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

AIR

Selecting Growth Measures for School and Teacher Evaluations: Should Proportionality Matter?

MARK EHLERT, CORY KOEDEL,
ERIC PARSONS
AND MICHAEL PODGURSKY

WORKING PAPER 80 • MAY 2013

Selecting Growth Measures for School and Teacher Evaluations: Should Proportionality Matter?

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 1
II. Models 3
III. Data 11
IV. Output from the Models.... 11
V. Model Selection 14
VI. Other Considerations.... 22
VII. Conclusion.... 24
References 25
Tables and Figures 29
Appendix A.... 33
Appendix B 34
Appendix C 37

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.

They thank Daniel McCaffrey, participants at the 2012 MARCES conference, and seminar participants at the Program for Education Policy and Governance at Harvard University for useful comments. They also gratefully acknowledge research support from CALDER (funded through Grant R305C120008 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education) 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: Should Proportionality Matter?

Mark Ehlert, Cory Koedel, Eric Parsons and Michael Podgursky

CALDER Working Paper No. 80
May 2013

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 under some conditions can be used 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 tied directly 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; Rockoff, 2004), 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.[2] We consider three competing approaches. The first approach, reflected in the popular student growth percentiles (SGPs) framework, 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 effects (Betebenner, 2009). Regression-based estimates similar in spirit to SGPs are

[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).
[2] Researchers have examined the potential to improve student achievement by using growth-based measures, mostly at the teacher level, to selectively shape the workforce (Boyd et al., 2011; Chetty et al., 2011; Goldhaber and Hansen, 2010; Hanushek, 2009; Staiger and Rockoff, 2010; Winters and Cowen, forthcoming). School-level measures could be used in a similar fashion. Growth-based measures can also be combined with other measures in evaluations; and used for other purposes like encouraging effort, targeting professional development, and sending useful signals about performance. We elaborate on these points below.

straightforward to construct, but in practice advocates of models that do not include student or schooling-environment controls have gravitated toward the SGP approach.[3]

The second approach that we consider is a one-step value-added model (VAM), which controls for student-background characteristics and schooling-environment factors while simultaneously estimating the growth measures for schools and/or teachers. The one-step VAM is widely used in the research literature, perhaps because under some conditions it can be used to disentangle the influence of context from school and teacher effects, in which case causal inference is possible.[4]

The third approach, also VAM-based, is designed to compare schools and teachers that serve observationally similar students. The rank ordering from an evaluation system based on the third approach need not be entirely consistent with a rank ordering based on the absolute "causal" achievement effects for schools and teachers. Rather, by construction, 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 (e.g., see Barlevy and Neal, 2012; Lazear and Rosen, 1981; Schotter and Weigelt, 1992).

We examine the appeal of these three modeling approaches in the context of an evaluation system for schools, although the substance of our findings will also apply to district- and teacher-level evaluations. We identify three key objectives of an evaluation system in education: (1) elicit optimal

[3] As with a sparse VAM, 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 to construct student peer groups.

[4] There is a large debate in the literature about whether causal inference can be supported for value-added estimates using a variety of model specifications. The issue of causal inference is not central to the model-selection argument we make in the present study, although it is of some comfort that a number of recent studies suggest that the scope for bias from properly-controlled growth models is small. A particularly compelling example is Chetty et al. (2011); also see Goldhaber and Chaplin (2012), Kane and Staiger (2008), Kinsler (forthcoming) and Koedel and Betts (2011). Studies that offer competing views include Briggs and Domingue (2011) and Rothstein (2009, 2010). Harris (2011) covers a number of issues related to value-added modeling, including the issue of bias.

effort from personnel, (2) improve system-wide instruction by providing useful performance signals, and (3) avoid exacerbating pre-existing inequities in the labor markets between advantaged and disadvantaged schools. When one considers any of these three objectives, we argue that the third modeling approach is preferable. The key distinguishing feature of the third approach – and the reason we advocate for its use in evaluation systems – is that it forces comparisons to be between equally-circumstanced schools and teachers. As a result of these forced comparisons, growth-based rankings are proportional to the evaluation sample throughout. For example, in a school-level evaluation, advantaged and disadvantaged schools are equally likely to be identified as top and bottom performers. We describe this feature of the model output as *proportionality*. When the question of model choice is framed within the context of designing an effective evaluation system, proportionality is a highly desirable modeling property.

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 for individual students. 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 under some conditions can be used 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 research literature and is motivated by 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.[5]

2.1 Student Growth Percentiles (SGPs)

Student Growth Percentiles (Betebenner, 2009) 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 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 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 outcome. 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. For students in higher grades, the SGP framework is extended to account for longer test-score histories, which determine students' comparison peer groups (the standard practice is to use same-subject score histories). 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 median SGPs can vary. In the subsequent analysis we use a median based on five years of student outcome data.

Although the SGP methodology differs from standard VAM methods in a number of ways, 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 (see Appendix B). Advocates of sparse growth models value their sparseness. They worry that, among other things,

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

conditioning on student or school-level characteristics lowers expectations for disadvantaged students.[6] We return to this point below (see Section 6).

2.2 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 (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. In the present study we model math test scores as outcomes and use communication-arts scores for the off subject.[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

[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, our study argues that the recommendation is misguided. It is also important to recognize that even sparse models, which only condition on students' prior scores, (empirically) lower expectations for disadvantaged students because lagged achievement is correlated with student demographics.

[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 (the standardized mean), 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.

data as within-year building switchers) and grade-level. The *S*-vector includes school-averaged student characteristics for these same variables.

