Abstract Title Page
Not included in page count.

Title: "*Defining School Value-Added: Do Schools that Appear Strong on One Measure Appear Strong on Another?*"

Author(s): Allison Atteberry

Abstract Body
Limit 5 pages single spaced.

Background / Context: *Description of prior research and its intellectual context.*

This project considers the common practice of using a single measure of school value-added as a sufficient characterization of the effectiveness of public high schools. By making the definition of a school's *causal effect* more explicit—and then modifying that definition—I explore whether a single value-added measure is suitable for policy purposes.

Beginning in 2002, *No Child Left Behind* (NCLB) required every state to implement a test-based accountability system. According to federal guidelines, each state articulated how schools experience sanctions when their performance on state-selected standardized tests consistently falls below a specified level. Yet precisely *how* test score data are translated into judgments about the effectiveness of teachers and/or schools[1] is a highly contentious matter.

In response to problems with the current approach of using a proficiency threshold,[2] researchers have begun to develop an alternative approach for translating test scores into assessments of schools, generally referred to as "value-added" (VA) modeling.[3] VA takes into account pre-existing student characteristics including (but not limited to) prior achievement so that, in theory, the "value" added by schools can be statistically disentangled from the characteristics of students that are outside the school's control. Several school systems are beginning to opt for some instantiation of "value-added" assessments of their schools, as opposed to the simple AYP model (e.g., districts in North Carolina, Tennessee, Chicago, New York, and Washington DC). Precisely what those school value-added estimates capture, however, often remains opaque.

The estimation of value-added can be framed as an attempt to more closely capture a *causal inference* about the effect of attending one school versus another (e.g., Reardon & Raudenbush, 2009; Rubin, Stuart, & Zanutto, 2004). Given the much-documented challenges of estimating causal effects in observational education settings, researchers have begun to use this framework to more closely investigate the statistical properties of purported school or teacher effect estimates (Baker, et al., 2010; Harris, 2009; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Reardon & Raudenbush, 2009; Rothstein, 2009).

However even beyond those concerns, I argue herein that the term "value-added" obscures fundamental choices about what the causal effect of interest *is* in the case of public schools. Furthermore, I demonstrate the extent to which any attempt to compare schools' effectiveness depends on precisely how one defines "value-added." Indeed, the term value-added is by itself quite vague. One might well wonder: Value-added… of what exactly? In comparison to what? As measured by what? For what kinds of students?

The extant body of literature on causal effects in social sciences is quite useful for articulating this problem. Given that "value-added" strives to capture a causal inference about the *effect of attending a given school,* the problem alluded to above can be framed as an underspecified definition of the causal effect we wish to estimate. Researchers in this area have converged upon a more complete definition of a *causal effect*, in which one must specify all of the following elements:

(1) The treatment of interest,
(2) The alternative to the treatment,
(3) The population of interest,
(4) The outcome on which the effect is estimated, and
(5) The timeline of exposure to the treatment.[4]

A researcher who seeks to capture schools' value-added (either explicitly or implicitly) makes choices about each of these elements of the causal effect. Thus, obtaining valid and stable

estimates of a school's capacity to add value is not simply a matter of statistical computation. While some of the methods being explored today require a high level of statistical expertise, this alone does not ensure that the estimates capture the desired latent phenomenon needed for accountability policies.[5]

The question arises: can one formulation of value-added capture the various *effects* we expect schools to *cause*? It is possible, for instance, that one specification of the causal effect misses that a particular school adds value in some subject areas (or for some students, or on some outcomes) but not others. Alternatively, perhaps schools that appear to excel according to one measure of value-added tend to excel on almost *any* measure of their contributions to students. This is an empirical question, and lies at the heart of the current investigation. Are determinations of school effectiveness robust against the choices researchers make about the hidden elements of the definition of school "value-added"?

