NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Selecting Growth Measures for School and Teacher Evaluations

MARK EHLERT, CORY KOEDEL,
ERIC PARSONS
AND MICHAEL PODGURSKY

WORKING PAPER 80 • AUGUST 2012

Selecting Growth Measures for School and Teacher Evaluations

Mark Ehlert
University of Missouri

Cory Koedel
University of Missouri

Eric Parsons
University of Missouri

Michael Podgursky
University of Missouri

Contents

Acknowledgements ii
Abstract iii
I. Introduction 4
II. Models 6
1) Student Growth Percentiles (SGPs) 6
2) One-Step VAMs 7
3) Two-step VAMs 9
III. Data 11
IV. Output from the Models 11
V. Model Selection 14
A Tournament Framework 15
Instructional Signals 17
Teacher Labor Markets 21
VI. Other Considerations 21
VII. Conclusion 23
References 24
Tables and Figures 28

Acknowledgements

The authors are in the Department of Economics at the University of Missouri – Columbia. In addition, Podgursky is a Fellow of the George W. Bush Institute at Southern Methodist University.

This research was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) funded through Grant R305A060018 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education. The authors also gratefully acknowledge research support from CALDER and a collaborative relationship with the Missouri Department of Elementary and Secondary Education.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Selecting Growth Measures for School and Teacher Evaluations

Mark Ehlert, Cory Koedel, Eric Parsons and Michael Podgursky

CALDER Working Paper No. 80
August 2012

Abstract

The specifics of how growth models should be constructed and used to evaluate schools and teachers is a topic of lively policy debate in states and school districts nationwide. In this paper we take up the question of model choice and examine three competing approaches. The first approach, reflected in the popular student growth percentiles (SGPs) framework, eschews all controls for student covariates and schooling environments. The second approach, typically associated with value-added models (VAMs), controls for student background characteristics and aims to identify the causal effects of schools and teachers. The third approach, also VAM-based, fully levels the playing field so that the correlation between school- and teacher-level growth measures and student demographics is essentially zero. We argue that the third approach is the most desirable for use in educational evaluation systems. Our case rests on personnel economics, incentive-design theory, and the potential role that growth measures can play in improving instruction in K-12 schools.

I. Introduction

School districts and state education agencies across the country are making increased use of growth-based performance measures in evaluation systems for schools and teachers, sometimes with high stakes attached. Performance metrics that are directly tied to student-achievement gains are appealing because (1) there is a large body of research showing that schools and teachers differ dramatically in terms of their effects on test-score growth (Betts, 1995; Hanushek and Rivkin, 2010), and (2) researchers have had great difficulty linking performance differences between schools and teachers to readily-observable characteristics (Betts 1995; Kane et al., 2008; Nye et al, 2004; Rivkin et al., 2005). The policy focus on growth-based evaluations has stimulated discussions concerning the properties of different statistical models that can be used to produce the growth measures (Goldschmidt et al., 2012).[1]

The question of how to model student test-score growth with the objective of evaluating schools and teachers has resulted in lively debates in states and school districts nationwide. One view, reflected in the popular student growth percentiles (SGPs) approach, eschews all controls for student covariates and other factors related to schooling environments. SGPs are student-level conditional performance percentiles, relative to a peer group. School- and teacher-level SGPs are median values of the student-level SGPs taken at the respective levels of aggregation. The developers of the SGP approach maintain that SGPs are descriptive measures designed to stimulate further investigation or discussion, and do not advocate their use for identifying causal "school effects" (Betebenner, 2011). Regression-based estimates similar in spirit to SGPs can also be constructed, but in practice advocates of models that do not include student or schooling-environment controls have gravitated toward the SGP approach.

[1] There is a large literature that examines the available alternatives for constructing statistical growth models, much of which predates, and/or forms the basis for, recent policy debates (e.g., see Braun, 2005; McCaffrey et al., 2003).

A second approach, usually associated with value-added models (VAMs), controls for student background characteristics and schooling environments with the objective, implicit or explicit, of identifying the causal effects of schools and teachers (Ballou et al., 2004, 2012; Harris, 2011).[2] Several notable studies have questioned whether VAMs can be used to identify causal effects, particularly at the teacher level (Briggs and Domingue, 2011; Rothstein, 2009, 2010). However, recent research suggests that some of the technical concerns of VAM critics may not be of great practical importance - the scope for bias in standard VAM estimates appears to be quite small (e.g., see Chetty et al., 2011).[3]

A third modeling approach focuses on comparing schools and teachers that serve observationally similar students. The rank ordering from an evaluation system based on this approach will not be entirely consistent with a rank ordering based on the absolute "causal" achievement effects for schools and teachers. Rather, it will reflect school and teacher performance relative to other schools and teachers in similar circumstances. This shift in emphasis is motivated by a literature in economics, developed mostly outside of the education context, which supports the idea that ranking systems used for evaluative purposes need not conform to causal effects across unequal classes of competing groups.

We examine the appeal of these three general approaches in the context of an evaluation system for schools, although the substance of our findings will also apply to district- or teacher-level evaluations. We identify three key objectives of any evaluation system in education: (1) elicit optimal effort from agents, (2) improve system-wide instruction by providing accurate performance signals, and (3) avoid exacerbating pre-existing inequities in the labor markets between advantaged and disadvantaged schools. When one considers any of these key objectives, we argue that the third

[2] There is nothing inherent in the SGP approach that prevents it from incorporating additional controls for student or schooling-environment factors. However, in application we are not aware of any implementations of the SGP method where controls beyond same-subject test score histories have been included.

[3] The issue of bias is a conceptually important concern, and researchers and policymakers should remain vigilant. That said, in addition to the Chetty et al. (2011) study, several other recent papers also suggest that the scope for bias in properly-controlled growth models is small. For example, see Goldhaber and Chaplin (2012), Kane and Staiger (2008), Kinsler (forthcoming) and Koedel and Betts (2011).

modeling approach, which has the distinguishing feature of comparing equally-circumstanced schools and teachers, is preferable.