Under some conditions, causal inference about school and/or teacher effects estimated by the one-step VAM can be supported.[8] However, a limitation of the one-step VAM is that by virtue of the one-step estimation, the coefficient vectors β_3 and β_4 are identified using within-school variation. For 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. An implicit assumption of the model is that students who differ in their eligibility for free/reduced-price lunch and attend the same school differ in the same way as students who differ in eligibility but attend different schools. That is, the one-step model uses performance differences between eligible and ineligible students at the same school to identify β_3^{FRL}, but this coefficient is also applied to predict performance differences between eligible and ineligible students *across schools*. There is no guarantee that the within- and between-school differences are the same, and no obvious way to test this assumption with available observational data.

The coefficients associated with the school-level variables (in the *S*-vector) are perhaps even more problematic. The purpose of including the school-level variables in the model, in the context of an evaluation system, is to control for schooling-environment factors that are outside of the influence of school personnel (Raudenbush and Willms, 1995). We note two issues with the incorporation of these controls in the one-step model that may limit their value in this role. First, like with the *X*-vector coefficients, the *S*-vector coefficients (β_4) are identified using 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

[8] In fact, Ballou et al. (2012) refer to a variant of the one-step model as the "true" model. Also see Ballou et al. (2004).

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. The same issue in extrapolating from the within-school variation to account for between-school differences arises. Because the school-composition variables are continuous, the identifying assumption can be described as requiring linearity in the effects of changes in school compositions on student outcomes. For example, 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.

A practical implication of the within-school identification approach is that the schooling-environment coefficients are unlikely to capture systematic differences in educational practice associated with school context.[9] For example, suppose that some instructional strategy *Q* is not feasible at schools that serve a large share of disadvantaged students. Within any particular school over a reasonably short timeframe, it is unlikely that the composition of the student body will change fast enough to lead to a change in the use of instructional strategy *Q*. Therefore, the coefficients in the one-step model attached to the schooling-environment controls, identified from within-school variation, will not capture variation in the use of instructional-strategy *Q*. At the same time, across-school differences in schooling environments will be associated with variation in the use of instructional-strategy *Q*. This is an example of schooling-environment information that the one-step model cannot capture.[10]

Another issue that may limit the value of using within-school variation to control for schooling-environment factors is measurement error. Causes of measurement error in the school-level variables likely include school-level roster inaccuracies, data-entry errors, etc. Because of measurement error, each observed school-composition variable is imprecise. The within-school estimation strategy in

[9] This concern is supported by Raudenbush and Willms (1995), who write "differences in school context must be assumed related to school practice." (p. 313)
[10] Below we use a labor-market example to illustrate this same general issue.

equation (1) is likely to exacerbate attenuation bias driven by measurement-error in the school-level variables for reasons similar to those discussed in previous work in other contexts (see Ashenfelter and Krueger, 1994; Griliches, 1979).

To illustrate the problem intuitively for the current application, in Appendix C we report variance decompositions for key school-level variables in our data panel. For the school-level share of students eligible for free/reduced-price lunch, Appendix C shows that nearly 90 percent of the total variance occurs between schools; for the disadvantaged-minority share, over 98 percent of the variance in the data is between schools.[11] While the vast majority of the total variance in these variables clearly occurs between schools, a much smaller share of the *measurement-error variance* is likely to occur between schools (given the nature of measurement error in these variables). This implies that the ratio of measurement-error variance to total identifying variation will increase through the reliance of the one-step model on within-school variation. Correspondingly, the coefficients on the school-level control variables will be attenuated, which will erode their value as schooling-environment controls.[12]

The issues we raise here are credible threats to the claim that the one-step model can be used to identify the causal effects of practice free from the confounding influence of contextual factors that are outside the control of education personnel. While mechanical identification of the model in (1) requires no more than multiple years of data per school, we submit that mechanical identification is not a sufficient condition for obtaining the parameters of interest (Mihaly et al. (forthcoming) raise a similar point in the context of evaluating teacher preparation programs). Ultimately, the research literature has yet to provide clear evidence showing that the parameter estimates in $\hat{\beta}_3$ and $\hat{\beta}_4$ sufficiently control

[11] One reason for the discrepancy across variables is that our data span the Great Recession. If we split our sample to look either pre- or post-2008, a much smaller share of the variation in the share of students eligible for free/reduced price lunch occurs within schools over time (that is, the variance decomposition suggests a within-between breakdown similar to what we find for the minority share).

[12] We are not aware of any evidence in the research literature to date that examines the ratio of measurement-error variance to total variance within and between schools using longitudinal education data. Such an analysis would be possible, but would require researchers to independently determine instances of measurement error in administrative education data files first.

for student characteristics and schooling environments. If these controls are insufficient, and if achieving academic growth in low-SES schools is truly more challenging, the likely bias in the output from the one-step VAM will favor high-SES schools at the expense of low-SES schools.[13]

2.3 Two-step VAMs

Our two-step VAMs are estimated as follows:

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

$$\eta_{isjt} = \delta_s + u_{isjt} \quad (3)$$

The variables in equation (2) are defined as in equation (1). There are two noteworthy differences between the one-step and two-step VAMs. 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 the schooling environment. In the formulation above, we incorporate the aggregate test-score controls into the school-level control vector S_{it}.[14] 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.

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

[13] Our discussion here presumes that the objective is to measure differences in school practice that lead to achievement gains. For other objectives, the one-step model, or even sparser models, may work well (e.g., if the objective is to inform parents of the schools that produce the highest growth for any reason, along the lines of what Raudenbush and Willms (1995) refer to as type-A school effects).

