

**Exploring Models of School Performance:
From Theory to Practice**

CSE Report 673

Kilchan Choi, Pete Goldschmidt, Kyo Yamashiro
University of Los Angeles, California/CRESST

March 2006

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE),
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2006 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), and the U.S. Department of Education.

This appeared as a chapter in the UC sponsored *The 104th Yearbook of the National Society for the Study of Education, Part 2-Uses and Misuses of Data for Educational Accountability and Improvement* (2005) by Joan L. Herman and Edward H. Haertel (ISSN: 0077-5762)

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.

EXPLORING MODELS OF SCHOOL PERFORMANCE: FROM THEORY TO PRACTICE

Kilchan Choi, Pete Goldschmidt, Kyo Yamashiro
University of Los Angeles, California/CRESST

Our purpose in this report is to present and discuss competing accountability approaches, or models, designed to systematically indicate how a school's students are performing academically. Within the framework of the current federally mandated accountability legislation, increased interest in models measuring school performance has caused educational policymakers to consider several key issues. These issues include whether results from different accountability models yield different inferences about a school's performance; what assumptions underlie each of the models; how different models are implemented; and ultimately which model is best suited for a particular context.

We address these issues by building a framework for accountability models and then explicitly comparing and contrasting these competing models. In order to accomplish this, we first need to examine two distinct pieces of the larger puzzle. With the first piece, we briefly summarize previous research on school performance. This is done in order to ground all of the accountability models and provide some reference for considering how an accountability model might be constructed. With the second piece, we present building blocks for accountability models. These building blocks include a) important properties of assessments, b) test metrics, c) ways of summarizing student achievement, and d) monitoring achievement growth over time; all of which need to be considered before they are incorporated into an accountability model.

Once we have the foundation and building blocks in place we can examine the continuum of accountability models, each of which results in a performance indicator. We consider the choice of model as lying on a continuum because accountability models range from simple calculations on the one end to complex statistical models on the other. At the upper end of the spectrum is a set of accountability models known as value-added models (VAM), which we compare separately. We also compare inferences based on one of these VAMs against inferences based on current federally mandated accountability models.

Examining competing accountability models and linking them back to the foundations and building blocks leads to both theoretical and practical implications that are central in considering which model is most appropriate for a given (physical and political) context. One fundamental concern is whether the accountability model can accurately capture the academic progress of under-privileged students (e.g., low socioeconomic status [SES]) and, by extension, under-privileged schools. Further, questions arise as to whether these students and schools can be fairly compared to more affluent students and schools. Based on this framework, we present empirical evidence on the effects of adjusting accountability models for student characteristics.

The report concludes with a brief summary and discussion of the salient issues surrounding accountability models.

Building a Framework for Accountability

Building a framework for accountability models requires us to briefly examine previous research on school performance. Divergent views persist on both which research methods and which research foci best capture school performance. Competing methods can be broadly categorized into qualitative and quantitative, while competing foci can be classified as emphasizing either school inputs or school outcomes. Inputs are factors such as the quality of teachers, the curriculum, and school policies and practices. Outcomes are often simply student assessment results. Both research strands and foci have implications for how an accountability model ought to be constructed. Part of the larger movement to align state policy and practice around high-quality content standards, the new accountability systems focus attention away from inputs and compliance to emphasize outcomes, or student performance on statewide standards-based assessments (O'Day & Smith, 1993). From the beginning of the standards-based movement, many have argued that if assessments, professional development, and curriculum were all aligned to rigorous academic content standards, there would be more hope for equity in schools (Goals 2000; IASA, 1994; O'Day & Smith, 1993). Increasingly, large-scale test results have become the primary indicator for measuring the condition of public education and have become the symbol of the new accountability era (Elmore, Abelman, & Fuhrman, 1996; Linn, 2000).

Not everyone agrees with this shift from inputs to outcomes. Many educators struggle against the reduction of their efforts into a single outcome indicator, contending that standardized tests cannot adequately or accurately measure school performance and believing that multiple outcome measures are needed. Some

researchers strongly advocate for school accountability models that encompass a broader spectrum of indicators than just large-scale assessment results (Oakes, 1989; Porter, 1988). Researchers argue for the need to monitor the quality and rigor of the academic curriculum (Oakes, 2002-03), the safety and cleanliness of school facilities (Oakes, 1989) and the equitable distribution of other resources, such as qualified teachers (Darling-Hammond, 2000).

Despite the continued lack of consensus, current federal accountability legislation, as enacted by the No Child Left Behind (NCLB) Act of 2001 (U.S. Public Law 107-110), emphasizes a quantitative accountability model focusing on student outcomes. In this reauthorization of education funding for low-income children (Title I), the federal accountability legislation focuses on holding schools accountable for improving performance, as measured by their success in making AYP towards 100% proficiency by 2013-14 on standards-based assessment results in reading and mathematics. NCLB requires states to incorporate into their accountability systems specific testing requirements that include annually testing at least 95% of students in all demographic subgroups in grades three through eight and once between Grades 10 and 12. Based on 2002-2003 baseline data, states set specific annual targets that culminate in 100% proficiency by 2013-2014. These intermediate targets vary from state to state. Meeting these annual targets, or making AYP, thus becomes the defacto accountability model for each state. Current NCLB legislation provides only one caveat for schools that do not meet their annual AYP target, but demonstrate at least a 10% gain in proficiency. These schools are considered compliant under the "Safe Harbor" provision.

The AYP accountability model is a high-stakes system that has several intended and unintended consequences. Novak and Fuller (2003) underline a statistical artifact of the construction of AYP in which schools that are more diverse and serve greater numbers of students from low socioeconomic standing are penalized because they have greater odds of failing to make AYP. Other researchers have criticized accountability-based test scores as highly susceptible to corruption when stakes are attached. This occurs through score inflation from familiarity with the test and a narrowing of the instructional focus by teaching to the test (Koretz, 1996; McNeil, 2000). In addition, calls have emerged from national education and civil rights groups to modify AYP to incorporate growth measures (The Alliance for Fair and Effective Accountability, 2004; National Conference of State Legislatures-NCLB Task Force Meeting, 2004). Many would like to explore alternative accountability models measuring school performance. This would be done either to validate results from the AYP model, as an alternative to

the AYP model, or to explore aspects of performance more extensively and specifically that are otherwise not addressed by AYP (Council of Chief State School Officers, Brain Trust on Value-Added and Growth Models, 2004). Questions currently raised regarding AYP point to the most salient issue with respect to all accountability models—do the results yield valid inferences concerning school performance?

Given that the current federal accountability model is based on quantitative evaluation of student outcomes, we focus our discussion on elements that should be considered in quantitative outcomes-based accountability models. By taking this route, we admittedly give qualitative methods short shrift. We do not, however, eliminate inputs from our discussion. In order to hold schools accountable for student performance, there is an implicit assumption that it is possible to isolate a school's effect from all other factors that might influence achievement (e.g., student background or inputs outside of a school's control). Without reliable and valid measures of both outcomes and inputs, attempts to estimate school performance will be limited and the inferences could be misleading.