Purpose / Objective / Research Question / Focus of Study: *Description of the focus of the research.*

This research project examines school performance when assessed according to competing conceptualizations what it means for a high school to "add value." In the paper, I turn to legislative resources to identify state-mandated educational outcomes for which we hold public school accountable. Based on that review, I posit five potential ways in which one could reasonably frame a school value-added causal effect. These are listed together in Appendix B, Figure 1. I do not argue that this is a complete universe of possible definitions—just an exploratory set (Please Insert Appendix B, Figure 1 here). The methods used to estimate these definitions of causal effects are elaborated in the Data Analysis section.

I examine the distributions of performance on these measures across 115 high schools in four large California districts. The objective of the study is to investigate the common notion that there is one universal school value-added measure. I argue that those who design accountability policies should be aware of precisely what causal effect is (and is not) captured in whatever strategy they adopt to translate student test scores into school value-added estimates.

Setting: *Description of the research location.*

I use student-level administrative data provided by four large California school districts. Because the dataset follows individual students longitudinally as they are exposed to public high schools, it provides an opportunity to examine how different *kinds* of students are served by the schools they attend. Since these districts are four of the ten largest in California, I am able to examine patterns across a large number of high schools, lending power to an analysis that focuses on the school as the treatment of interest.

Finally, high schools in California provide a unique opportunity to consider competing definitions of school value-added because, beginning with the Graduating Class of 2006, California requires students to pass an exit exam before they are awarded a high school diploma. The legislation that introduced the California High School Exit Exam (CAHSEE) defined a new goal toward which schools are expected to strive—that is, to prepare their students to initially take or eventually pass the CAHSEE. Schools may vary in terms of their effectiveness in terms of these two goals (initial or eventual passage). In addition, the existence of the CAHSEE provides a concurrent measure of student achievement against which performance on the statewide exam—the California Standards Test (CST)—can be compared.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

I use data from five cohorts of students—students scheduled to graduate in 2006 through 2010—for whom we have several key outcomes of interest including CST performance,

CAHSEE performance, and graduation. Students classified as special education students (roughly 10 percent of students) are excluded from the study, because these students were not subject to the same testing requirements in most of the years covered by our analyses.

Intervention / Program / Practice: *Description of the intervention, program or practice, including details of administration and duration.*

Using the causal framework invoked above, in this paper, the "treatment" of interest will always be one public high school versus some other public high school(s).[6] Implicitly, the effect of a "school" collectively includes the policies, resources, and practices of the school staff and leadership. In all definitions of high school causal effects that I posit below, the treatment will always be conceived of as one "school" relative to other schools. Only the other elements of the definition of a causal effect—i.e., the population, the treatment timeline, the outcome—will be manipulated below.

Research Design: *Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).*

The research design for this analysis is comprised of four basic steps:
(1) I posit and execute a statistical model to estimate an effect of each school according to the five definitions of high school value-added delineated in Appendix B, Figure 1. Each school in the sample thus receives one value-added estimate for each of these five competing definitions of high-school value-added. See the Data Analysis section below for more on this step.
(2) Second, I locate each school within the distribution of public schools in their own district. This will serve, loosely-speaking, to "rank" schools against one another in terms of each separate definition of value-added. I hypothesize that schools may differentially add value according to these five definitions.
(3) In the third step, I look across the various conceptualizations of high school value-added and examine how schools are distributed across them. I quantify the extent to which schools that perform well on one value-added measure perform well on others. This step addresses whether or not a single value-added measure may speak to the myriad educational objectives to which we expect schools to contribute.
(4) Finally, I compare how the five estimation approaches I introduce compare to the current API rankings currently used in California to summarize the quality of schools. This final step highlights the policy implications of making more explicit the definition of school value-added.
In the section that follows, I describe in general terms how the administrative data from these four districts was collected and analyzed in order to estimate five versions of value-added.