[14] 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. The correlation is much stronger when the time horizon is short (which is typically the case in these models). The problem in the one-step model is that with the school fixed effects and lagged school-averaged test scores in the same model, a low lagged average test score (t-1) suggests a stronger year for the school in year-t, conditional on the school fixed effects, which induces a positive association between low lagged aggregate achievement and current student scores. Per the above discussion, the model extrapolates this relationship across schools, which is problematic empirically. More details are available from the authors upon request. The two-step model circumvents this problem by sequentially attributing differences in student achievement to lagged aggregate test scores and current-year school effects, 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.

the measured dimensions to those characteristics. It is this aspect of the modeling structure that maintains proportionality in the output. It also allows for divergence between the estimates of δ_s and schools' causal effects. 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 this way, the two-step model has the potential to "overcorrect" for student disadvantage. In contrast, in equation (1), where the school effects and poverty effects are estimated simultaneously, the school-quality difference would not be absorbed by the poverty controls because the poverty effect would only be identified from within-school variation.

Recent research by Chetty et al. (2011) shows long-term effects of value added on student outcomes, where value-added is estimated using models that partial out the influence of the control variables first (like in our two-step model). Their findings lend credence to the general approach; however, whether or not the two-step model produces causal estimates is not central to the arguments that we make below.[15] Put differently, even if there is some divergence between schools' true value-added to student test scores and their growth ratings based on the two-step model, we argue that the two-step model is still the best choice for use in evaluation systems. The reason is that strict causal inference is not required – and in some contexts may actually be undesirable – in achieving the above-outlined system objectives.[16]

[15] Chetty et al. (2011) perform their analysis at the teacher level and include individual and classroom-level controls in their models. It is important to recognize that the Chetty et al. findings do not necessarily imply that the specification they use is the best specification for identifying causal estimates. Put differently, there may be a *better* causal model. Chetty et al.'s results are best interpreted as showing that their approach – which is analogous to our two-step model – produces estimates that are, at the least, strongly positively correlated with causal effects.

[16] A notable omission from our list of model choices is EVAAS®, a propriety modeling approach developed and marketed by SAS. EVAAS® does not fit easily into any of the three modeling classes that we consider; it is best viewed as a case somewhere in-between the sparse model and the two-step VAM. Regarding the issue of proportionality, which is central to our work, Sanders et al. (2009) note that proportionality is sometimes but not always maintained by the EVAAS® modeling approach. They indicate that correlations between EVAAS® estimates and aggregated student characteristics are "modest at worst and essentially zero at best" across a number of locations where the model is used. This is likely the case for a number of alternative modeling approaches and

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 linked 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 discussion, 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 education policymakers as a better way to assess performance 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

---

will depend heavily on the evaluation context. The key takeaway as it relates to our work is that, unlike in the two-step model, proportionality is not guaranteed by the EVAAS® approach. That said, the approach could be easily adjusted to make the output proportional. See Ballou et al. (2004) for more information about EVAAS® (note that Ballou et al. (2004) argue against imposing proportionality).

influence of factors outside of the control of schools to the "performance" measures. That is, some of the highest-ranked schools on the vertical axis may not be performing particularly well; and alternatively, some of the lowest-ranked 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.[17] 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 same subject test scores as are available for the student 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.

The last panel in Figure 2 plots the school-level growth estimates from the two-step VAM. By construction, the proportional 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

[17] 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 arguments do not depend substantively on whether the VAM estimates are shrunken or not.

representation comes from the fact that differences in schooling environments and school characteristics, including poverty share, are partialed out before the growth measures are estimated.[18] 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 observationally-similar schools, large differences in annual test-score growth are clearly visible.[19]

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 simple 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 share of disadvantaged schools in the analytic sample with the share in the top quartile of the rankings from each growth model. Disadvantaged schools in Table 2 are defined as those with at least 80 percent of students eligible for free or reduced-price lunch.[20] In the full evaluation sample, 13.3 percent of Missouri schools are identified as disadvantaged by this definition. Consistent with the visual

[18] The representation is not exactly even because of weighting. If we estimate the model using a single year of outcome data and assign a school-level growth measure to each student, then correlate the school-level growth measures and school characteristics using the student weights, the correlations are precisely zero by construction. However, when we correlate the aggregated school characteristics with the school-level growth estimates from our models, the weighting deviates from the student weights (there is one observation per school, rather than per student), which leads to small non-zero correlations. An additional complication is that the number of student observations in each school varies by year. The lack of precise proportionality in the two-step model derives from a combination of the above factors, along with the fact that we do not include a control for the five-year average of the aggregated school characteristics directly in the model (note that including five-year school averages in the first-stage estimation would lead to zero correlations weighted at the student-level, but would not lead to zero correlations at the school-level). As we discuss below, precise proportionality can be achieved with one additional regression step where the school effects themselves are the unit of analysis. Further information on this issue is available from the authors upon request.

[19] 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.

[20] The findings reported in Table 2 are not qualitatively sensitive to how disadvantaged schools are defined.

representation in Figure 2, clear differences emerge across the models. 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.[21] 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.[22]

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 proportional 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 that figures prominently in policy discussions in many states and school districts across the nation, and to which we now turn.

V. Model Selection

In this section we compare the three modeling options with respect to their alignment with the three policy objectives that we outlined above: (1) elicit optimal effort from personnel (i.e., teachers and administrators), (2) send signals to schools that will lead to improved instruction, and (3) avoid

[21] Again, the output from the two-step model is not precisely proportional because of weighting issues related to our use multiple years of data and our *ex post* evaluation of school-level estimates obtained from a student-level regression (see footnote 17). If desired, precise proportionality in the output could be achieved by using the residuals from an auxiliary school-level regression of the estimated school effects on panel-average school characteristics.