Much of the school performance literature draws upon educational production function or school productivity traditions (Hanushek, 1979; Meyer, 1996). Underlying many of the models used in this type of research is the basic assumption that causal claims can be made about school performance. In other words, by evaluating school performance, we assume that it is possible to measure a school's quality or effectiveness—the causal effect of a school's practice on a student's achievement. Results from accountability models are based on natural, quasi-experimental designs. These designs often suffer from confounded explanations for observed results. Some argue that we would actually need to randomly assign *schools* to various *practices or policies*, in order to adequately measure the causal effect of school policies and practices on achievement (Raudenbush, 2004b). This type of design would help to rule out rival hypotheses generated by potential confounding factors, such as the selection bias that might be present because certain kinds of students and teachers choose to attend or work in some schools over others (Campbell & Stanley, 1963). Still, there is some debate regarding experimental design and causal effects in education research: e.g., whether random assignment is necessary; and fitting the appropriate design to the questions of interest (Rubin, Stuart, & Zanutto, 2004; Raudenbush, 2004a; Shavelson & Towne, 2004).

Unlike some research on program effectiveness (where experimental design is often controllable), school accountability evaluates school performance given the existing grouping of students and teachers in schools; hence, random assignment

designs are, for the most part, moot. The approach since the late 1970s has been to use a regression-based analysis within a quasi-experimental framework in an attempt to disentangle student effects from school effects. In general, the idea is to relate current student achievement to the accumulated effects of student and family background factors, the accumulated effects of school inputs (e.g., class size, number of qualified teachers, percentage of students qualifying for free lunch), and the student's innate ability prior to schooling effects (Hanushek, 1979). This approach is theoretically viable but empirically difficult because we would ultimately need an innate ability measure that is uncorrelated with schooling inputs (Griliches & Mason, 1972), which generally requires measuring ability before any schooling begins. These data are generally unavailable, since testing at early ages is not common and is not a popular concept with early childhood educators. Further, measuring the impact of the *accumulation* of student and family background, schooling, and other factors to a given time point is thought to be a highly imprecise endeavor.

The framework for accountability models implies that performance indicators must address whether they accurately reflect the school's ability to facilitate student learning through specific actions rather than merely reflecting the aggregated effects of student background. Previous research has identified specific student and school inputs that need to be considered, as well as limitations to causal claims that need to be heeded. This framework allows us to next consider specific building blocks of accountability models.

Understanding the School Performance Indicator Landscape

Many methodologists have written extensively about the relative strengths and weaknesses of various school quality indicators (Koretz, 1996; Meyer, 1996; Wilms & Raudenbush, 1989). We build on these arguments below, to lay out the landscape of school performance—the proper measures, scales, summary statistics, and standards used to measure and judge school quality. Several preliminary criteria need to be taken into account with regards to the test construction and inferences made from the test before adequate school accountability models can be developed:

- Does the test measure what it is supposed to measure (i.e., are inferences based on the test results *valid*)? How accurately, consistently, or precisely does the test measure the concept we believe is important (i.e., is the assessment *reliable* in measuring student learning)?

- Can results from the assessment be meaningfully aggregated from students to schools (i.e., if the test measures student learning, do aggregations to the school level reliably and validly measure *school performance or school quality*)?
- How are changes in performance over time measured? If the accountability system intends to measure performance or quality over time, do changes in school performance over time, as measured with results from this assessment, have meaning (i.e., do changes indicate *increases or decreases in school quality*, or do they merely represent changes in enrollment or in the scale of the test from year-to-year or grade-to-grade)?
- Should performance standards be set by some absolute or relative criteria or both?

These concerns will be discussed briefly below, though a complete treatment of these issues is beyond the scope of this report.

Test Construction and Metric Matter

Regardless of the perspective one takes on school performance, in a high-stakes accountability setting it is crucial that valid inferences about school quality can be made. Valid inferences from the results of any school accountability model depend on the data that go into the model. School accountability criteria are most commonly based on student scores on state-mandated assessments. Test results are considered *reliable* when the test in question repeatedly yields consistent results. Test results are considered *valid* when they measure what we think they are supposed to be measuring.

If results from a particular assessment are not reliable and do not generate results from which valid inferences about student achievement can be made, results will, at best, be misleading. In other words, an accountability model that is premised on standards, but selects an assessment that is not adequately aligned to standards or measures an uneven selection of standards, could falsely identify schools for doing a good or bad job of facilitating student learning towards that set of standards.

From this point forward, our discussion assumes that the test selected demonstrates reliable and valid results. Selection of such an assessment is an essential starting point for any accountability system. There are many metrics and scales to choose from when reporting assessment results: raw scores, percent correct, national percentile ranks, normal curve equivalents, and scale scores. Many of these metrics have very different qualities—qualities that impact our ability to average across

students or to draw valid inferences about school performance. Further, accountability models that attempt to make inferences regarding student achievement growth over time must ensure that assessment results are meaningfully related across grades, so that the growth is also meaningful. Below we discuss the trade-offs of using some of the most common scales on which test results are reported for school accountability purposes.

Percentiles or National Percentile Ranks

Percentiles or National Percentile Ranks (NPR) perhaps represent one of the more familiar scales used in testing, since they are used in reporting on many of the national testing and certification programs such as the Scholastic Achievement Test (SAT) and the Graduate Record Examination (GRE). Percentiles measure a student's rank, or how well a student performed relative to a national population. A score placing a student in the 85th percentile, for example, indicates that she performed better than 85% of the students in the norming population; this provides that student with a ranking. One drawback often cited about percentile ranks is that they are not measured on an equal interval scale—meaning that changes at each point on the possible range do not represent equal changes in performance. In other words, a change from the 90th to the 91st percentile may be more significant (and more meaningful) than a change from the 50th to the 51st. Conducting calculations such as averages on percentile ranks across students in a classroom or school is considered inaccurate and inappropriate, because aggregating ordinal ranks is not a meaningful concept (Russell, 2000).

Normal Curve Equivalents

Averaging normal curve equivalents (NCE) is more meaningful for school accountability purposes. This metric represents an equal interval scale, so that a change of one increment in one segment of the score range is just as meaningful or significant as a change in another segment. This allows for arithmetic calculations such as averaging to be more meaningful. Much like percentile ranks, NCEs are useful for making relative comparisons among students or schools. If the accountability system is designed to measure absolute changes in individual students' performance, however, NCEs are inappropriate because they are norm-referenced, which means they describe a relative position compared to a norming population.

Scale Scores

Scale scores are a conversion of raw scores placed on a scale that allows for averaging across students. If the focus of the accountability model is on ranking schools,

NCEs and scale score results will be virtually identical (Goldschmidt, Choi, & Martinez, 2003). Scale scores, however, enable comparisons over time at the same grade level. So, for example, a 650 in the third grade this year is comparable to a 650 in the third grade the following year. This enables us to draw more meaningful conclusions about whether a school's performance in the third grade is improving. We cannot unequivocally say, however, whether the results are due to the third grade staff and instructional practices, or to the second grade staff the year before, or to the fact that the group of students in the third grade came in more or less prepared than in previous years. One important caveat: A scale score is not necessarily scaled across different grade levels. This means that a 5-point increase in scale scores on the third grade test is not the same as a 5-point increase on the fourth grade test.