Data Collection and Analysis: *Description of the methods for collecting and analyzing data.*

Beginning in spring of 2005, we requested individual student-level administrative records from four large California school districts. (Please insert Appendix B, Figure 3 here, which contains a complete listing of available data). Since 2005, we have continued to receive and integrate new cohorts of data in order to construct a single longitudinal dataset that contains students who entered all public high schools in these four districts in or since SY 2002-03.

In order to estimate school-level value-added estimates, I adopt a hierarchical linear model in which observations of students are nested within schools. I begin by describing the structure of the model for estimating school value-added according to the first definition posited in Appendix B, Figure 1. For each subsequent value-added definition, I make slight changes to that basic model to estimate the desired parameter.

Definition #1: A school adds value when achievement improves during the high school years (test score gains from spring of 8th to spring of 10th grade), regardless of students' skill level just before entering high school.

The model is written out in equation form in Appendix B, Figure 2. (Please Insert Appendix B, Figure 2 here). In this model, the outcome is the 10th grade ELA test score gain of student i in school s. At level-1, the 10th grade outcome is modeled as a function of the student's pre-high school achievement (as measured in spring of 8th grade by ELA test score, group-mean centered). Other pre-treatment covariates are included in the vector **X**. At level-2 (school-level), the intercept and slope from level-1 are allowed to vary randomly across school sites. These two estimated parameters constitute the school value-added estimates: $\widehat{\beta_{0s}}$ is the average 10th grade ELA test score gain *for a student of average 8th grade achievement in each school*. $\widehat{\beta_{1s}}$ captures the strength of the relationship between initial (just before entering high school) skill and 10th grade achievement gain. When $\widehat{\beta_{0s}}$ is relatively large and $\widehat{\beta_{1s}}$ is closer to zero, a school demonstrates both a high level of achievement gains, and that it does so regardless of a student's pre-high school skill level. The HLM model produces a set of estimated values for these two "value-added" parameters, estimates the variability within and between schools, and allows one to test the null hypothesis that schools do not differ in terms of this kind of value-added.

Definition #2: A school is expected to add value by stimulating gains in *both* math and English language arts (ELA) skills. Perhaps some schools are more effective in one of these subject areas than the other.

The model described above is now modified to be a multivariate growth model, in which both math and ELA test scores are concurrently estimated in the HLM framework above (Raudenbush & Bryk, 2002). This allows me to simultaneously estimate schools' effectiveness in these two subject areas and then quantify the extent to which these two kinds of value-added track together. That is, one can estimate the extent to which schools with high value-added in terms of one subject tend to exhibit high value-added in the other subject.

Definition #3: A school is expected to add value to their students' education, regardless of gender, race/ethnicity, socio-economic status, etc. However, schools may add value differentially for sub-populations of students.

In order to estimate whether schools serve different kinds of student at different levels, I add in student-level covariates into the level-1(student level) model and interact them with the pre-high school achievement variable. As a result, each student subgroup (e.g., for race/ethnicity, Black, White, Hispanic, Asian) obtains its own estimated $\widehat{\beta_{1s}}$ from (M1). I conduct post-estimation tests of the null hypothesis that there are no differences in value-added across sub-populations within the same school. This definition entails a change in the definition of the causal effect, in terms of the population of interest.

Definition #4(a): In settings where high school students are required to pass an exit exam, a school adds value by preparing its student take the test the first time (here, spring of 10th grade).

For definitions 4(a) and (b), we shift from focusing on CST achievement scores to high school exit exams, for which schools are mandated to prepare students (Senate Bill 2X, 1999). In order to do so, I use the same structure of the HLM model (M1), however I change the outcome to be an indicator variable which equals one if the student i in school s passes the CAHSEE exit exam on their first attempt, and 0 if he or she fails. When students perform well in the initial administration of the CAHSEE, even when taking into account their initial skill as they entered high school, schools may be considered effective at preparing their students to

meet the proficiency level the state requires for receipt of a diploma. This definition entails a change in the definition of the causal effect, in terms of the "outcome" element.