[22] 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 a proportional model.

exacerbating the weak labor-market position of disadvantaged schools. We argue that the proportional model is preferable for use in achieving each of these objectives.

5.1 Eliciting Optimal Effort from Personnel

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), which focuses on the efficient design of incentive pay for teachers.[23] 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. Indeed, Barlevy and Neal note favorably the attractive features of the SGP approach in this regard.[24]

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.[25] 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

[23] A seminal paper in the larger incentive design literature is Lazear and Rosen (1981). A widely cited, although somewhat dated, survey of this literature is Prendergast (1999).

[24] 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.

[25] 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.

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 the 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 proportional model achieves this objective. In contrast, the SGP and one-step VAM approaches do not result in balanced comparisons across school types and in fact, favor the advantaged group, which runs counter to the goal of eliciting optimal effort.[26]

From an incentive-design perspective, then, previous research in economics has established that comparisons from a proportional model are preferable. However, it is worth noting that the small body of evidence we have on the potential for incentives to improve educator effort in the United States is mixed. Springer et al. (2010) conduct an incentive experiment for teachers in Tennessee and find no discernible effort effects. One explanation for this result, consistent with teacher responses to surveys, is that teachers were already supplying considerable effort prior to enrolling in the incentive program. In contrast, Imberman and Lovenheim (2012) use variation across teachers in the strength of the performance incentives to which they are exposed to show that teachers who face stronger incentives are more productive than teachers who face weaker incentives. Although there is still much work to do in this area, it seems likely that policymakers will prefer a system that encourages more effort from education personnel over a system that encourages less, an objective which is best achieved by creating balanced comparisons.[27]

[26] Some 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.

[27] The mixed evidence on effort responses to incentives in the United States is in contrast with consistent international evidence that incentives increase effort, notably Duflo, Dupas and Kremer (2012) and Muralidharan and Sundararaman (2011).

5.2 Useful Instructional Signals

Growth models can be used to improve instruction in K-12 schools by reinforcing positive educational practices and discouraging negative ones.[28] For example, a positive performance signal from the growth model might encourage a school to continue to pursue and augment existing instructional strategies. 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. As an example, an underperforming school may benefit from observing a school that is performing better, but this benefit will only be attainable if the system provides useful information to direct educator-to-educator learning (i.e., if the system 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, by other means.

In terms of providing useful performance signals, we argue that the two-step model is again preferable. This is true regardless of whether the two-step estimates are "overcorrected" per the discussion in Section 2.3. Put differently, even if the two-step estimates mask differences in absolute performance across schools in different contexts, they still facilitate context-conditional comparisons, which are sufficient to send performance signals to schools that can be used to improve system-wide instruction.

To illustrate, assume for the moment that high-SES schools really are more effective at raising student achievement, a fact that would be masked by the output from the two-step model. A potential mechanism supported by previous research is that high-SES schools have access to a stronger labor

[28] In the context of teacher evaluations, recent evidence from Taylor and Tyler (2011) shows that evaluation systems can lead to improved educational performance. The teacher-evaluation system studied by Taylor and Tyler is significantly more involved than a system that simply provides growth measures to schools or teachers. Still, it establishes the potential for instructional improvements to arise from evaluation systems in education. The Talyor and Tyler (2011) study is agnostic about why the program they study was successful; their research design cannot identify the specific component(s) of the evaluation system that caused the improvements in teacher performance.

market (Boyd et al., 2005; Koedel et al., 2011; Jacob, 2007; Reininger, 2012). For this reason, it may be that high-SES schools really do have better teachers and administrators, and therefore, produce more test-score growth.

There are three primary factors that determine personnel quality in a given school: (1) the quality of the applicant pool, (2) conditional on the quality of the applicant pool, the success of the school in selecting the best applicants, and (3) teacher retention. Disadvantaged schools have access to lower-quality applicants for many reasons, most of which are outside of their control (Boyd et al., 2005; Jacob, 2007; Reininger, 2012). 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 applicant quality, 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. A similar argument holds if context is an important determinant of teacher retention, as is suggested by Hanushek et al. (2004) and Lankford et al. (2002).

Other examples can be readily imagined. One possibility touched on in Section 2.2 is that the learning environments in high-SES schools may facilitate the use of effective instructional practices that are infeasible at low-SES schools. While these examples are useful for their simplicity, the more-general issue is that advantaged and disadvantaged schools differ along many dimensions. These dimensions likely influence what constitutes effective "practice" in schools, which we define broadly to include curriculum implementation, instructional practice, personnel practice, and all other day-to-day decisions that combine to create the educational environment in schools. 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. The research literature does not provide concrete guidance for what different types of schools should be doing to improve achievement – in fact, it is not clear that this is the type of information that

researchers *can* provide. However, what researchers can do is help state and local education agencies evaluate performance conditional on the contexts in which schools are operating.

