, David. "Horizontal Accountability in England." Conference paper, University Council for Educational Administration, 2014 Annual Conference.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael Podgursky. "Selecting Growth Measures for School and Teacher Evaluations: Should Proportionality Matter?." *National Center for Analysis of Longitudinal Data in Education Research* 21 (2013).

Goldhaber, Dan, Joe Walch, and Brian Gabele. "Does the model matter? Exploring the relationship between different student achievement-based teacher assessments." *Statistics and Public Policy* 1, no. 1 (2014): 28-39.

Goldschmidt, Pete, Kilchan Choi, and J. P. Beaudoin. "Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance." *Council of Chief State School Officers* (2012).

Hanushek, Eric A., Steven G. Rivkin, and John F. Kain. "Teachers, schools, and academic achievement." *Econometrica* 73, no. 2 (2005): 417-458.

Harris, Douglas N. *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press. Cambridge, MA, 2011.

Holmstrom, Bengt, and Paul Milgrom. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics, & Organization* (1991): 24-52.

Hoxby, Caroline. "If Families Matter Most, Where Do Schools Come In?." *A Primer on America's Schools* (2001): 89-126.

Kennedy, Mary M. "Poverty, Achievement and the Distribution of Compensatory Education Services. An Interim Report from the National Assessment of Chapter 1." (1986).

Koedel, Cory, and Jiaxi Li. "The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce." *National Center for Analysis of Longitudinal Data in Education Research* 106 (2014).

Landis, J. Richard, and Gary G. Koch. "The measurement of observer agreement for categorical data." *Biometrics* (1977): 159-174.

Lissitz, Robert W., and Huynh Huynh. "Vertical Equating for the Arkansas ACTAAP Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability. A Report Submitted to the Arkansas Department of Education." (2003).

Lissitz, Robert W. “A Comparison of VAM Models.” University of Maryland (2012) Available http://www.marces.org/completed/FINALTechnicalReportVAM.doc

Loveless, Tom. "The 2012 Brown Center Report on American Education: How Well Are American Students Learning? With Sections on Predicting the Effect of the Common Core State Standards, Achievement Gaps on the Two NAEP Tests, and Misinterpreting International Test Scores. Volume III, Number 1." *Brookings Institution* (2012).

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton. *Evaluating Value-Added Models for Teacher Accountability. Monograph*. RAND Corporation. Santa Monica, CA, 2003.

Mihaly, Kata, Daniel F. McCaffrey, J. R. Lockwood, and Tim R. Sass. "Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg." *Stata Journal* 10, no. 1 (2010): 82.

Neal, Derek. "Aiming for Efficiency Rather than Proficiency." *The Journal of Economic Perspectives* 24, no. 3 (2010): 119-131.

Patz, Richard J., and Lihua Yao. "Methods and models for vertical scaling." In *Linking and Aligning Scores and Scales*, pp. 253-272. Springer New York, 2007.

Polikoff, Morgan S., Andrew J. McEachin, Stephani L. Wrabel, and Matthew Duque. "The waive of the future? School accountability in the waiver era." *Educational Researcher* 43, no. 1 (2014): 45-54.

Reardon, Sean F., and Stephen W. Raudenbush. "Assumptions of value-added models for estimating school effects." *Education Finance and Policy,* vol. 4, no. 4 (2009): 492-519.

Reback, Randall. "Teaching to the rating: School accountability and the distribution of student achievement." *Journal of Public Economics* 92, no. 5 (2008): 1394-1415.

Rothstein, Jesse. "Student sorting and bias in value-added estimation: Selection on observables and unobservables." *Education Finance & Policy,* Vol. 4, Issue 4 (2009): 537-571.

Rothstein, Richard, Rebecca Jacobsen, and Tamara Wilder. *Grading Education: Getting Accountability Right*. Washington, DC: Economic Policy Institute, 2008.

Sanders, William L., and Sandra P. Horn. "The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment." *Journal of Personnel Evaluation in education* 8, no. 3 (1994): 299-311.

Sanders, William L., and June C. Rivers. "Cumulative and residual effects of teachers on future student academic achievement." (1996)

Sanders, William L., and Sandra P. Horn. "Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research." *Journal of Personnel Evaluation in Education* 12, no. 3 (1998): 247-256.

Schochet, Peter and Hanley Chiang. "What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?" *Journal of Educational and Behavioral Statistics*, Vol. 38, Number 2 (2013): 142-171.

Stuit, David, Megan J. Austin, Mark Berends, and R. Dean Gerdeman. "Comparing Estimates of Teacher Value-Added Based on Criterion-and Norm-Referenced Tests. REL 2014-004." *Regional Educational Laboratory Midwest* (2014).

Winters, M. "Why the Gap? Special Education and New York City Charter Schools." Manhattan Institute for Policy Research, Civic Report No. 80, October 2013.

Winters, M. "Why the Gap? English Language Learners and New York City Charter Schools." Manhattan Institute for Policy Research, Civic Report No. 93, October 2014.

Wolf, Patrick J., John F. Witte, and David J. Fleming. "Special choices: Do voucher schools serve students with disabilities?." *Education Next* 12, no. 3 (2012): 16.

Wright, S. Paul. "An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education."*Cary, NC: SAS Institute, Inc.* (2010).

Xu, Zeyu, Umut Ozek, and Matthew Corritore. "Portability of Teacher Effectiveness across School Settings. Working Paper 77."*National Center for Analysis of Longitudinal Data in Education Research* (2012).