II. Models

Although there are many ways to model student-achievement growth, most models can be categorized into one of three broad classes. The first class of models are what we call "sparse" models – these models purposefully omit available information about students and schooling environments and condition only on prior test score histories. In our study, sparse models are represented by median SGPs taken at the school level. The second class of models comes from the academic literature on the education production function and aims to identify the causal impacts of schools. The representative model for this approach in our study is a one-step fixed effects model, which we estimate as a regression-based VAM. The third class of models is less common in the empirical literature and is motivated with the purpose of building an effective evaluation system. This modeling class is represented in our work by a two-step fixed effects model, which is also based on linear regression. Below we provide additional details about each approach.[4]

1) Student Growth Percentiles (SGPs)

Student Growth Percentiles (Betebenner, 2011) have been adopted for use in evaluation systems in several states. SGPs are calculated using a flexible, non-parametric curve-fitting procedure designed to identify growth percentile curves for student test scores that are analogous to growth charts for children. Imagine a scatter diagram with grade-4 scale scores on the ordinate and grade-3 scores on the abscissa. The SGP procedure fits non-linear quantile regressions for each percentile of the distribution. Thus, for any given third grade scale score, the resulting chart identifies a conditional

[4] The authors produced the VAM measures used in this study. The SGP measures were produced by another group for the Missouri Department of Elementary and Secondary Education (DESE) and were provided for this work.

density function of fourth grade scores. For a student with grade-3 and grade-4 scores, for example, the chart would identify the percentile of the grade-4 score conditional on the student's grade-3 score. Here, an SGP of 67 would indicate that the student's grade-4 score is in the 67th percentile among her peers with the same grade-3 scale score. Aggregated SGPs, when reported, are median percentiles for all of the students assigned to the relevant unit (e.g., district, school or teacher). The number of years of student-level data used to calculate these median SGPs can vary. In the subsequent analysis we use a median based on five years of student level observations.[5]

Although the SGP methodology differs from standard VAM methods in many respects, the relevant distinguishing feature for the purposes of the present study is that SGPs include no controls for student characteristics or schooling environments. Similar VAMs can be constructed. Advocates of sparse growth models such as these value their sparseness; they worry that conditioning on student or school-level characteristics would implicitly "lower the bar" for disadvantaged students.[6] We discuss this issue in more detail below.

2) One-Step VAMs

The one-step VAM is by far the most prevalent modeling structure among research studies designed to estimate school and/or teacher effects (see, e.g., Aaronson, Barrow and Sander, 2007; Goldhaber and Hansen, 2010; Hanushek et al., 2005; Harris and Sass, 2012; Koedel and Betts, 2011; Rockoff, 2004; Rothstein, 2010). Many variants exist. The version that we estimate is shown in Equation

[5] One-year SGP (or VAM) estimates at the school level are highly unstable, which is consistent with findings reported by Kane and Staiger (2002). In spite of this well-known result, the Colorado Department of Education nonetheless reports school-level estimates based on a single year of data. Interestingly, their own web-based visualization tool makes it easy to observe this instability simply by picking a set of schools or districts and changing the testing year. http://www.schoolview.org/ColoradoGrowthModel.asp

[6] Federal guidelines issued in 2009 recommend against using student or school demographic covariates in growth models (U.S. Department of Education, 2009). SGPs, or comparably sparse VAMs, satisfy this recommendation; however, we argue that the recommendation is misguided. We also note that even sparse models, which only condition on students' prior scores, also "lower the bar" for disadvantaged students because lagged achievement is correlated with student demographics.

(1):

$$Y_{isjt} = \beta_0 + Y_{isjt-1}\beta_1 + Y_{iskt-1}\beta_2 + X_{it}\beta_3 + S_{it}\beta_4 + \theta_s + \varepsilon_{ijst} \quad (1)$$

In (1), Y_{isjt} is a test score for student *i* in subject *j* (math or communication arts) in year *t*, X_{it} is a vector of student characteristics for student *i*, S_{it} is a vector of school characteristics for the school attended by student *i* in time *t*, θ_s is a vector of school fixed effects, and ε_{isjt} is the error term.

The model in (1) controls for lagged same-subject and off-subject scores. We show results for math models where communication-arts is the off subject, but our findings are qualitatively similar if we model communication-arts scores instead.[7] The *X*-vector includes information about student race, gender, free/reduced-price lunch eligibility, English-language-learner status, special education status, mobility status (mobile students are defined in the data as within-year building switchers) and grade-level. The *S*-vector includes school-averaged student characteristics for these same variables.

By virtue of the one-step estimation, the coefficient vectors $\hat{\beta}_3$ and $\hat{\beta}_4$ are identified using within-school variation. As a concrete example, the coefficient on free/reduced-price lunch eligibility for individual students is identified by comparing students in the same school who differ in eligibility status. In principle, the model in (1) can identify the causal effects of schools, free from confounding factors, by exploiting within-school variation in the *X*- and *S*-variables. Complications can arise, though. For example, students who differ in terms of their eligibility for free/reduced-price lunch at the same school may not be as different as students who differ in eligibility across schools. Also consider the variables in the *S*-vector, which reflect school-level compositions. In the one-step model the estimates in $\hat{\beta}_4$ are

[7] If a student's lagged off-subject score (communication arts in the mathematics model) is missing, the missing test value is set to zero, and a dummy variable indicating the presence of the missing score is set to one. Moreover, the model also contains an interaction term between the missing test score dummy variable and the student's lagged same subject score. That is, we upweight the predictive value of the same subject lagged score when the off-subject lagged score is not available. This procedure allows us to keep students in the analytic sample if they are only missing the prior score in the off-subject. In cases where students' *same subject* lagged scores are missing, the observations are dropped.

identified by variation in the composition of the student body within schools over time. So, for example, the share of students eligible for free/reduced price lunch may range from 0.80 to 0.85 to 0.78 over a three year period at a particular school. It is this variability that is used to estimate the "effect" of compositional changes on test scores, rather than differences in the shares of students eligible for free/reduced price lunch *across* schools. Within-school identification will work well as long as the effects of compositional changes are linear. Put differently, it must be the case that the effect of a 5-percentage point change in the free/reduced-lunch share is one-tenth the size of the effect of a 50-percentage point change. Changes of the latter magnitude will be commonly observed across schools, but are unlikely to be observed within schools.

We could extend this discussion, but it is not central to our argument. In fact, we will move forward under the assumption that the model in Equation (1) *can* recover the causal effects of schools.[8] That is, we assume that the estimates of θ_s are unbiased estimates of the causal school effects. The implication would be that if we take a random student from school *p* and move her to school *q*, where $\theta_p > \theta_q$, her achievement growth will suffer.