Figure 3 provides a concrete example of the types of problems that can arise from an evaluation system that does not maintain proportionality. 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 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 signal from either of these models, Rough Diamond might feel pressure to make substantial changes to the delivery of instruction in response to the negative performance rating.[29] However, whatever Rough Diamond is doing seems to be working quite well in the environment in

[29] At the other end of the spectrum, Rough Diamond could choose to ignore the performance signals provided by the non-proportional models. Although doing so may well be preferred in Rough Diamond's case, clearly any widespread dismissal of performance signals is antagonistic to the goal of improving system-wide instruction as part of an evaluation system. We also note that any real-world system will be based on multiple performance measures. We are not advocating that schools and/or teachers respond entirely to output from growth models without confirmation from alternative measures. Nonetheless, for the purpose of determining what types of measures should be incorporated into evaluation systems, it is useful to consider how schools and/or teachers might respond to the information contained by each measure in isolation.

which it is operating; put differently, if Rough Diamond were to start over completely, in expectation it would do worse.[30]

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 encourage Rough Diamond to continue to pursue and refine its current instructional and personnel strategies, which have placed it well above average relative to observationally similar schools.

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 underwhelming. Similarly, it determines whether principals from other 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 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 might 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.[31] More generally, 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

[30] Here we use the word "expectation" in the statistical sense. To say this differently, if the personnel at Rough Diamond were to be completely relocated and the instructional strategy completely rebuilt from the ground up, the data suggest that, conditional on the types of students who attend Rough Diamond, its performance would decline.
[31] This, of course, is one of many possibilities, per the above discussion. Also note that differences in applicant-pool quality are likely to be relevant in school-leader labor markets (Koedel et al., 2011).

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 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. A potentially harmful consequence of such a system is that it could result in a perpetuating cycle of the destruction and re-invention of instructional practices at disadvantaged schools, whether these practices are effective or not (conditional on circumstance). Alternatively, a system that differentiates schools conditional on disadvantage can highlight the large performance differences among observationally similar schools across the school spectrum. 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.[32] 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 useful 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.

5.3 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

[32] Deming et al. (2012) show that quality differences between observationally similar schools meaningfully impact long-term student outcomes.

associated with working in challenging educational environments. An important benefit of the proportional model is that the "winners" and "losers" from the evaluation will be broadly representative of the system as a whole (see Table 2).[33]

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. Although we understand and appreciate the spirit of this concern, in our view it is 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 policymakers to clearly see absolute differences in achievement across schools, and proficiency gaps that are unadjusted for student or school characteristics, regardless of which growth model is adopted. Dual reporting of similarly-circumstanced comparisons along the lines of those produced by the two-step model, in conjunction with information about absolute achievement levels, 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.[34]

[33] If the current personnel at disadvantaged schools truly are less effective, the two-step model will provide equity-enhancing incentives that encourage teachers and principals to move into disadvantaged schools. The reason is that it will be easier for personnel in disadvantaged schools to positively distinguish themselves within an evaluation system that relies on a proportional model.

[34] This type of dual reporting may also help highlight the need for policy intervention at a higher level. For example, one could imagine a system that combines the information in Figure 1 with the information in the third panel of Figure 2. Side-by-side reporting of this information would reveal that even the most effective disadvantaged schools, per the similarly-circumstanced comparisons, are falling short of achieving targeted test score levels. An implication is that closing the achievement gap at these schools will likely require outside intervention; put differently, the data

A related concern is that a proportional model will lower expectations for students in disadvantaged schools (and additionally, for disadvantaged students through the use of student covariates). However, it is important to recognize that setting expectations for individual students is not the purpose of the models that we consider here. The purpose is solely to achieve the three evaluation-system objectives outlined above. Philosophically, policymakers may not want to lower expectations for disadvantaged students. If this is the case, then the proper approach to student-level evaluation is to set fixed success benchmarks for all students and evaluate progress toward those benchmarks. None of the three growth models that we consider here are designed to achieve this objective. For example, even the sparse model allows for different growth targets for different types of students by conditioning on individual prior achievement.[35] This issue highlights the importance of framing the purpose of the growth model. Given the objectives outlined above, proportionality is a desirable property of an effective growth model couched within the context of an evaluation system for education personnel. Models used for other purposes may need to be designed differently.

We also note that there are a number of ways to achieve proportionality beyond following the exact two-step approach we use here. For example, any output from an initial model that does not achieve proportionality can always be regression-adjusted *ex post*.[36] Also, one could design proportional models within the percentiles framework used by Barlevy and Neal (2012) and Betebenner (2011), which would facilitate ordinal rather than cardinal comparisons between students. This could be achieved by converting all student scores to percentiles before the first-step regression and would be

---

indicate that behavioral changes by the personnel currently working in these schools are unlikely to be sufficient to raise achievement to the desired level if they are not accompanied by other interventions.

[35] This is an empirically accurate statement, although conceptually policymakers may find it more palatable to distinguish between conditioning on prior achievement and conditioning on student demographics. For example, empirically, African American students have lower test scores on average than white students. By conditioning on prior individual achievement, even the sparsest growth model is implicitly allowing for variable growth targets between African American and white students.

[36] In fact, our two-step model is not precisely proportional due to weighting issues (see footnote 17). If we wanted to achieve precise proportionality (where the correlations between school characteristics and the growth measures were all exactly zero), we could perform a school-level regression of the estimated two-step growth measures on panel-average school characteristics and use the residuals.

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.

VII. Conclusion

We examine three approaches to modeling student test score growth – Student Growth Percentiles (SGPs), a one-step VAM, and a two-step VAM. These models reflect the spectrum of choices for policymakers in their efforts to design evaluation systems for schools and teachers. All three approaches produce growth measures that are highly correlated (0.82 – 0.85). The high correlations, however, mask an important difference. A key distinguishing feature of the two-step approach is that it produces growth rankings that are proportional to the evaluation sample. Put differently, the two-step approach 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 evaluation systems in education, the proportionality property of the two-step model is highly desirable.

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.

