Abstract Title Page
Not included in page count.

Title: Alternative Methods for Estimating Achievement Trends and School Effects: When is Simple Good Enough?

Authors and Affiliations:
Siri Warkentien, RTI International, swarkentien@rti.org
David Silver, RTI International, dsilver@rti.org

Abstract Body

Background / Context:

Public schools with impressive records of serving lower-performing students are often overlooked because their average test scores, even when students are growing quickly, are lower than scores in schools that serve higher-performing students. Schools may appear to be doing poorly either because baseline achievement is not easily accounted for or because changing demographic trends result in successive cohorts of students that are not comparable. These situations are common and problematic for practitioners, policymakers, and researchers who are increasingly tasked with identifying, replicating and communicating effective educational practice. Blunt measures of school effectiveness lead to misguided best-practices studies; misidentification of effective teachers and principals for promotion and retention efforts; and damage to the public image of schools, which in turn can result in low morale, difficulty recruiting staff, and flight. Moreover, poor measures of school effectiveness measures are the most common sources of information for families faced with school choice, leading many to make choices that are not in the best interests of students (Glynn & Waldeck, 2013). Finally, biased measures of school quality may pose broader social threats such as undermining preferences for racial and economic diversity among middle-class families.

Prior research has identified these issues as problematic (for brief overviews, see Boast & Field, 2013; Glazerman & Potamites 2011), pointing out reckless and widespread use of average test scores as measures of school quality, not only in media accounts but also in district and state accountability systems. In the last two decades, value-added approaches that try to isolate the independent contribution of schools to student achievement have exploded, promising a more accurate way to evaluate the effectiveness of schools while accounting for characteristics of the student population being served. Under the right circumstances and assumptions (Amrein-Beardsley 2014; Rubin, Stuart, Zanutto 2004), these methods can provide improved estimates of school effectiveness, but implementing such methods are often unrealistic for school districts and states that have limited time, budgets, and research staff capable of conducting such analyses.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is to (1) explore alternatives to value-added models that are simple to calculate and easy to interpret, and (2) describe the conditions that must be met for such alternatives to provide results that are similar to those found with value-added methods. We focus on three alternative methods that have been the subject of prior reports (see Glazerman & Potamites 2011; Castellano & Ho 2013): average gain scores, calculated as the difference between a school's average score in a given grade in one year and that school's average score in the previous grade the previous year; average cohort differences, calculated as the difference between a school's average score in a given grade in one year and that school's average score in the *same* grade the previous year; and residual gain scores, calculated as the difference between the observed and expected average score in a given grade in one year given the previous year's average score. What differentiates this paper from previous studies is its focus on establishing the robustness of imperfect, practical methods with known limitations, rather than on establishing the supremacy of methods that are not currently practical for widespread use. Our study asks the following questions:

- How similar are results obtained from average gain, cohort difference, and residual gain methods to results obtained using value-added methods?
- Under what conditions can results from average gain, cohort difference, or residual gain scores provide an acceptable substitute for value-added results? We focus on the relationship between student background characteristics and growth and student attrition (or student composition changes) across years.

The first question will be addressed through the use of empirical school district data and the second through simulation studies. The goal of this analysis is to provide practical evidence for districts and states that can inform analyses of achievement growth trends. The methods examined can be used broadly—with school data of interest to practitioners in the state, public/aggregated or statewide longitudinal data—and can quickly produce growth trends that account for the baseline achievement of students by tracking cohorts of students over time.

Population / Participants / Subjects:

The empirical data used to explore our research questions comes from a large urban school district. We use the reading and math scores for students in grades 3 through 8 and estimate school effects for schools with more than 15 students per grade per year. For the value-added estimates, we include measures of student background characteristics, including eligibility for National School Lunch Program (NSLP), disability status, parental education, and race/ethnicity—all found in the school district's administrative data. The alternative methods do not rely on any student background characteristics.

Research Design:

This study presents a combination of empirical investigation using achievement data from a large urban school district and simulation studies. The simulation studies use data generating processes informed by the empirical school district data, making the simulated data as realistic as possible. In particular, simulations use the observed sample sizes and covariates from the empirical data, but with simulated outcome values so that the true school effects are known. We assess the bias, mean square error, and coverage rates (i.e., the proportion of replications where the true parameter values are covered by 95% confidence interval of the parameter estimate.) We run 500 replications for each of the scenarios defined by (1) the strength of relationship between student covariates and academic growth, and (2) levels of student mobility/attrition from one year to the next. Simulations are conducted in Mplus and R.

Data Collection and Analysis:

In this study, we calculate school effects in four ways. The first method is a simple value-added model that includes the student's prior year test score, student background characteristics (National School Lunch Program (NLSP) eligibility, disability status, parental education, race/ethnicity), and an indicator variable for the school attended:

$$Y_{i,j,g} = \beta_1 Y_{i,j,g-1} + \beta_2 X_{i,j,g} + \gamma_{j,g} I_{j,g} + e_{i,j}$$

where $Y_{i,j,g}$ is the achievement score for student *i* in school *j* in grade *g;* $Y_{i,j,g-1}$ is that same student's score in the previous grade; $X_{i,j,g}$ is a vector of student characteristics; and $I_{j,g}$ is an indicator variable for whether the student attended school *j* in grade *g*. The coefficient of interest is the effect of a given school $\gamma_{j,g}$.