Ultimately, the optimal metric to use when examining change is a vertically equated Item Response Theory (IRT)-based scale score because it is on an interval scale and is comparable across grade levels and across time (Hambleton & Swaminathan, 1987). Thus, a change on a vertically equated scale score from one year to the next is an absolute measure of academic progress, irrespective of grade. Examining student achievement growth over spans longer than one grade level, however, can be difficult because equating is generally designed to compare contiguous grade pairs (Yen, 1986).

Despite the usefulness of scale scores, and the optimal qualities of vertically-equated IRT-based scale scores, these metrics are far less prevalent than others mentioned here and are often not available to those who might benefit the most—school personnel. Assuming these types of metrics were made available for accountability systems by test publishers, however, one of the most important questions still remains: Can valid inferences concerning school performance be drawn from summaries of our measure of student performance? This can, in part, depend on how scores will be aggregated or statistically modeled. If the metric and test issues outlined above have not been adequately addressed, however, the methods discussed below for aggregating or modeling achievement will be extremely limited.

Summarizing Achievement at the School Level

The next step is to define the model for aggregating or summarizing student-level test score information at the school level. We begin with two simple models of school performance summaries, both of which capture a snapshot of performance, are simple to compute, and are intuitively understood by the public: the percent proficient criteria and the average school achievement score. We highlight why these measures

may be inadequate for measuring school performance. Following these models, we progress to statistical models that measure change in achievement over time, or VAMs.

Percent Proficient

AYP models of school performance across the country are based upon the percent proficient in a school, which essentially reflects the percent of students who have scored at or above a particular proficiency score. Though simple and intuitive, a percent proficient indicator is, in many ways, one of the weakest indicators of performance. Percent proficient is a status indicator, and primarily measures movement around the proficiency cut score. In other words, if a student must score 680 on a test to be considered proficient, a percent proficient indicator is mostly measuring movement around that 680 score. Thus, students at the very bottom of the scale that improved from a 400 to a 600 would not count in this measure. Similarly, those students who fell from an 800 to 690 would not count in this model either. The percent proficient model misses a great deal of movement (both progress and regress) at other ends of the range, and is, for this reason, very limited. In addition, percent proficient scores—as with average or mean score discussed below—are considered “contaminated” by prior performance and other confounding factors (e.g., student and family background characteristics, selection, etc.). This means that we have no way of separating out whether or not the high percentage of proficient students in a given year is due to the teaching and learning that happened in that year or to an influx of new, better-prepared students.

Average or Mean Performance

The school average (or mean) achievement level is another oft-used measure of school performance. Similar to percent proficient scores, averages suffer from loss of information due to aggregation. In addition, averages tend to be contaminated by student background characteristics (e.g., disadvantaging schools with large populations of poor students; see Aitkin & Longford, 1986; Hanushek, 1979; Raudenbush, 2004b). As Table 1 indicates, based on correlations between school means from one year to the next, means are quite stable over time. However, it is important to note that a middle school’s sixth-grade average reading score is heavily influenced by how all of those students fared in their fifth-grade classes the year before and/or whether those students came from families who read frequently at home. Background characteristics may artificially inflate the year-to-year correlations and may have more to do with the stability of school populations than the quality of the school.

Table 1

Correlations of Mean School Performance by Year

	Math 1994	Math 1995	Math 1996	Math 1997	Math 1998	Math 1999	Math 2000	Math 2001	Math 2002
Math 1993	0.96	0.91	0.89	0.87	0.90	0.90	0.88	0.88	0.87
Math 1994		0.94	0.92	0.90	0.92	0.91	0.90	0.89	0.89
Math 1995			0.93	0.90	0.91	0.90	0.88	0.87	0.88
Math 1996				0.96	0.93	0.90	0.90	0.87	0.89
Math 1997					0.94	0.92	0.91	0.89	0.89
Math 1998						0.97	0.95	0.94	0.93
Math 1999							0.97	0.95	0.94
Math 2000								0.96	0.95
Math 2001									0.96

Aggregated, single-point-in-time status indicators that are not adjusted for confounding factors share a common weakness: changes in results from these models may have little to do with changes in school performance, per se, but may have more to do with changes in student populations. Adjusted (or conditional) means are a more precise model for summarizing school performance, but require statistical modeling, in order to compare schools with similar conditions. Adjusted or conditional models of performance will be discussed further in the next section. Once a summary statistic is selected (e.g., adjusted means), questions still remain about how changes in these summary statistics will be measured and reported over time.

Tracking Changes in Performance Over Time

Assuming we agree on the test, the metric, and the summary statistic, those designing an accountability system must carefully consider how changes in performance over time will be taken into account. Detecting changes in school performance over time requires yet another set of considerations. Users of performance accountability information must first determine whether they believe that judgments about school performance should be based solely on a school's *status*—the level at which students are currently performing—or based on a school's *growth*—the improvements in students' performance over time—or based on a combination of both status and growth. For instance, the simplest model for tracking changes in performance over time has historically been to track a series of *status* indicators (e.g., school averages or percent proficient across multiple years). This type of "time series" data would allow for trend data (from year-to-year) to be reported, but would attach with it many of the concerns raised above because the summary statistic chosen (e.g., unadjusted averages) is confounded by factors other than school performance. Time series data using adjusted averages would also be an option for tracking change over time.

However, other more sophisticated models use "panel data," which links individual students from year-to-year. This panel data enables longitudinal studies of student growth over time. Models that use growth based on individual student changes represent more appropriate ways of measuring change over time, and can control for individual differences over time as well. There are several ways of combining status and growth measures in one accountability system, though the intricacies of including

both kinds of information is still a work in progress.¹ AYP is one such way, setting annual growth targets toward an absolute benchmark of 100% proficiency. It is important to note that, though AYP tracks changes over time by setting intermediate targets, it is not a model that can accurately measure growth because changes are not based on the same individuals over time. In essence, AYP employs growth targets within a *time series* framework, carrying with it many of the drawbacks discussed.

Comparing Performance Standards

In order to make qualitative judgments about the changes in performance that are detected, policymakers and practitioners must make clear choices about the standards they set for performance. Whether performance is measured based upon attainment at a single point in time or multiple time points, criteria for success (the performance standard) must be set. Should a school's performance be judged against an *absolute* standard of performance (100% proficient or e.g., a gain of 30 points or more), or based on a *relative* standard of performance (highest score among schools starting out in the first decile). Figure 1 compares many of the performance standards alluded to thus far. School A would be considered the top performer from an absolute *status* standard of performance, as well as relative to the rest of the sample, while School B would be the top performer from an absolute *growth* standard of performance, as well as relative to the rest of the sample. We would argue that *both* absolute and relative criteria are essential components to measuring and judging performance.

The kind of growth that we suggest is more appropriate for high-stakes accountability settings requires statistical models, and falls under the general framework of VAMs.

¹ Other methods involve statistical models that measure a school's or student's distance to target and calculate the probability of meeting the absolute target, given their growth to-date (see Doran & Izumi, 2004; Thum, 2003b).