Ashenfelter, Orley and Alan Krueger. 1994. Estimates of the Economic Return to Schooling from a New Sample of Twins. *American Economic Review* 84(5), 1157-1173.

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, Damian W. 2009. Norm- and Criterion-Referenced Student Growth. *Educational Measurement: Issues and Practice* 28(4), 42-51.

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, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2011. Teacher Layoffs: An Empirical Illustration of Seniority v. Measures of Effectiveness. *Education Finance and Policy* 6(3), 439-454.

Boyd, Donald, Hamilton Lankford, Susanna Loeb. 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.

Deming, David, Justine S. Hastings, Thomas J. Kane and Douglas O. Staiger. School Choice, School Quality and Postsecondary Attainment. NBER Working Paper No. 17438.

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

Griliches, Zvi. 1979. Sibling Models and Data in Economics: Beginnings of a Survey. *Journal of Political Economy* 87(5), Part 2, S37-S64.

Hanushek, Eric A. 2009. Teacher Deselection, in *Creating a New Teaching Profession* eds. Dan Goldhaber and Jane Hannaway. Urban Institute, Washington, DC.

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, 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.

Imberman, Scott A. and Michael F. Lovenheim. 2012. Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System. NBER Working Paper No. 18439.

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.

Lankford, Hamilton, Susanna Loeb, James Wyckoff. 2002. Teacher Sorting and the Plight of Urban Schools. *Educational Evaluation and Policy Analysis.* 24(1), 37-62.

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.

Mihaly, Kata, Daniel McCaffrey, Tim R. Sass and J.R. Lockwood (forthcoming). Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. *Education Finance and Policy*.

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.

Raudenbush, Stephen and J. Douglas Willms. 1995. The Estimation of School Effects. *Journal of Educational and Behavioral Statistics* 20(4): 307-335.

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.

Sanders, William L., S. Paul Wright, June C. Rivers and Jill G. Leandro. 2009. A Response to Criticisms of SAS® EVAAS®. SAS Institute white paper (November).

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.

Staiger, Douglas, Jonah Rockoff (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.

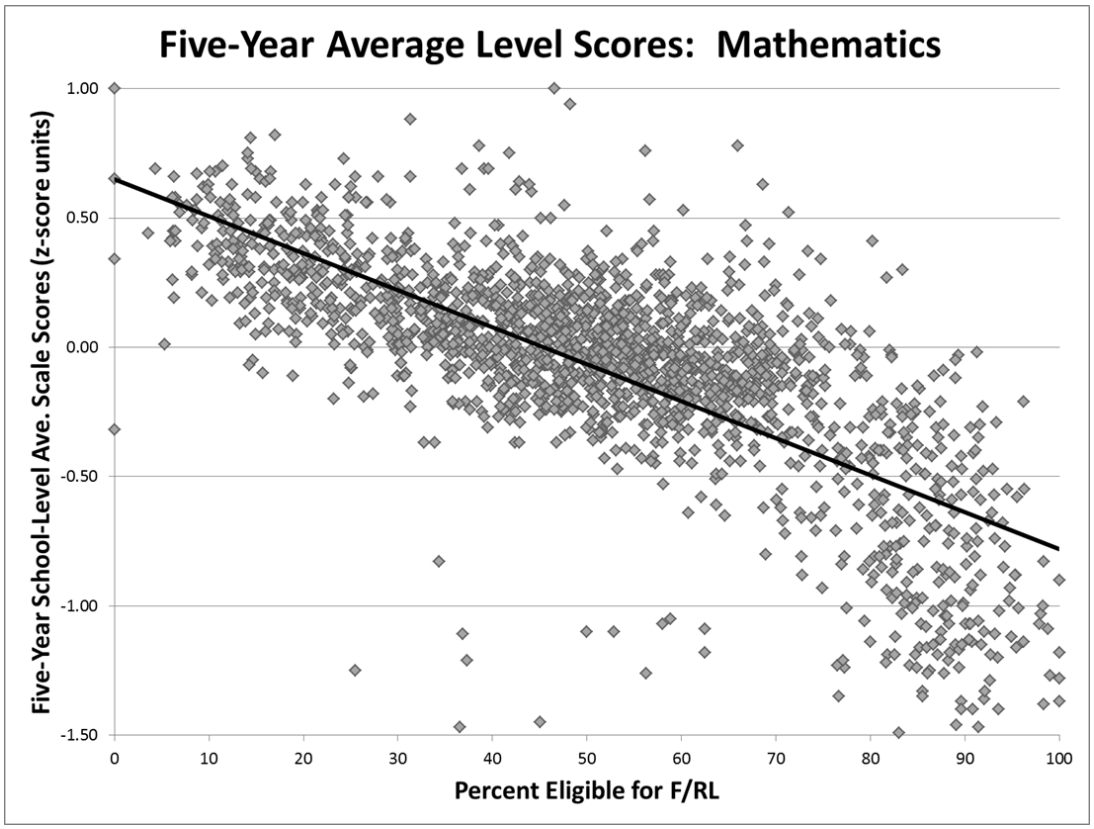
Taylor, Eric S. and John H. Tyler. 2011. The Effect of Evaluation on Performance: Evidence from Student Achievement Data of Mid-Career Teachers. NBER Working Paper No. 16877.

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

Winters, Marcus A. and Joshua M. Cowen (forthcoming). Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. Journal of Policy Analysis and Management.