3) Two-step VAMs

Our two-step VAMs are estimated as follows:

$$Y_{isjt} = \gamma_0 + Y_{isjt-1}\gamma_1 + Y_{iskt-1}\gamma_2 + X_{it}\gamma_3 + S_{it}\gamma_4 + \eta_{isjt} \quad (2)$$

$$\hat{\eta}_{isjt} = \delta_s + u_{isjt} \quad (3)$$

The variables in equation (2) are defined as in equation (1). There are two noteworthy differences between the one-step and two-step fixed effects models. First, estimating the model in two steps allows us to control for lagged school-average prior test scores, which is a substantively important control for

[8] Again, recent evidence in the teacher-quality literature suggests that this is a reasonable assumption. For example, see Chetty et al. (2011), Goldhaber and Chaplin (2012), Kane and Staiger (2008), Kinsler (forthcoming), Koedel and Betts (2011); with Rothstein (2010) being the exception. Ballou et al. (2012) refer to a variant of the one-step model as the "true" model.

the schooling environment. In the formulation above, we incorporate the aggregate test-score controls into the school-level control vector S_{it}.[9] Second, the two-step approach partials out differences in test-score performance between students with different characteristics, and in different schooling environments, *before* estimating the school effects. In this way, the two-step model is conceptually similar to a model where the school effects are specified as random instead of fixed.

By partialing out the predictive effects of student and school characteristics prior to estimating the school-level growth measures, the two-step model attributes all differences between students along the measured dimensions to those characteristics. This allows for divergence between the estimates of δ_s from equation (3) and θ_s from equation (1). For example, suppose that all free/reduced-lunch eligible students attend schools that are truly inferior in quality, on average, to the schools attended by ineligible students. The average gap in school quality between these groups in the two-step model would be fully absorbed in the first step. In equation (1), where the school effects and poverty effects are estimated simultaneously, the school-quality difference would not be absorbed by the poverty control because the poverty effect would only be identified from within-school variation. As mentioned previously, if we assume ideal estimation conditions for the one-step model, the one-step estimates can be interpreted as estimates of the causal achievement-growth effects; the two-step estimates do not carry the same interpretation.

[9] There is a mechanical negative correlation between prior school-averaged achievement and year-*t* student outcomes in the one-step model if lagged school-level average test scores are included along with the school fixed effects and the time horizon is short (which is typically the case in these models). The two-step model circumvents this problem by attributing differences in student achievement to lagged aggregate test scores and the current-year school effect sequentially, rather than simultaneously. It does not *solve* the problem. That terminology would be too strong; it simply assigns attribution of the effects sequentially, leaving the current-year school effects only to explain the residual variance after student scores have already been adjusted for the lagged average achievement in schools.

III. Data

The data available for our study are from the Missouri Assessment Program (MAP) test results, linked longitudinally using a statewide student identifier. Like many other state education agencies, the Missouri Department of Elementary and Secondary Education (DESE) has been grappling with ways to incorporate growth measures into its statewide evaluation system. The growth measures compared in this paper were developed in part as a result of those efforts. The administrative data panel contains nearly 1.6 million test-score growth records for students (where a growth record consists of a current and prior score) covering the 5-year time span from 2007 to 2011 (2006 scores are used as lagged scores for the 2007 cohort). We evaluate 1846 schools serving students in grades 4-8 in Missouri. Descriptive statistics regarding the administrative dataset can be found in Appendix A.

IV. Output from the Models

As a point of entry into our discussions, Figure 1 plots school-averaged test scores, in levels, against the share of students eligible for free/reduced-price lunch. The clear negative relationship between student poverty and test score levels, combined with the uncontroversial role that non-schooling factors play in determining student success, has contributed greatly to the migration toward growth measures in education. Is it the case that nearly every high-poverty school in Missouri performs poorly, as is implied by Figure 1, or are at least some of these schools actually performing well only to have their performance masked by their general disadvantage?

Growth modeling has gained considerable traction among researchers and educational policymakers as a way to improve inference for performance evaluations in education. It appeals to the intuitive notion that a system of rewards and sanctions built around the rankings in Figure 1 would wrongly attribute the influence of factors outside of the control of schools to the "performance" measures. That is, some of the highest-placing schools on the vertical axis may not be performing

particularly well; and alternatively, some of the lowest-placing schools may be performing quite well when one considers the context in which they are operating.

By way of comparison, Figure 2 displays growth metrics for the same schools as in Figure 1 using the three different approaches discussed above.[10] Again, we order schools along the horizontal axis by their share of students eligible for free/reduced-price lunch. The first panel shows schools' median student growth percentiles (SGPs). As noted previously, the SGP framework is a commonly-adopted version of what we refer to as a "sparse" growth model. In the typically-estimated SGP framework, the model conditions on as many prior test scores as are available for the student, in the same subject, to construct a comparison "peer group." Students with as few as one prior test score are included. The SGP plot shows that a substantial portion of the negative relationship between test score levels and student disadvantage disappears when prior test scores are accounted for – that is, when we move from a levels-based to a growth-based evaluation. Nonetheless, a clear negative relationship between growth and student poverty remains.

The next panel shows analogous output from the one-step fixed effect VAM. The scale on the vertical axis changes as we move from estimates measured as percentiles to estimates measured in standard-deviation units (as in the VAM), but the negative relationship remains. Assuming that the conditions required for causal identification are satisfied in the one-step model, the implication is that all else equal, high-poverty schools produce less growth in achievement relative to their low-poverty counterparts. This would not be far-fetched given the labor-market dynamics in education, among other factors.[11]

[10] None of the estimates in Figures 2 or 3 are shrunken. The reason is that it is not clear how one would shrink the SGP estimates, and for illustrative purposes we want to maintain as much comparability as possible across models. Our argument does not depend substantively on whether the VAM estimates are shrunken or not.

[11] For example, several recent studies demonstrate teachers' preferences to work in more-advantaged schools (one reason is that these schools are closer to where teachers live – see Boyd et al., 2005; Reininger, 2012). Jacob (2007) discusses general recruiting difficulties faced by urban districts, which are often disadvantaged. The education sector is generally devoid of compensating wage differentials to offset the recruiting difficulties faced by disadvantaged schools.

The last panel in Figure 2 plots the school-level growth estimates from the two-step VAM. By construction, the two-step model breaks the correlation between achievement and the poverty measure, resulting in the flat-lined picture shown in the figure. That is, high- and low-poverty schools are roughly evenly represented throughout the school rankings based on the two-step model. The even representation comes from the fact that differences in schooling environments and school characteristics, including poverty share, are partialed out before the school-level growth measures are estimated.[12] A notable feature of the flat-lined picture is that there is still considerable variability in the estimates within any vertical slice in the graph. That is, even when schools are compared to other similar-looking schools, large differences in annual test-score growth are visible.[13]

Overall, the scatter plots in Figure 2 show that there are important differences in the results from the three classes of models that we consider. This fact is somewhat masked by the raw correlations between the estimates from the different models, which we present in Table 1. The correlations in the table may seem high; however, the differences in output across the models corresponding to these correlations are nontrivial because the types of schools that do well (or poorly) in each model differ in systematic ways.