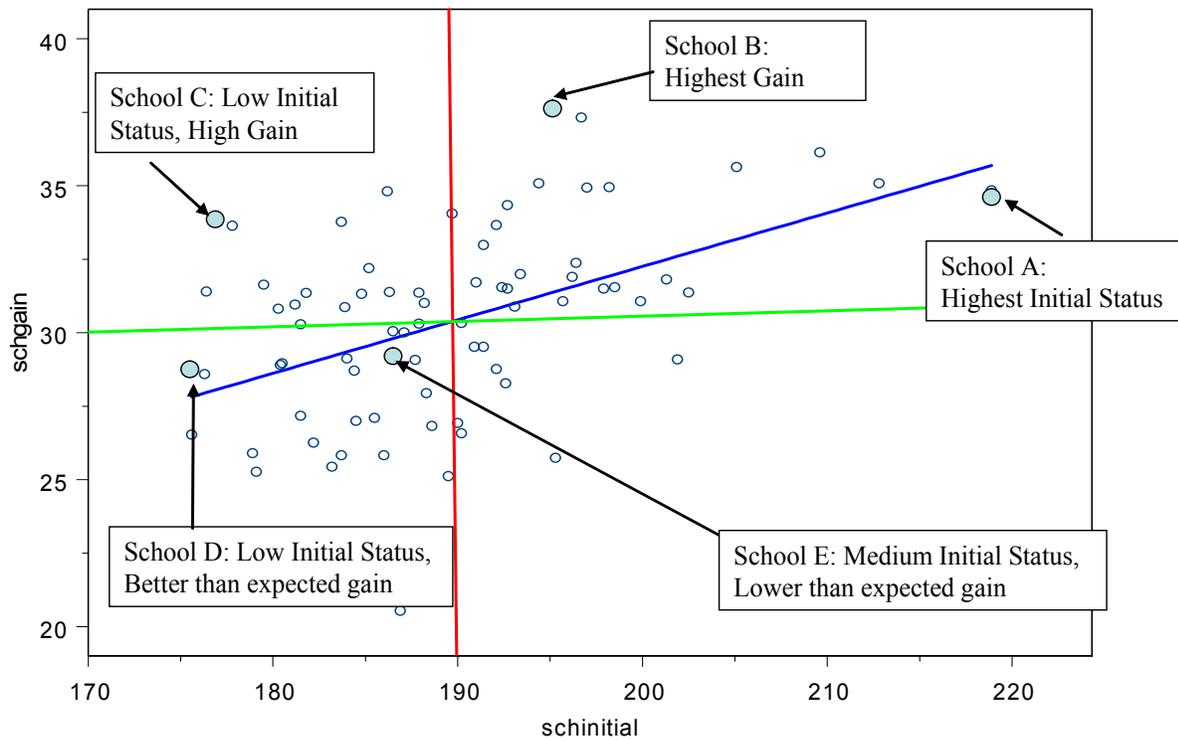


Figure 1. Effect of different criteria for determining school performance: School rank by different criteria.

- 1) Status (x-axis): $A > B > E > C > D$
- 2) Gain (y-axis): $B > C > A > E > D$
- 3) Conditional gain (regression line): $B > C > D > E > A$

Note that the vertical line and the horizontal line represent district average initial status and district average gain, respectively.

Value-Added Models: Measuring Growth

VAMs represent a specific subset of accountability models that provide estimates of students' progress over time. Student gains can simply be measured with *actual* gains between two time points on a given measure (e.g., Student A's score in Year 2 minus his or her score in Year 1).² However, when student gains are summarized to the school level for estimates of school growth, this *actual gain* can be an incomplete measure of school change, and ignores the nested structure of students within classrooms and classrooms within schools. For this reason, we define *school value-added models* to be regression-based models, *estimating* school gains between at least two time points, typically using multi-level models.³

Most VAMs either implicitly or explicitly compare a school's gain against other like, or similar, schools. Comparing "like with like" can mean that schools are compared with other schools that started out at similar levels of performance (e.g., those that start in the lowest 10%), or that serve similar populations of students (e.g., 85% English Language Learners). Many would argue that comparing "like with like" makes as much sense philosophically as it does methodologically (Goldstein & Spiegelhalter, 1996; Rumberger & Palardy, 2004). Researchers and educators alike acknowledge that adjusted models may be necessary to level the playing field (e.g., Good, Biddle, & Brody, 1975; Linn & Slinde, 1977; Thum, 2003a). To not use these adjustments in high-stakes accountability settings may risk exacerbating teacher and principal distribution issues. That is, highly qualified teachers and principals may stay away from "difficult" schools, knowing that they will not be adequately acknowledged for the difficult situation they are tackling. Further, the implicit assumption for all VAMs is that anything not explicitly accounted for in the model (e.g., background characteristics) will be absorbed in what is left over, which constitutes the value-added estimate. Leaving these characteristics unadjusted for would, in effect, "contaminate" the value-added estimate.

² These actual gains are often described as the "unbiased estimate" of a student's gain in achievement (Bryk & Weisberg, 1976; Rogosa, 1995; Rogosa, Brandt, & Zimowski, 1982).

³ Multi-level decomposition of variance within and between schools is considered state-of-the-art by many researchers, given that it recognizes the non-random sampling within schools—that students in a particular classroom or school are more like each other than students in any other school in the sample.

As an example of what we mean by comparing similar schools, if a given school's gain is higher than expected (e.g., higher than the average gain for all schools starting out at that same level), the school can be said to have added value, as exemplified by School C in Figure 1. Note under this criteria that School A is no longer a top performer. The concept that schools make gains but do not add value is not intuitive. We emphasize that schools add value when their gains are greater than a set criteria that can be either absolute or relative. Absolute criteria are those developed a priori, such as an external target; while relative criteria are data driven, such as the district average (the horizontal line in Figure 2) or regression-based expected gains (the fitted regression line in Figure 2). Without an external absolute standard for school gains and value added estimates, deciding whether either is *sufficient* is difficult. More research on value-added standards setting is needed.

The simplest VAM—gain score—is defined below, with specific examples of more complex VAMs following.

Gain Scores or Adjusted Gains

We have defined VAMs to encompass those models that estimate gains through statistical modeling; this includes gain score and adjusted gain models. Both approaches model the outcome as a function of a set of inputs, which might include student or school background characteristics (if they are adjusted gains). Unlike a typical regression model where the outcome of interest is a score from a particular year, in gain score models, the outcome of interest is the change in scores from one time point to the next ($Y_2 - Y_1$). Adjusted gains models are motivated by the aforementioned notion of comparing like with like and, as a result, the results provide a relative measure of gain, conditioned by the set of variables chosen for the model, such as student or school background variables or initial starting points.

Many believe—and studies have often shown—that there can be an important relationship between where a student starts out and how much he or she grows (see Choi & Seltzer, 2004; Muthen & Curran, 1997; Raudenbush & Bryk, 2002, chpt. 11; Seltzer, Choi, & Thum, 2003). To address this relationship, some use the initial starting point as a predictor in the gains model. Using the starting point as a predictor raises

some concerns, since Year 1 scores contain measurement error.⁴ Ignoring the measurement error leads to attenuated estimates of gains. This measurement error can be addressed through the modeling process.⁵ Gain score models are perhaps the most common type of VAM. More complex modeling processes have also been employed in some high profile settings in recent years; examples of such models are discussed below.