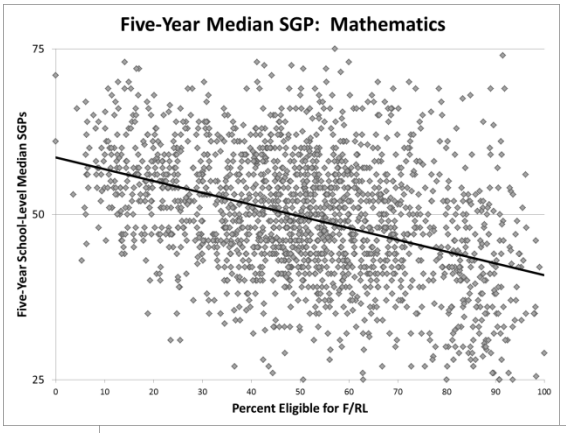
Tables and Figures

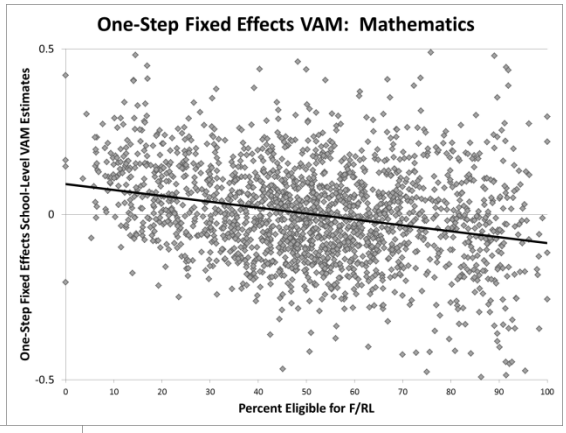
Figure 1. School-Average Test Scores Plotted Against School Shares Eligible for Free/Reduced-Price Lunch

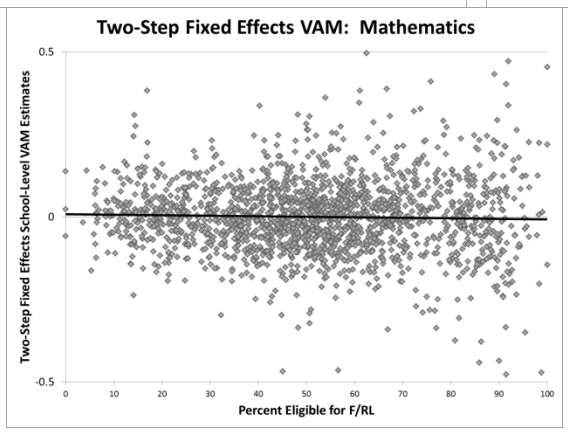


Correlation: -0.68

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

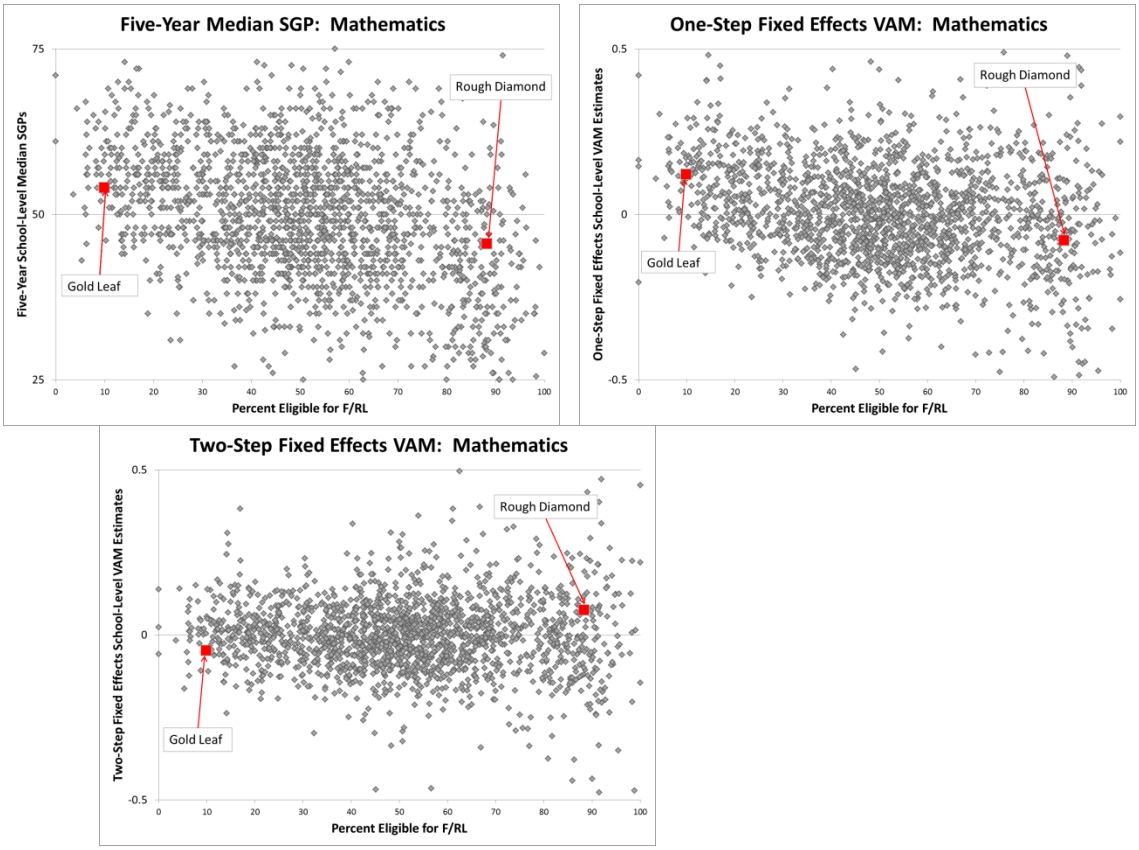






Correlation: -0.37 Correlation: -0.25 Correlation: - 0.03

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



Note: Correlations are the same as in Figure 2.