Tables 2 and 3 provide more details about the differences in results across models. First, Table 2 contrasts the proportion of disadvantaged schools in the analytic sample with the proportion in the top quartile of the rankings from each growth model. Disadvantaged schools are defined as those with at

[12] The representation is not exactly even because of weighting. For example, if we assign a school growth measure to each student, then correlate the school-growth measures and school characteristics using the student weights, the correlations are zero by construction. However, when we correlate the school-aggregated measures the weighting deviates from the student weights (there is now one observation per school, rather than per student), which in our application results in small non-zero correlations between the school-level growth measures and aggregated demographics.

[13] The graphs in Figure 2 are also consistent with recent evidence from Sass et al. (2012), who analyze teacher quality at high and low poverty schools. They find that there is more variation in teacher effectiveness at high-poverty schools. Our school-level growth measures show a similar pattern of increasing variability at higher levels of student poverty.

least 80 percent of students eligible for free or reduced-price lunch.[14] Consistent with the visual representation in Figure 2, clear differences emerge. Using median SGPs, disadvantaged schools are meaningfully underrepresented in the top quartile. Alternatively, the two-step model produces rankings where disadvantaged schools are slightly overrepresented. The one-step model is an in-between case where high-poverty schools are somewhat underrepresented in the top quartile of performance, but less so than in the SGP rankings.[15]

Next, in Table 3, for each model we identify all top-quartile schools that are not identified as top-quartile schools in the other models. These "non-overlapping winners" provide an alternative illustration of the differences in output. Again, consistent with what can be seen from Figure 2, the first two models are much more likely to identify advantaged schools as being in the top quartile relative to the two-step model. For example, top-quartile schools as identified by SGPs that are not identified as top-quartile schools by the two-step model have, on average, 32.8 percent of their students eligible for free/reduced-price lunch. Conversely, 69.7 percent of students are eligible for the lunch program at top-quartile schools as identified by the two-step model but not by SGPs.

The substantive differences illustrated in this section naturally lead to the question of which model should be used for evaluating school performance. It is this question to which we now turn our attention.

V. Model Selection

As a quick summary of the statistical characteristics presented above, the SGP output is descriptive and designed to foster discussion. In contrast, the one-step fixed effects model is the only model where direct causal inference is potentially supported. Some assumptions are required for causal

[14] The findings reported in Table 2 are not qualitatively sensitive to how disadvantaged schools are defined.
[15] Goldhaber et al. (2012) report qualitatively similar findings in a series of comparisons between SGPs and one-step VAMs. Their analysis focuses on growth measures for individual teachers. They do not evaluate two-step specifications (or substantively similar models).

identification but recent evidence, most-notably from Chetty et al. (2011), suggests that these assumptions are likely to hold (at least roughly). Finally, the two-step model almost surely "over-corrects" relative to the actual causal effects of schools. Any student or school characteristics that are systematically correlated with schooling effectiveness will absorb the school-effect differences in the first step.

We argue that despite the potential for "overcorrection" in the two-step model, it is still preferable when the objectives of the evaluation system include the following: (1) eliciting optimal effort from agents (i.e., teachers and administrators), (2) sending signals to schools that will improve instruction, and (3) avoiding exacerbating inequities in the labor markets faced by advantaged and disadvantaged schools.

A Tournament Framework

A large and growing literature in personnel economics focuses on the importance of sending the right performance signals to employees (or more generally, "agents"). An important paper that links this literature to K-12 education is Barlevy and Neal (2012).[16] These authors focus on the efficient design of incentive pay for teachers. One finding from their study is that systems based on percentile rankings, which are ordinal, are in many contexts preferred to systems that incorporate cardinal information, such as those discussed in the VAMs above. The primary advantage of the ordinal percentile measures is that they do not depend on the scaling of the exam. This reduces the need to worry about vertical alignment and, according to the authors, reduces the incentive to "corrupt" the testing measures by teaching to particular forms of tests, for example. Indeed, Barlevy and Neal note favorably the attractive features of the SGP approach in this regard.[17]

[16] A seminal paper in this literature is Lazear and Rosen (1981). A widely cited, but somewhat dated survey is Prendergast (1999).

[17] The argument is that the move to ordinal performance measures will allow test makers to become less predictable by freeing them from attempting to align scores vertically across tests. Although not common practice currently, it would be straightforward to estimate VAMs that are designed for ordinal comparisons.

However, a key finding in the larger personnel economics literature, noted by Barlevy and Neal, is that it is of great importance to set up the right comparison groups for the evaluation.[18] The intuitive argument is that if competitors are placed in competition with players against whom they have no hope of winning, incentives will weaken for everyone. Experimental evidence on tournaments supports this thesis. For example, Schotter and Weigelt (1992) draw on the tournament literature to examine the incentive effects of affirmative action programs. They employ games designed to mimic tournaments that "level the playing field" and deter disadvantaged agents from dropping out. Done properly, these types of "asymmetric tournaments," as they are called in the literature, have the effect of raising the effort level of all agents, including those in advantaged groups.

A central lesson from Barlevy and Neal and related studies in this literature is that the right signal must be sent to agents in different circumstances. This signal need not be a direct measure of absolute productivity; instead, it should be an indicator of performance relative to equally-circumstanced peers. By leveling the playing field, the two-stage model discussed above achieves this objective. In contrast, the SGP approach or one-step VAM would not create a balanced league table and, in fact, would favor the advantaged group, which runs counter to the goal of eliciting optimal effort.[19]

From an incentive-design perspective, then, a long-standing prior research literature has established that comparisons like those produced by the two-step model are preferable.[20] Still, skeptics may point to several recent studies questioning the margin for meaningful improvement in K-12 schools in the United States from eliciting additional effort from educators. For example, Springer et al. (2010)

[18] Barlevy and Neal (2012) routinely refer to equally-circumstanced peers as peers with *similar prior achievement*, but this is somewhat misleading. In several places, they elaborate on what they actually mean by this terminology, which is that peers should be in similar circumstances in general, not just in terms of prior test scores.