Definition #4(b): For those students who initially fail the exit exam, a school adds value by subsequently helping students ultimately succeed by spring of 12th grade.

Definition 4(b) builds upon the model used for definition 4(a), but I make two key changes. First, the population of students to whom these value-added estimates apply is limited to *initial CAHSEE failers*. Second, the outcome here becomes the probability of *eventually* succeeding on the CAHSEE before spring of 12th grade (a dichotomous dummy variable). Note that this shift in population and outcome also entails a shift in the timing of treatment; that is, in Definition 4(a) I consider the effect of schools in grade 9 and 10, whereas in Definition 4(b), the model is limited to what occurs between the end of 10th grade and the end of 12th grade. Here, value-added refers to each school's ability to contribute to its struggling students' probability of meeting the requirement and obtaining a high school diploma.

Definition #5: A school adds value to students' educational achievement by stimulating the kind of learning that extends beyond performance on a particular standardized test. That is, observed achievement gains should be substantial enough to register across test metrics.

For this last question, I investigate whether each school's contribution to learning are discernable both on the CAHSEE and the CST. Again, I use a multiple outcomes HLM model to simultaneously estimate both the intercepts and slopes for each school in terms the overlapping test metrics. These alternative sources of information about student achievement provide an opportunity to examine more closely whether gains on one test metric also appear on the other. When schools prove effective at increasing CAHSEE test scores, we would expect to observe corresponding gains on students' CST scores. If, for instance, the kinds of gains students make on the CAHSEE are test-specific—that is, the gains do not translate to the CST—one might worry that those gains are not authentic, or might be due to CAHSEE-specific teaching strategies.

Findings / Results: *Description of the main findings with specific details.*

Complete results are not shown here, for the sake of brevity, however Figure 4 illustrates the kind of analytic comparisons made in the paper. In the left-hand panel of Figure 4, I depict the variability in school-value added according to *Definition 1*. Each line represents one of the schools in the sample, each defined by the estimated intercept ($\widehat{\beta_{0s}}$) and slope ($\widehat{\beta_{1s}}$) from (M1). One can observed notable variability in both slopes and outcomes (a chi-square test indicates that the variability is larger than one would expect by chance). This panel portrays a distribution of schools in terms of their ability to "add-value" to students' CST scores based on their initial skill. The right-hand panel of Figure 4 is of the same form, however the outcome is now 10th grade CAHSEE scores, rather than 10th grade CST scores. Figure 5 juxtaposes the two value-added measures. On the x-axis, I plot the CAHSEE slope coefficient from the right panel of Figure 4, and on the Y-axis, I plot the accompanying slope coefficient from the CST model. There is a clear, positive correlation. However, the two are by no means collinear.

Conclusions: *Description of conclusions, recommendations, and limitations based on findings.*

Preliminary results indicate that the full specification of the school causal effect does change the rankings of schools within these four districts. If one believes that each of these definitions captures an element of the public high school's responsibility, one single measure of school value added may be insufficient for accountability policy purposes.

Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers: Economic Policy Institute.

Cohen, D., Raudenbush, S., & Ball, D. (2003). Resources, instruction, and research. Educational Evaluation and Policy Analysis, 25(2), 119-142.

Harris, D. (2009). Would Accountability Based on Teacher Value Added Be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives. Education Finance and Policy, 4(4), 319-350.

McCaffrey, D. (2003). Evaluating value-added models for teacher accountability: Rand Corp.

McCaffrey, D., Sass, T., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. Education Finance and Policy, 4(4), 572-606.

Raudenbush, S., & Bryk, A. (2002). Hierarchical linear models: Applications and data analysis methods: Sage Publications, Inc.

Reardon (2010). The Counterfactual Model of Causality and the Benefits of Randomization. Causal Inference in Quantitative Educational and Social Science Research, Lecture Notes: Week 1.