The second method is the cohort difference (CD), calculated as the average score of students in grade *g* in school *j* in year *t* minus the average score of students in the same grade *g* in school *j* in the previous year *t-1*.

$$CD_{jg} = \bar{Y}_{j,g,t} - \bar{Y}_{j,g,t-1}$$

The third method is the average gain (AG), calculated as the average score of students in grade *g* in school *j* in year *t* minus the average score of students in grade *g-1* in school *j* in year *t-1*.

$$AG_{jg} = \bar{Y}_{j,g,t} - \bar{Y}_{j,g-1,t-1}$$

The final method is the residual gain, calculated as the residual of the regression model where the average score of school *j* in grade *g* at time *t* is regressed on the average score of school *j* in grade *g-1* at time *t-1*. The residual then indicates whether the school performed better or worse than expected in a given year conditional on student achievement levels the prior year.

$$Y_{i,j,g} = \beta_1 Y_{i,j,g-1} + e_{i,j}$$

Following these calculations, we rank each school in order of their score, such that the school with the largest effect has the rank of 1, second largest has a rank of 2, and so on. For the empirical analysis, following the paper by Glazerman and Potamites 2011, we use the value-added method as our "gold-standard" and compare the ranking of schools under value-added to their ranking according to the other three methods to assess the alignment or divergence of each across five different metrics (see Table 1).

Following the presentation of these metrics, we then test how well these results hold while varying common confounders of school effects, including the strength of relationship between student background characteristics and academic growth and levels of student attrition across years.

Findings / Results:

In the current study, we demonstrate the usefulness of simpler alternatives to value-added methods using empirical data from a large urban school district. In this particular setting, we estimate the effectiveness of schools (as measured by the growth in student achievement on state standardized tests over time) for students in grades 3 through 8 across 465 schools.

In general, we find that cohort difference methods perform the worst, average gains perform relatively well, and residual gains perform the best when assessing how schools rank with these methods compared to their ranks with value-added methods. For instance, figures 1a, 1b, and 1c show the correlations of school ranks for grade 4 mathematics with each of the three methods against their value-added rank. The correlations between value-added ranks are highest for residual gain (0.947), in the middle for average gain (0.798), and lowest for cohort difference (0.446). Conclusions from the remaining five metrics (displayed in Table 2) are substantively similar and consistent with respect to the best-performing alternative to the value-added approach.

Conclusions:
Static rankings of schools and school districts are misleading proxies for school quality, yet such methods continue to be used in state evaluation systems and form the basis of most media accounts of school improvement and school effectiveness. This study shows the prevalence of misinformation that is perpetuated about schools or districts when practitioners, policymakers, and the media rely on badly-flawed methods of tracking student achievement over time. In particular, our analyses demonstrate with empirical data how often families who make decisions based on prevalent, well-intentioned static school rankings (such as those provided by greatschools.org) or cohort difference estimates will make sub-optimal school and housing choices.

As the prevalence and funding for State Longitudinal Data Systems (SLDS) continues to grow, and as the demand for data-informed decision making expands among parents and stakeholders, there in an urgent need for states, districts, and schools to have rapid *and* accurate analyses of student achievement growth over time. Our study speaks directly to how researchers can address the needs of policymakers and practitioners by providing them with useful guidelines as to what simpler alternative methods for calculating school effects or student achievement growth over time are available and when using such alternatives is acceptable. Our results demonstrate that among the alternatives to value-added, cohort differences perform the worst, average gain scores perform relatively well, and residual gain scores perform the best. We provide concrete guidance to stakeholders regarding when gain scores and residual gain scores should be presented to parents and practitioners as reliable estimates, when they should be presented with reservations or warnings, and when they should not be used at all.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education. Critical perspectives on tests and assessment-based accountability.* New York: Routledge.

Boast, L. and Field, T. (2013). Quality School Ratings: Trends in Evaluating School Academic Quality. *National Alliance for Public Charter Schools.* Retrieved from http://www.publiccharters.org/wp-content/uploads/2014/01/Quality-Ratings-Report_20131010T114517.pdf

Castellano, K. & Ho, A. (2013). A Practitioner's Guide to Growth Models. *Council of Chief State School Officers.* Retrieved from http://www.ccsso.org/Resources/Publications/A_Practitioners_Guide_to_Growth_Models.html

Glazerman, S. M., & Potamites, L. (2011). False Performance Gains: A Critique of Successive Cohort Indicators. Working Paper. *Mathematica Policy Research, Inc.* Retrieved from http://eric.ed.gov/?id=ED528389