[19] Most states using SGPs as part of their accountability systems attempt to deal with this issue by creating league tables *ex post*, i.e., only comparing similar looking schools to one another after the fact. Although such a system is not an unreasonable compromise given the nature of the growth measure chosen, the field leveling is likely to be more statistically accurate and comprehensive if done *ex ante* (such as in the two-step model), as well as being less susceptible to corruption from political pressure.

[20] The separate issue raised by Barlevy and Neal (2012) of whether evaluations should be based on ordinal or cardinal rankings can be addressed within any modeling framework.

conduct an incentive experiment for teachers in Tennessee and find no discernible effort effects. One hypothesis put forth by the authors, consistent with teacher responses to surveys, is that teachers were already supplying considerable effort prior to enrolling in the incentive program. The Springer et al. (2010) findings are corroborated by a recent study by The New Teacher Project (2012), which shows no difference in weekly work hours between high-performing and low-performing teachers.[21] These results might fairly be interpreted to suggest that the margin for improving K-12 instruction from eliciting additional educator effort is limited. That said, recent evidence from Taylor and Tyler (2011) shows that high-quality teacher evaluation systems can lead to improved performance. The potential for instructional improvements to arise from an evaluation system will depend on proper signals being sent to the relevant actors. It is to this subject that we now turn.

Instructional Signals

Growth models can be used to improve instruction if they provide accurate performance signals. A positive performance signal, for example, might encourage a school to continue to pursue and augment existing instructional practices. Alternatively, a negative signal can provide a point of departure for instructional change and/or intervention. Furthermore, informative signals throughout the system can be used to improve system-wide instruction. For example, an underperforming school may benefit from observing a school that is performing much better, but this benefit will only be attainable if the system provides useful information to direct educator-to-educator learning (i.e., a system that tells educators who should be learning from whom). The signaling value of an evaluation system is particularly important when it is difficult for individual schools to assess their performance, and the performance of others, accurately.

[21] This evidence from the United States is at odds with evidence from several studies in the developing world, notably Duflo, Dupas and Kremer (2012) and Muralidharan and Sundararaman (2011).

In terms of providing effective performance signals, the two-step model is again preferable. This is true even under the maintained assumption that the one-step model produces "causal" estimates. How can a model that produces estimates that deviate (potentially) from the causal estimates be preferred? The answer is that "causality" is too narrowly defined in the one-step model. That is, even if advantaged schools really do produce more test-score growth, on average, when compared to disadvantaged schools, the outcomes at advantaged schools do not properly reflect the counterfactual outcomes for their disadvantaged counterparts. One obvious reason is that the degree to which a school's students are disadvantaged affects educator labor supply (Boyd et al., 2005; Jacob, 2007; Reininger, 2012).

To illustrate, suppose for simplicity that the school effects from the one-step model entirely reflect the quality of teachers and principals. The differences in rankings in this simple world, then, reflect differences across schools in personnel quality. Continuing to assume that the one-step model is properly ranking schools in terms of growth produced, this would imply that advantaged schools have higher-quality teachers and administrators. There are two primary factors that determine the quality of the personnel in a given school: (1) the quality of the applicant pool and (2) conditional on the quality of the applicant pool, the success of the school in selecting the best applicants. Based on observational measures, we know that disadvantaged schools have access to lower-quality applicants (Boyd et al., 2005; Jacob, 2007; Koedel et al., 2011; Reininger, 2012). There are many explanations for this, most of which are outside of the control of the schools themselves. Hence, if there are large disparities in applicant-pool quality between advantaged and disadvantaged schools, and disadvantaged schools do not have any levers to pull to meaningfully improve the quality of their applicant pools, an evaluation system that provides "performance signals" based in part on this feature of the labor market will be of little value for improving instruction.

The above labor-market example is useful for its simplicity, but more generally, advantaged and disadvantaged schools are likely to differ along many dimensions. Designing an evaluation system that sends performance signals to schools that can be predicted ex ante with easily observable measures of student disadvantage ignores all of the dimensions by which different types of schools are segmented. Figure 3 provides a concrete example of the types of problems that can arise from an evaluation system that does not maintain proper seeding. In the figure, we take one high-poverty school and one low-poverty school and highlight the placement of each school in each plot from Figure 2. To protect the anonymity of the actual schools we call the high-poverty school "Rough Diamond" and the low-poverty school "Gold Leaf." Beginning with Rough Diamond, if we take a vertical slice in the area of Rough Diamond in any of the pictures in Figure 3, it is clear that Rough Diamond is performing well compared to similar schools. Few schools that look anything like Rough Diamond in terms of student poverty do meaningfully better, and many do worse. A concrete question that should be at the forefront in the design of the evaluation system is this: What signal should be sent to Rough Diamond?

The SGP and one-step fixed effects models send similar performance signals to Rough Diamond, both negative. For example, the median SGP for Rough Diamond, coupled with its status in test score levels (not shown), would put it in the "needs improvement" quadrant in the standard SGP bubble chart. Similarly, in the one-step fixed effects model, Rough Diamond would get a growth rating that is below average. Given the signals from either of these models, Rough Diamond might be expected to respond in one of two ways. At best, it would dismiss the model output. At worst, it would be prompted to make wholesale changes to the delivery of instruction in response to a negative performance rating from the system. Whatever Rough Diamond is doing seems to be working quite well in the environment in which it is operating; put differently, if Rough Diamond were to re-tool and start over, in expectation it would do worse.

In contrast, the two-step model sends a positive signal to Rough Diamond. We argue that this is the right signal, in the sense that it should inspire Rough Diamond to continue to pursue and refine its current instructional and personnel strategies, which have placed it well above average relative to peer schools.

Hence, by virtue of choosing a growth model, the signal that is sent to Rough Diamond (or any other school in similar circumstances) is largely at the discretion of policymakers. The choice of model determines whether Rough Diamond receives a signal from the evaluation system that reinforces current practices, or whether the system indicates to Rough Diamond that its performance is unsatisfactory. Similarly, it determines whether principals from low-performing schools in similar contexts will be encouraged to look at Rough Diamond as an example for how they might improve instruction at their own schools.

For Gold Leaf, on the other hand, the opposite story holds. In the SGP and one-step models, Gold Leaf is identified as a school with above-average growth. As a result, other schools would be encouraged to look to Gold Leaf as an example school given the output from these models. However, as can be seen clearly in Figure 3, Gold Leaf is among the lowest performing advantaged schools in the state. A plausible scenario is that Gold Leaf is doing a poor job hiring effective educators conditional on the quality of its applicant pool, but continues to perform better than average because its pool is so strong.[22] The fact that Gold Leaf outperforms Rough Diamond need not be informative at all about the context-specific performance of either school. Put differently, would it make sense to bus teachers and administrators from Rough Diamond to Gold Leaf so that they can observe a "high performing" school? Is Gold Leaf really an appropriate model school for Rough Diamond to emulate?