Reardon, S., & Raudenbush, S. (2009). Assumptions of value-added models for estimating school effects. Education Finance and Policy, 4(4), 492-519.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. Education Finance and Policy, 4(4), 537-571.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688-701.

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.

Senate Bill 2X, California Senate(1999).

West, S., Biesanz, J., & Pitts, S. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. Handbook of research methods in social and personality psychology, 40–84.

_____
Endnotes

[1] This discussion can pertain to the estimation of teacher or school value added estimation, and the majority of research has focused on the former. In this paper, I focus specifically on the estimation of *public high school* value-added. Arguably, high schools have different objectives than those of their primary counterparts (e.g., graduation, dropout prevention, in addition to contributing to student achievement). This will be important for the analysis at hand.

[2] In its current form, *NCLB* requires schools to make "adequate yearly progress" (AYP) in terms of the percentage of students who demonstrate proficiency on statewide exams. This approach does not take into account that some schools tend to receive students that are initially farther from the state's definition of proficiency. Since traditional accountability formulae do not take these initial differences into account, schools that serve students far from statewide targets are less likely to make AYP and more likely to be sanctioned. In effect, only gains made with students near or above the proficiency targets are rewarded, while gains with the most disadvantaged students may go unacknowledged or ultimately punished.

[3] The term "value-added" refers to a whole body of estimation approaches wherein prior performance is taken into account when examining current performance (McCaffrey, 2003). There is a wide range of such models, both in terms of the underlying strategy and statistical sophistication.

[4] The effect, δ, [on some outcome Y] [for some unit i] [of some treatment condition t relative to some other condition c] is defined as the difference between the value of Y that would be observed if unit i were exposed to treatment t and the value of Y that would be observed if unit i were exposed to treatment (Reardon, 2010; Rubin, 1974; West, Biesanz, & Pitts, 2000).

[5] McCaffrey and others have pointed to this relative shortcoming in the value-added literature: "…Little attention has been paid to the various ways that such a [teacher or school] effect may be defined… The definition of teacher effect involves the specification of a plausible alternative, … as well as an indication of which students are being considered…and what outcome is used to quantify achievement" (McCaffrey, 2003, p. 16).

[6] Some have argued that this treatment may be insufficiently defined to be useful (Cohen, Raudenbush, & Ball, 2003; S. Reardon & Raudenbush, 2009; Rubin, et al., 2004). This critique is very much in line with the spirit of the investigation at hand, as it too draws attention to the underlying definition of school causal effects. However, the current study begins by assuming that it is useful to conceptualize public high schools as a set of competing treatments.

Appendix B. Tables and Figures
Not included in page count.

Appendix B, Figure 1

Possible Dimensions for High School Value-Added:

- *Definition #1*:
  A school adds value when achievement improves during the high school years, regardless of students' level of skill just before entering high school.
- *Definition #2*:
  A school is expected to add value by stimulating gains in *both* math and English language arts (ELA) skills. Perhaps some schools are more effective in one of these subject areas than the other.
- *Definition #3*:
  A school is expected to add value to their students' education regardless of gender, race/ethnicity, socio-economic status, etc. However, schools may add value differentially for sub-populations of students.
- *Definition #4(a)*:
  In settings where high school students are required to pass an exit exam, a school adds value by preparing its student take the test the first time (here, spring of 10th grade).
- *Definition #4(b)*:
  For those students who initially fail the exit exam, a school adds value by helping students ultimately succeed by spring of 12th grade.
- *Definition #5*:
  A school adds value to students' educational achievement by stimulating the kind of learning that extends beyond performance on a particular standardized test. That is, observed achievement gains should be substantial enough to register across test metrics.