Some Current Value-Added Examples

We present some of the strengths and weaknesses of commonly used VAMs and present empirical results from some of our own work in this area. Though the models below are more similar to each other than they are to, say, an AYP model, each has a slightly different focus and takes a slightly different approach to: what to adjust for and how; scaling issues and measurement error; assumptions about (linear) growth patterns; treating school effects as varying across schools or fixed; and how to represent cumulative effects of schools over time. We examine the following models and their approach to addressing these modeling issues:

- Sanders’s teacher effectiveness model (Tennessee Value-Added Model)
- McCaffrey’s more general version of Sanders’s model (RAND model)
- Chicago Public School Productivity Model (CPSP)
- CRESST Student Growth Distribution model (CRESST Model)

Sanders’s Model

This model—often referred to as the TVAAS, the Tennessee Value-Added Assessment System (Sanders, Saxton, & Horn, 1997)—incorporates multiple content measures as well as multiple cohorts and panel data (linked student data over time). It does not assume simple linear growth, and it implicitly adjusts for prior achievement by using the gain score method. That is, prior years’ achievement is layered onto each subsequent year’s achievement. As such, this model assumes that the effect of prior

⁴ Measurement error is the degree of uncertainty or variation we are likely to find in any measurement we take; test scores are no exception, since they represent an approximation of the skill or concept they are attempting to measure.

⁵ Some researchers do this through latent variable modeling, where the initial starting point is a latent variable, estimated within the model.

years' achievement remains stable as time goes by. Sanders's approach does not attempt to model the interaction between where a school starts and how much it grows. The model does not adjust for student background characteristics, though an extension of the model proposed by Ballou, Sanders, and Wright (2004) uses a two-stage procedure to adjust for differences in student and school background characteristics. Some of the challenges of the model are its tremendous requirements for data and computing capacity and its cost.

Further, because of proprietary estimation procedures, more broad applications of this model are not currently possible.

McCaffrey et al's RAND Model

This model specifies a more general form of Sanders's model, i.e., a multivariate, longitudinal mixed model that incorporates the complex nesting structure of student longitudinal data linked to teachers (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). This model allows one to include adjustments for student and school characteristics. In addition, it does not assume that prior teacher effects are persistent or static (an assumption of Sanders's model), which allows us to test whether prior teacher effects differ from current teacher effects on current student outcomes and to what degree teacher effects fade with time. Some of the shortcomings both in this model and Sanders's model are that ranking teachers using this model is likely to be a challenge, because obtaining precise estimates of teacher effects requires a large data set (Ballou et al, 2004; also see Lockwood, Louis, & MaCaffrey, 2002). The RAND study found that both McCaffrey's and Sanders's model identified only one-fourth to one-third of the teachers as distinct from the mean. This problem is, in part, due to the fact that the two approaches are modeling teacher effects, not school effects, and are therefore hampered by smaller sample sizes per classroom. The implication of these findings is that making distinctions between teachers based on their value-added estimates could be challenging. This model also does not address the interaction of where a school starts and how much it grows. In addition, wider application of this model could be limited by the complex estimation procedure used, model convergence problems, available software, and extensive computing time.

Chicago Public School Productivity Model

This model uses a "productivity profile" for the Chicago Public Schools, in which initial status trends (input trends) and gain trends (learning gain trends), as well as output trends (adding input and learning gain together), are estimated for each grade

level (Bryk, Thum, Easton, & Luppescu, 1998). Productivity is judged by both the learning gain trend and the output trend, to address situations where gains are up but inputs and outputs are down (gap is spreading). One of the key strengths of the Chicago model is that the system is designed around a well-designed testing system (e.g., vertically-equated test scores). This model does estimate both initial status as well as the gain, and can be applied to multiple subject areas and multiple cohort data. The model also explicitly takes measurement error into account. It is possible to adjust for student- and school-level characteristics. The model can be fitted using an accessible and well-known software program (HLM6, 2004). However, this model does not specify modeling gain trends as a function of initial status trends.

CRESST Student Growth Distribution Model

The CRESST model uses latent variable regression in a hierarchical modeling framework (Choi, Seltzer, Herman, & Yamashiro, 2004; see also Choi & Seltzer, 2004 and Seltzer et al., 2003). The key strengths of this model are that it estimates average school growth as well as the distribution of student growth within a school by explicitly modeling student growth as a function of a student's initial status. Modeling this relationship between where a student starts out and how much he or she grows allows us to provide complementary information about how equitably student growth is distributed within a school for particular performance subgroups (e.g., above or below average performers). Like the CPSP model, the CRESST model explicitly takes measurement error into account where standard errors of measurement are available. Similar to the other models here, the CRESST model can be extended to incorporate multiple measures and multiple cohorts; it is possible to adjust for observed student- and school-level covariates; and it is possible to adjust for student and school initial differences. Computation limitation with very large data sets might be a shortcoming of this model.

Two Empirical Studies

Below we present empirical results that address some of the more salient modeling and philosophical issues discussed above. The first study explores a comparison of results from status versus gains models, comparing performance classifications based on AYP results to those based on results from a VAM. The second study compares school performance results from an array of VAMs using different types of adjustments for background characteristics.

Comparing Performance Classifications Based on Status and Gain

We illustrate how results based on AYP and VAMs can differ based on an alternative way of measuring progress using the CRESST VAM as described in the previous section. We do so through analyses of a longitudinal dataset from an urban school district in the Pacific Northwest. The outcomes of interest in this analysis are Iowa Test of Basic Skills (ITBS) reading scores for third graders in 2001 and those same students' fifth-grade scores in 2003. We examine what these growth patterns tell us about schools that meet their AYP targets and those that do not.

Our focus is to compare the state's AYP classifications (e.g., met or did not meet AYP) against the results that we obtained by fitting each school's data to the expected gains for each of the three different levels of initial status (i.e., expected gain for students 15 points below the school mean, expected school mean gain, and expected gain for students 15 points above the school mean). First, as can be seen in Figure 2, 15 schools among 51 schools meeting AYP (AYP Schools) have an estimated gain that is smaller than the district average gain. Furthermore, among the remaining 36 schools, only 12 have gains that are statistically greater than the district mean gain. In contrast, almost half of the schools not meeting AYP (Non-AYP Schools; in Figure 3) have gains that are higher than the district average. Among those, we have two exceptionally well performing schools. Schools #23 and #55 have gains that are higher than the district mean gain by approximately 5-10 points and are statistically greater than the district average gain. As such, we clearly see that there are remarkably large numbers of non-AYP schools making sizable (and close to district average) gains.