The challenges facing disadvantaged schools include the direct difficulties associated with teaching students who receive lower quality non-schooling inputs, and also the indirect challenges

[22] This, of course, is one of many possibilities. Also note that differences in applicant-pool quality are also likely to be relevant in school-leader labor markets (Koedel et al., 2011).

related to educator labor markets, funding discrepancies, etc. that come with being in a disadvantaged area. It is difficult to understand how a system that ignores these issues and attempts to signal to all (or nearly all) disadvantaged schools that they must perform better will help improve instruction. Alternatively, a system that differentiates schools conditional on disadvantage can highlight the large performance differences among observationally similar schools across all types of schools. These differences clearly exist and are illustrated by the large variation within any vertical slice in any of the plots in Figures 2 and 3. There is the potential for much learning to occur across observationally similar schools and for subsequent improvements in overall instruction, but only if the output from the evaluation system provides proper signals with regard to which schools are performing well and which schools are performing poorly, conditional on the real-world contexts in which they operate.

Teacher Labor Markets

As noted above, it is well-established that schools in poor areas are at a competitive disadvantage in the labor market. As stakes become attached to school rankings based on growth models, systems that disproportionately identify poor schools as "losers" will make positions at these schools even less desirable to prospective educators. Policymakers should proceed cautiously with implementing an evaluation system that will further degrade the pecuniary and non-pecuniary benefits associated with working in challenging educational environments. An important benefit of the two-step model is that the "winners" and "losers" from the evaluation will be broadly representative of the system as a whole (see Table 2).

VI. Other Considerations

One concern with the two-step model, or other models that level the playing field across schools, is that it will "hide" inferior performance at disadvantaged schools. That is, in the sparse model one can clearly see that lower growth, on average, is achieved at low-income schools. One might worry

that moving to a model similar to the two-step model would mask this feature of the data and give the impression that school performance in disadvantaged areas is better than it actually is.

Our view is that this concern is largely misguided. A model along the lines of the two-step VAM can be adopted in conjunction with reporting on test scores levels, and in fact, state- and district-level evaluation systems that incorporate test-score growth also typically have a test-score-levels component. The reporting on test-score levels will allow state administrators and policymakers to clearly see absolute differences in achievement across schools, regardless of which growth model is adopted. It is also important to recognize that the information from the growth model need not crowd out other useful information. For example, schools could receive a growth-model report indicating their performance relative to similarly-circumstanced schools that also includes information about absolute achievement levels. Again, this type of dual reporting is desirable because it allows for the transmission of useful instructional signals. For example, a poor school that is performing well, like Rough Diamond, can be encouraged to continue to refine and improve an already-effective instructional strategy (in terms of raising test scores compared to similar schools) but still be reminded that their students are not scoring sufficiently high relative to an absolute benchmark. The latter information need not disappear in any evaluation framework.

We also note that there are many potential variants of the two-step model that we estimate here. A notable alternative would convert all scores to percentiles as suggested by Barlevy and Neal (2012) and Betebenner (2011), which would facilitate ordinal rather than cardinal comparisons between students. This would be particularly desirable in circumstances where the scaling properties of the exam are suspect, which prior research suggests is a common problem (Ballou, 2009). If scores were converted to percentiles, then the output from the first step of the model would be interpretable as conditional performance percentiles for students. The lagged-score controls could also be reconstructed

and/or expanded in various ways. For example, rather than using linear predictors as in Equation (2), one could construct indicator variables for different lagged-score values or bins.[23]

VII. Conclusion

We examine three general approaches to modeling student test score growth – Student Growth Percentiles (SGPs), a one-step VAM model, and a two-step VAM model. Broadly speaking, these models represent the choice set for policymakers in their efforts to design evaluation systems for schools and/or teachers. Although SGPs are currently employed for this purpose by several states, we argue that they (a) cannot be used for causal inference (nor were they designed to be used as such) and (b) are the least successful of the three models in leveling the playing field across schools. Neither of these results is surprising given the sparse nature of the measure. One-step VAM estimates are the most likely of the three approaches to produce causal estimates and are the most strongly-represented in the research literature. However, the two-step VAM approach, and general variants, will be most desirable for use in educational evaluation systems even if the output is at odds with schools' absolute causal effects on achievement. The key feature of the two-step model is that it levels the playing field across schools so that "winners" and "losers" are representative of the system as a whole. When one considers the key objectives of an evaluation system, this feature of the two-step model is desirable along all fronts.

[23] The indicator-variable approach could be used to produce estimates very similar to what Betebenner's SGP software produces, with a benefit being that it would be straightforward to compute standard errors and confidence intervals.

References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.

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

Ballou, Dale, Christine G. Mokher and Linda Cavalluzzo. 2012. Using Value-Added Assessment for Personnel Decisions: How Omitted Variables and Model Specification Influence Teachers' Outcomes. Unpublished manuscript.

Ballou, Dale, William Sanders and Paul Wright. 2004. Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics* 29(1), 37-65.

Barlevy, Gary and Derek Neal. 2012. Pay for Percentile. *American Economic Review* 102(5), 1805-31.

Betebenner, Damien W. 2011. A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories.

Betts, Julian. 1995. Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth. *Review of Economics and Statistics* 77(2), 231-250.

Boyd et al. 2005. The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. *Journal of Policy Analysis and Management* 24(1), 113-132.

Braun, Henry. 2005. Using Student Progress to Evaluate Teaching: A Primer on Value-added Models. Princeton, NJ: Policy Information Center, Educational Testing Service.

Briggs, Derek and Ben Domingue. 2011. Due Diligence and the Evaluation of Teachers. National Education Policy Center Report.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2011. The Long-Term Impacts of Teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper No. 17699.

Duflo, Esther, Pascaline Dupas and Michael Kremer. 2012. School governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools. NBER Working Paper No. 17939.

Goldhaber, Dan and Duncan Chaplin. 2012. Assessing the "Rothstein Falsification Test". Does it Really Show Teacher Value-Added Models are Biased? CEDR Working Paper.

Goldhaber, Dan and Michael Hansen. 2010. Assessing the Potential of Using Value-Added

Estimates of Teacher Job Performance for Making Tenure Decisions. CALDER Working Paper No. 31.