Appendix B, Figure 2

Model (M1)

Level 1:
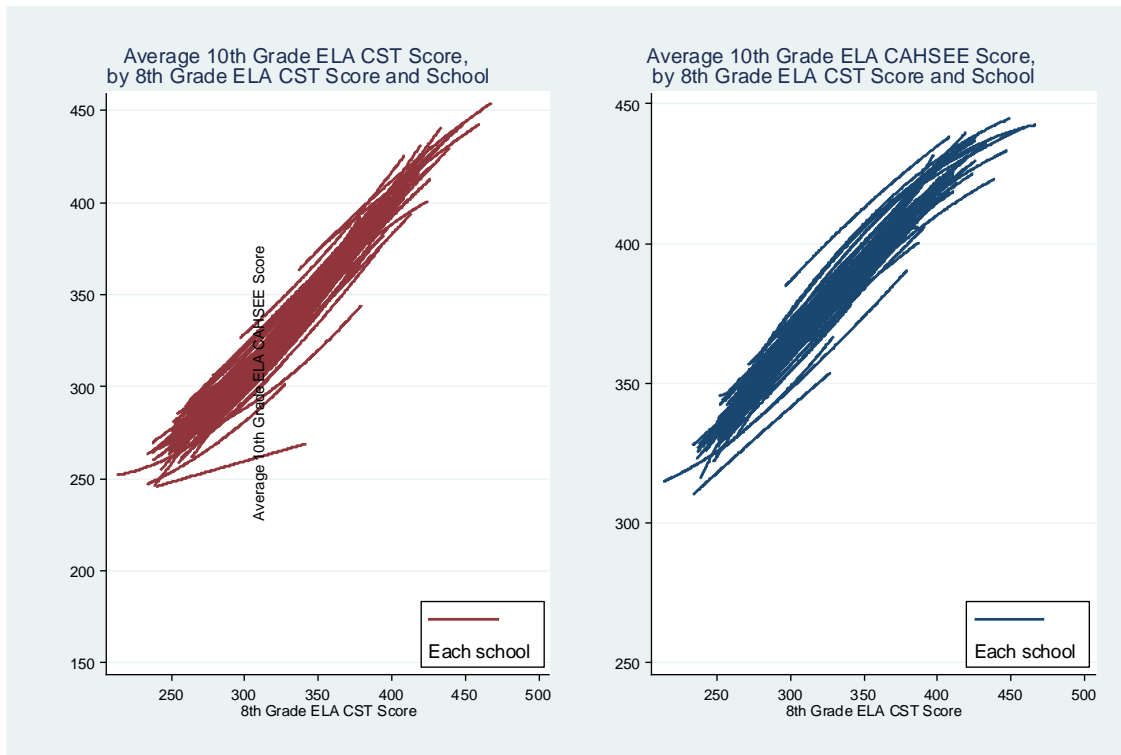$$TESTSCORE\_grade10_{is} = \beta_{0s} + \beta_{1s}(TESTSCORE\_grade8_{is}) + (\boldsymbol{X}_{is})\boldsymbol{\beta_s} + \varepsilon_{is}$$
Level 2:
$$\beta_{0s} = \gamma_{00} + \gamma_{01}(W_s) + u_s$$
$$\beta_{1s} = \gamma_{10} + \gamma_{11}(W_s) + u_s$$

Appendix B, Figure 3

We requested and received, the following data:
(1) a unique id (anonymized) for each student, linked to a unique school identifier;
(2) CAHSEE scale scores for each student for each administration;
(3) CST scale scores & proficiency levels each year, since the first CST administration in 2002;
(4) information on active/ withdrawn/ dropout/ graduation status;
(5) diploma receipt and date, and
(6) basic student demographic characteristics such as age, race/ethnicity, gender, free/reduced lunch status, special education status, and English language status.

Appendix B, Figure 4

Average 10th Grade ELA CST Score, by 8th Grade ELA CST Score and School

Average 10th Grade ELA CAHSEE Score, by 8th Grade ELA CST Score and School

Appendix B, Figure 5

Comparing Betas on 10th Grade ELA Test Scores

CST Standardized Beta
Relationship Between 8th ELA CST and 10th CST Scores