Rubin, D., Stuart, E., Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

Glynn, T. P., & Waldeck, S. E. (2013). Penalizing Diversity: How School Rankings Mislead the Market. JL & Educ., 42, 417.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Metrics for assessing alignment of school rankings across methods

Metric	Description
(1) Rank correlation	Calculate the correlations of the rankings of the value-added method vs. each of the other three methods
(2) Percent changing rank	Calculate the percentage of schools whose rank order changed more than 5, 10, 50, or 100 places from the value-added rank to the rank according to each of the other three methods
(3) Top or bottom performers	Calculate the number of schools ranked in the top ten (or bottom ten) according to each of the three methods that were actually ranked in the bottom half (or top half) of the distribution according to value-added
(4) Misclassified across median	Calculate the percentage of schools that were misclassified across the median (i.e., schools that were in the top half of the distribution according to value-added but in the bottom half of the distribution according to each of the three methods and vice versa)
(5) Misclassified by more than a decile	Calculate the percentage of schools that were misclassified by more than one decile (i.e., schools that were ranked in a given decile according to value-added but were classified in the decile above or below according to each of the three methods)

Table 2. Performance metrics of school effective for grade 4 mathematics achievement: Cohort Difference, Average Gain, and Residual Gain Method Rankings vs. Value-Added Method Ranking

	Cohort Difference		Average Gain		Residual Gain	
(1) Correlation with Value-Added Rank	0.45		0.79		0.95	
(2) Percentage of Schools With Ranks Different than Value-Added by	n	%	n	%	n	%
> 5 ranks	440	94.6	423	91.0	384	82.6
> 10 ranks	416	89.5	383	82.4	336	72.3
> 25 ranks	370	79.6	313	67.3	217	46.7
> 50 ranks	311	66.9	225	48.4	107	23.0
> 100 ranks	220	47.3	107	23.0	17	3.7
(3.1) Number of schools ranked in top 10 according to Value-Added but below median for	4	0.9	0	0.0	0	0.0
(3.2) Number of schools ranked in bottom 10 according to Value-Added but above median for	0	0.0	0	0.0	0	0.0
(4) Percentage of Schools Ranked in top half by Value-Added and bottom half according to	81	17.4	46	9.9	22	4.7
(5) Percentage of schools With Ranks more than a Decile Different than Value-Added	315	67.7	239	51.4	124	26.7

NOTE: For all metrics, the 465 schools were ranked according to the size of the achievement growth (school effect), from largest to smallest. The ranking on schools used value-added methods are compared to the rankings of schools using each of the three alternate methods.

Figure 1a. Cohort Difference vs. Value-added Rankings

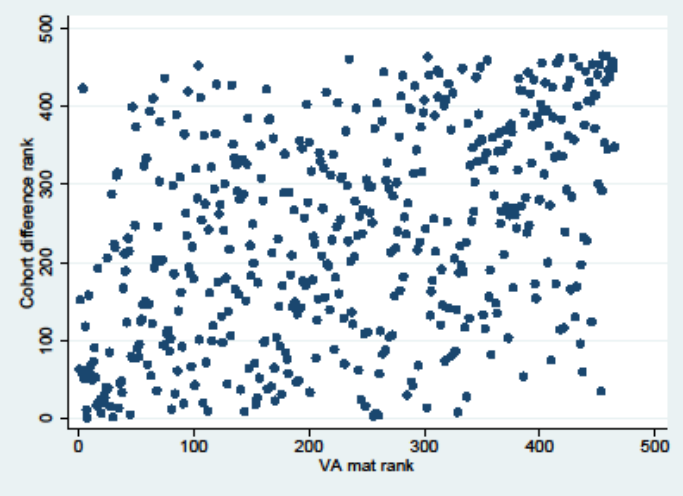

Figure 1b. Average Gain vs. Value-added Rankings

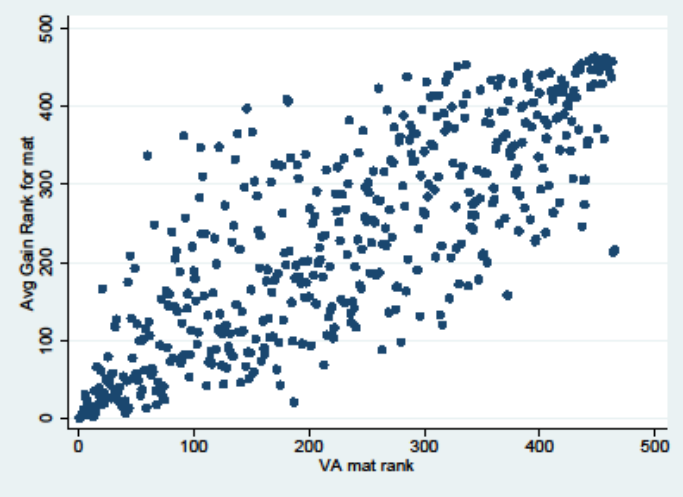

Figure 1c. Residual Gain vs. Value-added Rankings