Goldhaber, Dan, Joe Walch and Brian Gabele. 2012. Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. CEDR Working Paper.

Goldschmidt, Pete, Kilchan Choi, and J. P. Beaudoin. (2012) A Comparison of Growth Models: Summary of Results. Washington DC: CCSSO

Hanushek, Eric A. 2011. Valuing Teachers. *Education Next* 11 (3), 41-45.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien and Steven G. Rivkin. 2005. The Market for Teacher Quality. NBER Working Paper No. 11154.

Hanushek, Eric A., John F. Kain and Steven G. Rivkin. 2004. Why Public Schools Lose Teachers. *Journal of Human Resources* 39(2), 326-354.

Hanushek, Eric A and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100(2), 267-271.

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

Harris, Douglas N. and Tim R. Sass. 2012. Skills, Productivity and the Evaluation of Teacher Performance. Unpublished manuscript.

Jacob, Brian. 2007. The Challenges of Staffing Urban Schools with Effective Teachers. *Future of Children* 17(1), 129-153.

Kane, Tom J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. What Does Certification Tell us about Teacher Effectiveness? Evidence from New York City. *Economics of Education Review* 27(6), 615-631.

Kane, Tom J. and Douglas O. Staiger. 2002. The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives* 4(1), 91-114.

--. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, NBER Working Paper No. 14607.

Kinsler, Joshua. Forthcoming. Assessing Rothstein's Critique of Teacher Value-Added Models. *Quantitative Economics*.

Koedel, Cory and Julian R. Betts (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy* 6(1): 18-42.

Koedel, Cory, Jason A. Grissom, Shawn Ni and Michael Podgursky. 2011. Pension-Induced Rigidities in the Labor Market for School Leaders. CALDER Working Paper No. 62.

Lazear, E. P. and S. Rosen. 1981. Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy* 89(5), 841-864.

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

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy* 119(1), 39-77.

Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. 2004. How Large are Teacher Effects? *Educational Evaluation and Policy Analysis* 26(3), 237-257.

Prendergast, Candice. 1999. The Provision of Incentives in Firms. *Journal of Economic Literature*. 37(1), 7-63.

Reininger, Michelle. 2012. Hometown Disadvantage? It Depends on Where You're From: Teachers' location preferences and the implications for staffing schools. *Educational Evaluation and Policy Analysis* 34(2), 127-145.

Rivkin, Steven G., Eric A. Hanushek and John F. Kain. 2005. Teachers, Schools and Academic Achievement. *Econometrica* 73(2), 417-58.

Rockoff, Jonah. 2004. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review (P&P)* 94(2), 247-252.

Rothstein, Jesse. 2009. Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy* 4(4), 537-571.

--. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* 125(1), 175-214.

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio and Li Feng. 2012. Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools. *Journal of Urban Economics* 72(2-3), 104-122.

Schotter, Andrew and Keith Weigelt. 1992. Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results. *Quarterly Journal of Economics* 107 (2), 511-539.

Springer, Matthew, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper and Brian M. Stecher. 2010. Teacher Pay for Performance:

Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives Report.

Taylor, Eric S. and John H. Tyler. 2011. The Effect of Evaluation on Performance: Evidence from student achievement data of mid-career teachers. NBER Working Paper No. 16877.

The New Teacher Project. 2012. The Irreplaceables: Understanding the Real Retention Crisis in America's Urban Schools. Policy Report.

U.S. Department of Education. 2009. Growth Models: Non-Regulatory Guidance. (January 12). *www2.ed.gov/admins/lead/lead/growthmodel/0109gmguidance.doc*

Tables and Figures

Figure 1. School-Average Test Scores Plotted Against School Shares Eligible for Free/Reduced-Price Lunch

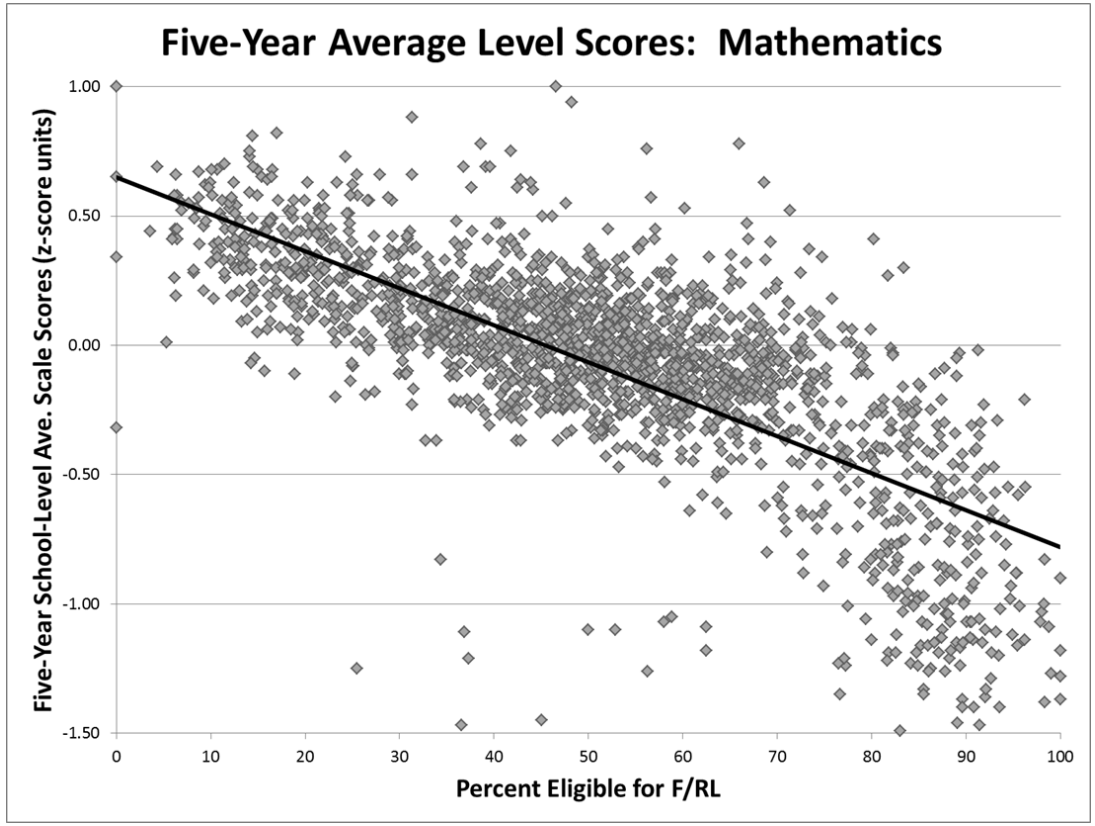

Figure 2. School Growth Measures from Each Model Plotted Against School Shares Eligible for Free/Reduced-Price Lunch.