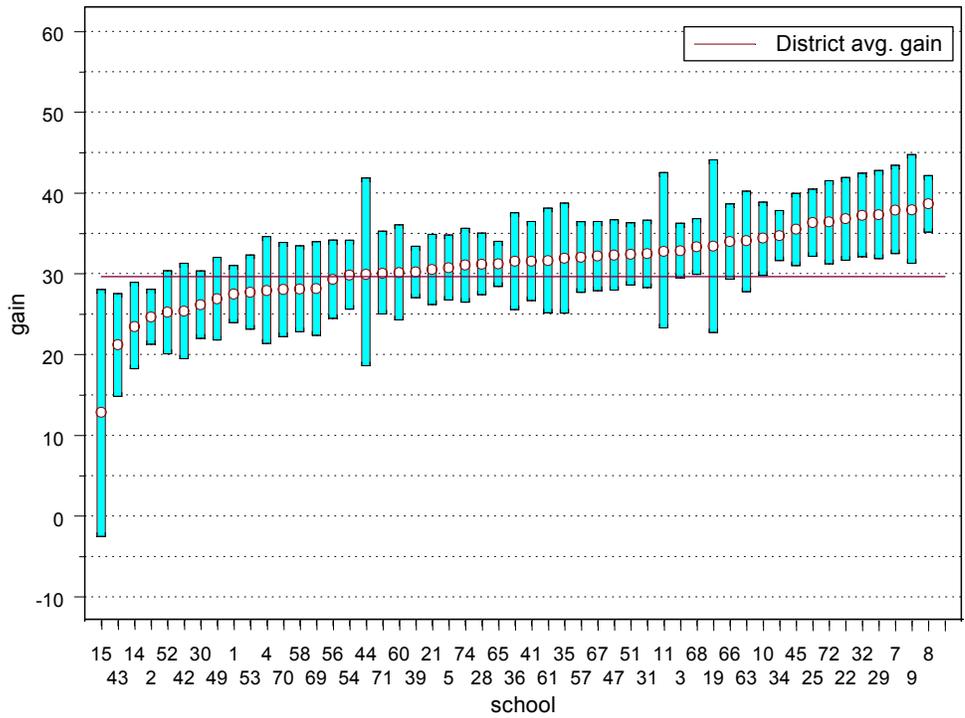


Figure 2. Expected mean gain in ITBS reading scores for AYP schools. The horizontal line represents the district average gain for our sampled 72 schools. The top line, middle circle, and bottom line of each bar represent the 95% confidence interval around the expected mean gain for a given school.

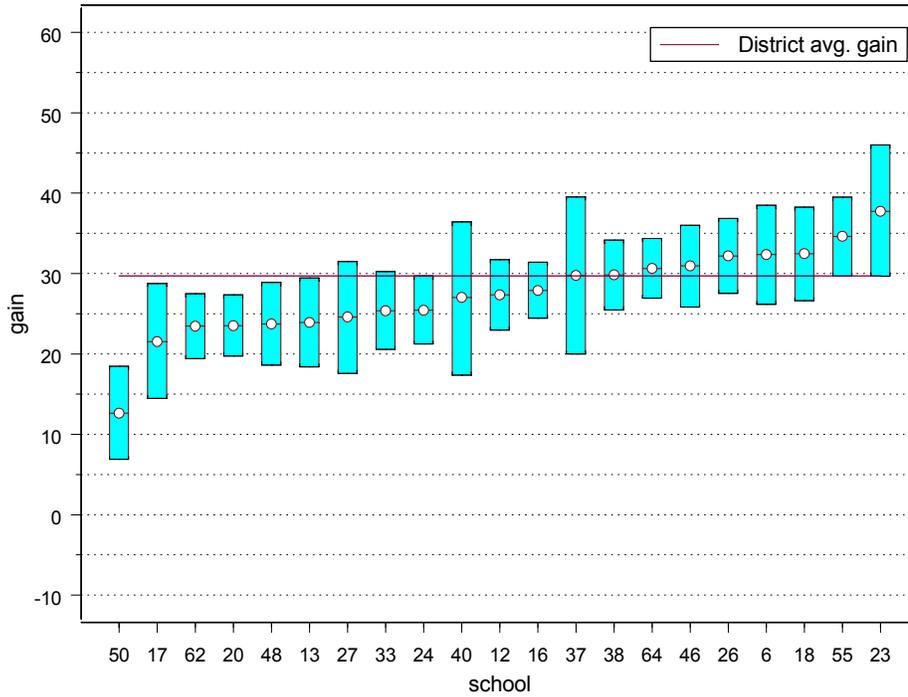


Figure 3. Expected mean gain in ITBS reading scores for Non-AYP schools. The horizontal line represents the district average gain for our sampled 72 schools. The top line, middle circle, and bottom line of each bar represent the 95% confidence interval around the expected mean gain for a given school.

Next, we compare AYP Schools and Non-AYP Schools based on the magnitude of the expected gain for students 15 points below and 15 points above the school average. Analyzing growth patterns among these subgroups allows us to note another trend, one that involves the closing or widening of achievement gaps. As can be seen in Figure 4, Type I schools show growth trajectories for the three performance subgroups that are parallel, meaning that all subgroups are growing at the same, upward rate. Type II and Type III schools show growth trajectories that either converge or diverge.

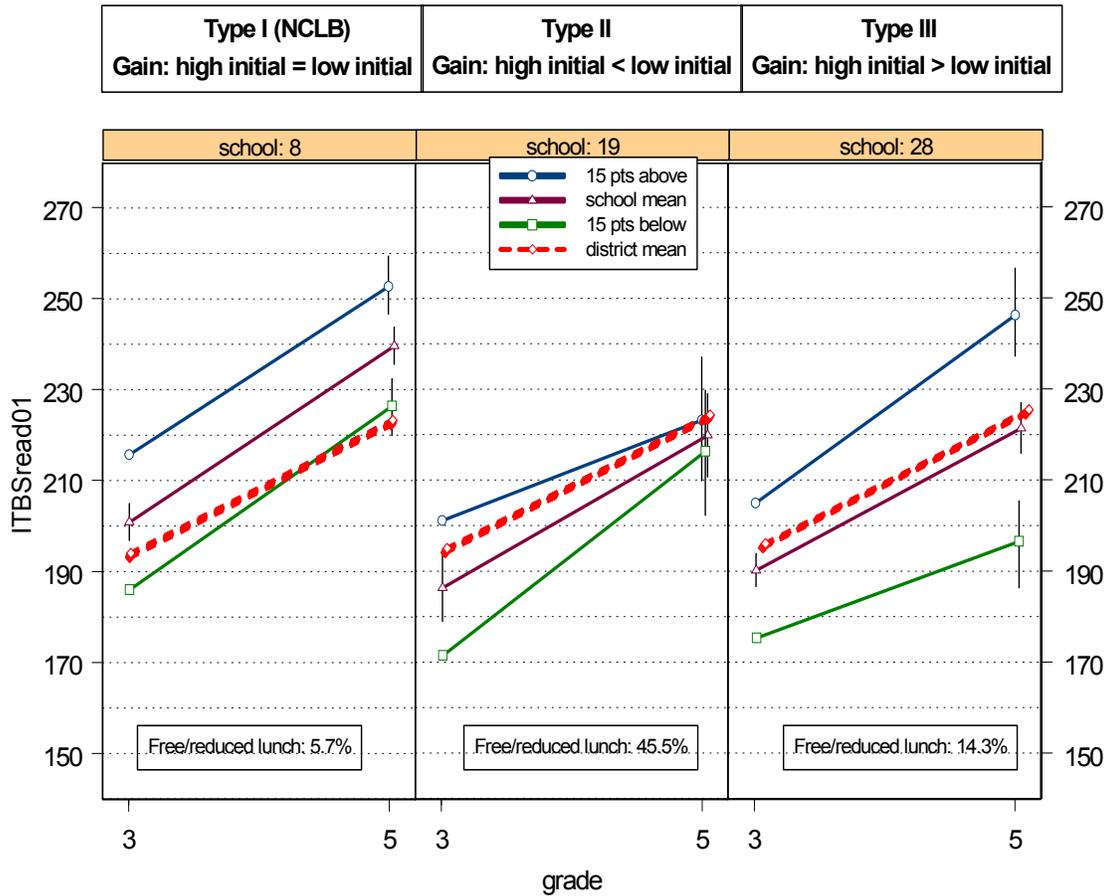


Figure 4. Distribution of student gain. *Type I School*: NCLB School with a substantial gain of more than 30 points across *all* performance subgroups. *Type II School*: Low adequate gain (less than 30 points) for students initially starting 15 points above the average, while substantial gain (larger than 30 points) for students initially starting 15 point below the average. *Type III School*: Low adequate gain (less than 30 points) for students initially starting 15 points below the average, while substantial gain (larger than 30 points) for students initially starting 15 point above the average.