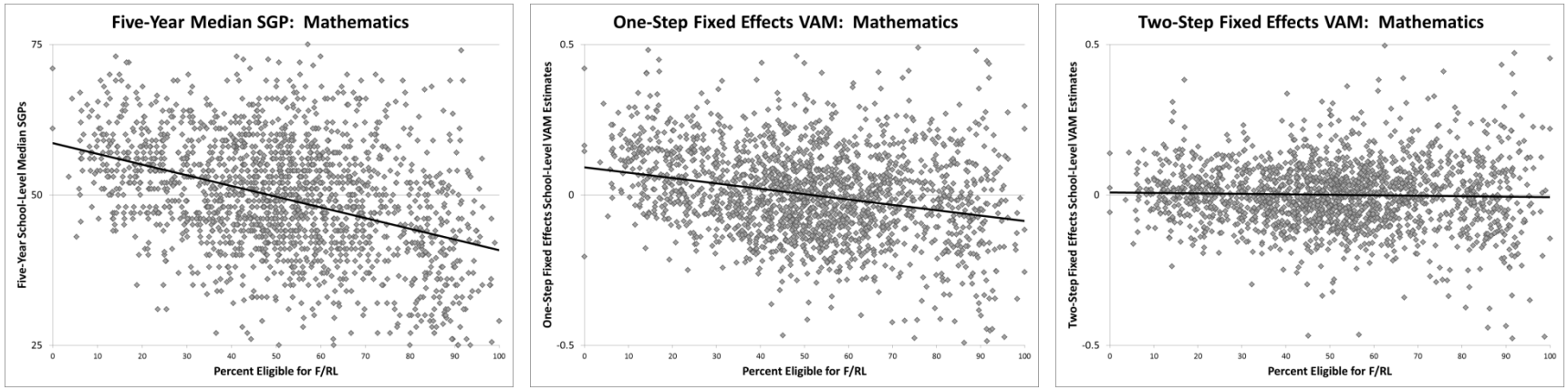

Figure 3. School Growth Measures from Each Model Plotted Against School Shares Eligible for Free/Reduced-Price Lunch, with Highlighted Example Schools.

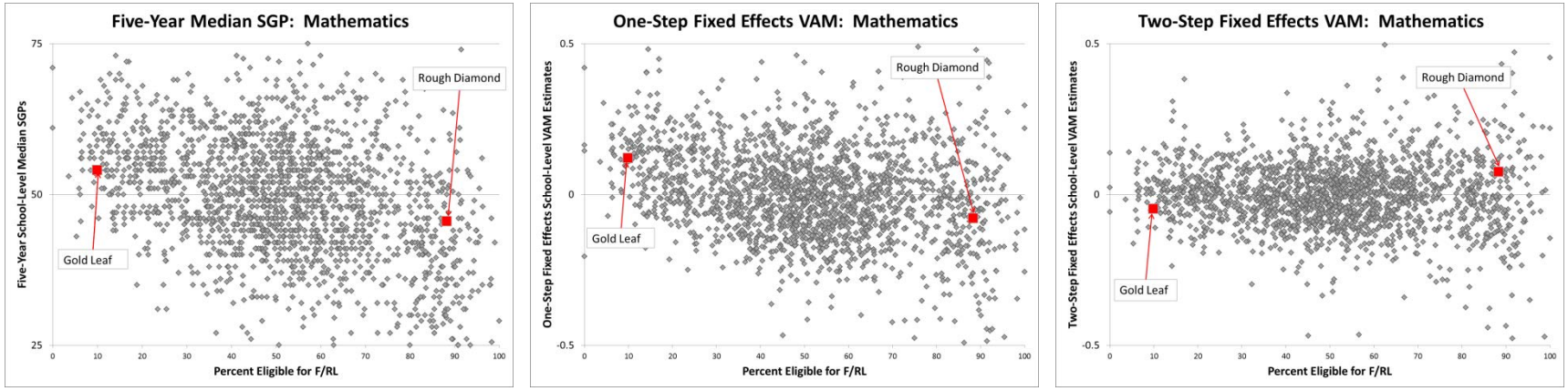

Table 1. Correlations in School-Level Estimates Across Models.

	SGP	One-step fixed effects	Two-step fixed effects
SGP	1.00	0.82	0.85
One-step fixed effects	--	1.00	0.84
Two-step fixed effects	--	--	1.00

Table 2. Representation of High-Poverty Schools in Top Quartile of Growth Estimates.

	SGP	One-step fixed effects	Two-step fixed effects
Share of high-poverty schools	0.042	0.104	0.152

Note: The share of high poverty schools in the overall analytic sample is 0.133.

Table 3. Average Share of Students Eligible for Free/Reduced-Price Lunch in Non-Overlapping Top-Quartile Schools Across Models.

	Outside of Top Quartile: SGP	Outside of Top Quartile: One-step FE	Outside of Top Quartile: Two-step FE
Top-Quartile: SGP	--	47.7	32.8
Top-Quartile: One-step FE	52.4	--	29.2
Top-Quartile: Two-step FE	69.7	60.5	--

Note: See text for a description of "non-overlapping top-quartile schools."

Appendix A
Data Description

Table A1. Data Details.

Student-Level	
Number of student test score pairs used in the model	1,572,601
Percent free/reduced-price lunch eligible	45.1%
Percent American Indian	0.4%
Percent Asian/Pacific Islander	1.8%
Percent Black	17.4%
Percent Hispanic	3.7%
Percent White	76.4%
Percent Multi-Racial	0.3%
Percent Female	48.9%
Percent of students with an individualized education plan (IEP)	13.2%
Percent of students with limited English proficiency	2.4%
Percent of mid-year building switchers	4.2%
Percent of students with missing lagged mathematics MAP score	0.3%
Percent of students with missing lagged communication arts MAP score	0.5%
School-Level	
Number of schools for whom a school effect was estimated	1,846
Average percent F/RL eligible	48.2%
Average percent minority	22.4%
Average percent female	48.3%
Average percent of students with an IEP	15.0%
Average percent of students with limited English proficiency	2.0%
Average percent of students who switched buildings mid-year	6.7%