In Type II schools, the initial gap between above average and below average students in the school closes over time, while the initial gap gets magnified in Type III schools. While this phrase is most often used to refer to the racial gap in achievement that exists in many schools, Type II schools in our sample are making progress in closing the gap between an important pair of subgroups—those who start out below average and those who start out above average. When evaluating a school’s effectiveness or quality, however, it is difficult from an ethical point of view to choose the better of Type II or Type III schools. From a policy point of view, more resources

and sanctions (e.g., through Title I, NCLB) are often targeted toward helping the lowest performing students improve. Thus, it could be argued that those schools succeeding with below average groups (e.g., the Type II schools) are more successfully addressing the spirit of those legislative initiatives. Models like the CRESST model, which analyzes different performance subgroups' growth, might be more effective in stimulating discussion among teachers and administrators to identify a school's success or weakness in addressing the needs of struggling students so that appropriate interventions can be taken (Seltzer et al., 2003).

Adjusting for Background Characteristics in Value-Added Models

We return to the issue of omitted variables and their effect on interpreting model results. Methodologists and education policy researchers continue to search for and debate the most appropriate ways to adjust for background characteristics. Some researchers describe the type of school effect measured in a multilevel setting in terms of adjustments made for school-level or student-level background characteristics (Raudenbush & Wilms, 1995; Wilms & Raudenbush, 1989).

We examined adjustment practices through a series of models, comparing the school performance results from each (Choi, Yamashiro, Seltzer, & Herman, 2004). The models included:

1. Three-level Unconditional Gains (no adjustment)
2. Adjusting for Student Initial Status
3. Adjusting for Student Socioeconomic Status
4. Adjusting for Student Initial Status + Student SES
5. Adjusting for Student Initial Status + Student SES + School Mean Initial Status
6. Adjusting for Student Initial Status + Student SES + School Mean Initial Status + School Mean SES

Each model represents a step in adjusting for background characteristics. Models 2 through 4 are adjusting for only student background characteristics. In contrast, Models 5 and 6 are adjusting for not only student background characteristics but also for school contextual effects. Schools were ranked according to the gain estimated from each model and correlations were calculated across these rankings (See Table 2). The first model (Model 1, with no adjustments) serves as a reference point.

Table 2

Correlation between Value-Added Estimates for Different Levels of Adjustments

	No Adj	Student IS	Student SES	Student IS_SES	Student IS_SES & School IS	Student IS_SES & School IS-SES
No Adjustment	1					
Student IS	0.93	1				
Student SES	0.98	0.97	1			
Student IS_SES	0.91	1.00	0.96	1		
Student IS_SES & School IS	0.86	0.98	0.92	0.99	1	
Student IS_SES & School IS-SES	0.75	0.88	0.85	0.90	0.90	1

Note: IS = Initial Status (initial starting point), SES = Socioeconomic Status

The correlations above demonstrate several important findings about adjusting for student- and school-level background characteristics. First, adding in an adjustment for student SES (as measured by eligibility for free- or reduced-price lunch) adds very little once a student's initial status is controlled. The rank orderings from these two models (Models 2 and 4) are perfectly correlated (1.00). This indicates that student initial status captures many of the effects that SES is attempting to measure. In other words, by controlling for initial status, the model already captures the preceding effects that SES might have on students. Once school-level contextual effects are added into the model (mean initial status and mean SES, as measured by percent qualifying for free- and reduced-price lunch), the rank ordering of schools in terms of their estimated gains differs significantly. After adjusting for the student-level characteristics (student initial status and SES), the school contextual effect of school mean initial status turns out to be insignificant, while the contextual effect of mean SES is significant. The correlations in Table 2 demonstrate this, as the results from adding in school mean initial status

correlate highly (0.99) with results from the model without school mean initial status (Model 4 and 5); whereas, the results from the addition of school mean SES and initial status are not as highly correlated with the model with no school-level effects included (Model 4 and 6). Some researchers caution that adding in school mean SES may be over-adjusting and removing some of the school policy and practice effects we are attempting to estimate (Raudenbush, 2004a). In summary, this result indicates that there is much sensitivity to the type of school and student background characteristics adjusted for in the model, and how these adjustments are made.

Conclusions

In the current high stakes accountability environment, attention should be paid to the fact that, no matter what accountability model is chosen, performance indicators are merely *estimates* of true school performance. Using statistical estimates of growth, as opposed to raw data—whether means, percent proficient, or raw year-to-year gain scores—without adjustment produces misleading estimates of school performance. These estimates are unduly influenced by movement around the proficiency cut score or average, and can be attributed to chance or changes in enrollment, rather than true student achievement growth or school performance.

We believe that VAMs provide both the most informative and the most valid picture of school performance. Information gleaned from growth patterns based on individual students is more conducive to principal, teacher, or student use as formative and diagnostic tools about progress being made. Results of VAMs potentially lead to more valid inferences because they can account for differences in student initial academic status or other known achievement-moderating factors such as SES or language proficiency.

Despite the potential benefits of VAMs, there are still political and logistical obstacles to implementing these models. The most prominent political obstacle to value-added and longitudinal growth models is the NCLB emphasis on measuring school performance with a percent proficient indicator. Except for the Safe Harbor provision, the AYP model is not a true growth model, as it entails making a series of yearly static comparisons of actual to targeted proficiency levels.

The logistical obstacle is that most state assessment data systems are not currently designed to enable longitudinal analyses. Key features required to be in place, at a minimum, would be unique student identifiers to track students longitudinally; testing in multiple (if not all) grades per school level (elementary, middle, high school);

vertically equated test scores (to compare scores across grades as well as over time); and consistency in testing from year-to-year (no changes in tests) and across grade and subject areas (no alternating between reading in Grade 3 and math in Grade 4). Though state assessment systems have not typically incorporated the key features needed to conduct longitudinal growth analyses based on individual gains, NCLB testing requirements will go a long way in bolstering state assessment systems in this regard.

Until the logistical challenges posed by the structure of state assessment systems and technical issues with regard to test construction are resolved, we suggest that models of school performance using student input adjustments are preferable to models without adjustments. However, state and other advocacy groups should continue to insist that alongside status indicators, measures of growth (gains) should be incorporated into future accountability models. Combining or weighing results from different performance models (e.g., status versus gains) is no easy task. In addition, more work must be done on the development of performance standards, both absolute and relative. Policymakers and methodologists will need to work together to answer some of the technical and philosophical questions behind many of the questions raised in this paper, basing their conclusions on sound theory and a clearly defined purpose.

The technical differences between VAMs are important, and may impact our judgments of school quality. However, these technical differences are often ultimately confusing to practitioners and turn them away from the many advantages these models have to offer. It is important to note, however, that the differences between these VAMs are much smaller than the differences between all VAMs and AYP models. Making judgments about a school's quality should be done with much reserve and with great attention to how confident we feel about the measures, metrics, and models we have chosen to represent school quality. No single model will be able to address every possible concern; rather, judicious use of results from multiple performance models allow us to look at what is surely a multi-dimensional construct—school quality—through a multi-dimensional perspective.

References

- Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, 149(1), 1-43.
- The Alliance for Fair and Effective Accountability. (October 21, 2004). *Joint Organizational Statement on No Child Left Behind (NCLB) Act*.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-65.
- Bryk, A.S., Thum, Y.M., Easton, J.Q., & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education*, 2, 103-142.
- Bryk, A.S., & Weisberg, H I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1, 127-155.
- Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for Research*. Chicago: Rand McNally.
- Choi, K., & Seltzer, M. (2004). Modeling heterogeneity in relationships between initial status and rates of change: Treating latent variable regression coefficients as random coefficients in a three-level hierarchical model. *Journal of Educational and Behavioral Statistics*. Under review.
- Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2004). *Children left behind in AYP and non-AYP Schools: Using student progress and the distribution of student gains to validate AYP*. CSE Technical Report. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing. University of California, Los Angeles.
- Choi, K., Yamashiro, K., Seltzer, M., & Herman, J. (2004). *Comparing like with like: The role of student and school characteristics in value-added models*. Paper presented at the Center for Research on Evaluation, Standards and Student Testing (CRESST) Annual Conference, September, 2004.
- Council of Chief State School Officers (2004). *Brain Trust on Value Added Growth Models*, Washington, DC, November.

- Darling-Hammond, L. (2000, January). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, 8(1). Retrieved October 10, 2004 from <http://olam.ed.asu.edu/epaa/v8n1/>
- Doran, H., & Izumi, L. (2004). *Putting education to the test: A value-added model for California*. San Francisco: Pacific Research Institute.
- Elmore, R.F., Abelman, C.H., & Fuhrman, S.H. (1996). The new accountability in state education reform: From process to performance. In H.F. Ladd (Ed.), *Holding Schools Accountable: Performance-based Reform in Education* (pp. 65-98). Washington, DC: The Brookings Institution.
- Goals 2000: Educate America Act of 1994, Pub. L. 103-227, Stat. 125
- Goldschmidt, P., Choi, K., & Martinez-Fernandez, F. (2003). *Using hierarchical growth models to monitor school performance over time: Comparing NCE to scale score results*. CSE Technical Report. Los Angeles, CA: Center for the Research on Evaluation, Standards, and Student Testing, University of California Los Angeles.
- Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159, 384-443.
- Good, T.L., Biddle, B. J., & Brody, J.E. (1975). *Teachers make a difference*. Lanham, MD: University Press of America.
- Griliches, Z., & Mason, W. (1972). Education, income, and ability. *Journal of Political Economy*, 80 (3), Part 2 (May-June), S74-S103.
- Hambleton, R. K., & Swaminathan, H. (1987). *Item response theory: Principles and Applications*. Boston, MA: Kluwer.
- Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of education production functions. *Journal of Human Resources*, 14 (3), 351-388.
- Koretz, D. (1996). Using student assessments for educational accountability. In E.A. Hanushek & D.W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171-195). Washington, DC: National Academy Press.
- Linn, R.L., & Slinde, J.A. (1977). The determination of the significance of change between pre- and post-periods. *Review of Educational Research*, 47, 121-150.
- Linn, R.L. (2000) Assessments and accountability. *Educational Researcher*, 29(2), 2000.

- Lockwood, J.R., Louis, T.A., McCaffrey, D. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, 27, 255-270.
- McNeil, L.M. (2000) *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.
- McCaffrey, D., Lockwood, J.R., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.
- Meyer, R.H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D.W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197-223). Washington, DC: National Academic Press.
- Muthen, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power. *Psychological Methods*, 2, 371-402.
- National Conference of State Legislatures (2004). NCLB Task Force Meeting, Portland, Oregon: October 10-11.
- Novak, J., & Fuller, B. (2003). *Penalizing diverse schools? Similar test scores, but different students, bring federal sanction*. PACE Policy brief. Berkeley, CA: Policy Analysis for California Education.
- Oakes, J. (Summer, 1989). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis*, 11(2), 181-199.
- Oakes, J. (2002-2003). Ensuring equity in college preparation: What K-16 partnerships can do. *Teaching to change L.A.'s school report card*, 3(1-7).
- Oakes, J. (2003) *Education inadequacy, inequality, and failed state policy: A synthesis of expert reports prepared for Williams v. State of California*. Retrieved September 12, 2004 from www.mofo.com/decent-schools/expert_reports/oakes_report.pdf
- O'Day, J., & Smith, M. (1993). Systemic reform and educational opportunity. In S. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 250-312). NY: Jossey-Bass.
- Porter, A. (March, 1988). Indicators: Objective data or political tool? *Phi Delta Kappan*, 69(7), 503-508.

- Raudenbush, S.W. (2004a). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121-129.
- Raudenbush, S.W. (2004b). *School, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing Service, Policy Evaluation and Research Center.
- Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Newbury Park, CA: Sage Press.
- Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2004). *HLM6: Hierarchical linear and non-linear modeling*. Homewood, IL: Scientific Software International.
- Raudenbush, S.W. & Wilms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307-335.
- Rogosa, D.R. (1995). Myths and methods: Myths about longitudinal research, plus supplemental questions. In J.M. Gottman (Ed.), *The analysis of change* (pp 3-65). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Rogosa, D.R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 92, 726-74.
- Rubin, D., Stuart, E., & Zanutto, E. (2004). A potential outcomes view of value-added assessment in education, *Journal of Educational and Behavioral Statistics*, 29(1), 103-116.
- Rumberger, R.W., & Palardy, G.J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 235-258). Thousand Oaks, CA: Sage Publications.
- Russell, M. (2000). Summarizing change in test scores: Shortcomings of three common methods. *Practical Assessment, Research & Evaluation*, 7(5), <http://pareonline.net/getvn.asp?v=7&n=1> .
- Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin.

- Seltzer, M., Choi, K., & Thum, Y.M. (2003). Examining relationships between where students start and how rapidly they progress: Implications for conducting analyses that help illuminate the distribution of achievement within schools. *Educational Evaluation and Policy Analysis*, 25(3), 263-286.
- Shavelson, R.J., & Towne, L. (2004). What drives scientific research in education? *American Psychological Society*, 17(4), 27-30.
- Thum, Y.M. (2003a). Measuring progress toward a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods and Research*, 32(2), 153-207.
- Thum, Y.M. (2003b). *No Child Left Behind: Methodological challenges & recommendations for measuring adequate yearly progress*. CSE Technical Report. Los Angeles, CA: Center for the Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
- U.S. Public Law 107-110. 107th Congress. 2nd Session. 8th October, 2002. The No Child Left Behind Act of 2001.
- U.S. Public Law 103-227. 103rd Congress. 2nd Session. 31st March, 1994. Goals 2000: Educate America Act of 1994.
- U.S. Public Law 103-382. 103rd Congress. 2nd Session. 10th October, 1994. Improving America's Schools Act of 1994.
- Wilms, J.D., & Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26(3), 209-232.
- Yen, W.M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23(4), 299-325.