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 which 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%

Appendix B

Constructing a Sparse VAM to Approximate Median SGPs

In this appendix we briefly show that a simple, sparse VAM can be constructed – within the linear-regression framework – to produce school-level growth measures that are very similar to the median SGPs shown in the text. We construct our linear sparse model to contain the same information that is used for the median SGPs. Specifically, we predict students' current scores as a function of their prior score histories (in the same subject) and include indicator variables for missing test-score information. The model is run separately by year and grade and estimated via ordinary least squares (OLS). The fullest specification of the model is given by the following equation (estimated for grade-8 students in the final year of the data panel):

$$\begin{aligned} Y_{ijt} = {} & \beta_0 + Y_{ij(t-1)}\beta_1 + Y_{ij(t-2)}\beta_2 + Y_{ij(t-3)}\beta_3 + Y_{ij(t-4)}\beta_4 + Y_{ij(t-5)}\beta_5 + \\ & MS_{ij(t-1)}\beta_6 + MS_{ij(t-2)}\beta_7 + MS_{ij(t-3)}\beta_8 + MS_{ij(t-4)}\beta_9 + MS_{ij(t-5)}\beta_{10} + \varepsilon_{ijt} \end{aligned} \quad \text{(B.1)}$$

In equation (B.1) Y_{ijt} is a test score for student *i* in subject *j* in year *t*, and MS_{ijt} is an indicator variable equal to one if the score in that year and subject is missing for student *i*.[37] As mentioned previously, the equation shows the full model that we estimate for grade-8 students in the final year of our data panel; for earlier years and grades we remove lagged-score controls as necessary.

In a second step we group the residuals from equation (B.1) by school and take the median residual for each school, which we use as the growth measure. Note that one could construct a sparse VAM that reflects SGPs even more closely (for example, one could use indicator variables for "peers" with similar score histories in place of the lagged-score controls),

[37] In cases where MS_{kjt} is equal to one for period *k*, Y_{ijk} is set to the sample average score.

but as we show below, the specification above is sufficient to produce estimates that are very similar to the median SGPs.

Appendix Table B.1 shows that the correlation between the median SGPs and sparse-VAM estimates is 0.97. Appendix Table B.2 replicates Table 2 from the main text, showing that the degree of underrepresentation for high-poverty schools in the top quartile of school rankings is similar using the sparse-VAM and SGP frameworks. Finally, Appendix Table B.3 shows correlations between the estimated growth measures from each model and the school-level share of students eligible for free/reduced-price lunch (the correlations for SGPs and the one- and two-step VAMs are as reported in Figure 2). The strong similarity between median SGPs and the sparse-VAM estimates is again apparent.

Overall, this appendix shows that there is nothing particularly special about the SGP framework in the sense that it does not produce output that meaningfully differs from a simple value-added model based on similar information. Thus, the proportionality issue raised in our study generalizes beyond SGPs to other sparse models.

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

	SGP	One-step fixed effects	Two-step fixed effects	Sparse VAM
SGP	1.00	0.82	0.85	0.97
One-step FE	--	1.00	0.84	0.80
Two-step FE	--	--	1.00	0.82
Sparse VAM	--	--	--	1.00

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

	SGP	One-step fixed effects	Two-step fixed effects	Sparse VAM
Share of high-poverty schools	0.042	0.104	0.152	0.050

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

Appendix Table B.3. Correlations Between the School-level Share of Students Who are Eligible for Free/Reduced-Price Lunch and School-Effect Estimates from Four Different Models.

	SGP	One-step fixed effects	Two-step fixed effects	Sparse VAM
Correlation with FRL share	-0.37	-0.25	-0.03	-0.38

Appendix C

Within/Between Variance Decompositions for Key School-Level Composition Variables

In this appendix we decompose the total variance for two school-level control variables: (1) the share of students who are eligible for free/reduced-price lunch, and (2) the share classified as disadvantaged minorities (African American or Hispanic). Note that the one-step VAM identifies the parameters on the school-level control variables using only within-school variance.

Appendix Table C.1 shows that most of the variance in both of these variables occurs between schools. This means that only a small fraction of the total variance in these variables is used for identification in the one-step VAM.[38] Alone, these decompositions do not provide conclusive evidence to dismiss the possibility that the one-step VAM can adequately control for schooling environments. However, in conjunction with the discussion we provide in Section 2.2, they do raise questions about the extent to which this variation can truly be used to capture key aspects of schooling environments that are likely to influence student achievement. As we note in the text, mechanical identification is not a sufficient condition to ensure that the parameters of interest are properly identified in standard value-added models.

[38] Although most of the variance occurs between schools for both variables, the between variance share is much larger for the FRL variable. A key reason is that our data panel spans the Great Recession. If we perform similar variance decompositions using data entirely from either before or after the Great Recession, the within-between decompositions look similar for both variables, and much more like what we show for the minority share in Appendix Table C.1.

Appendix Table C.1. Within- and Between-School Variance Decompositions for the School-Level Shares of Students Who are Eligible for Free/Reduced-Price Lunch and Who are Disadvantaged Minorities.

	Total Variance	Between Variance	Between Variance Share	Within Variance	Within Variance Share
FRL Share	520.5	463.9	89.1%	56.6	10.9%
Minority Share	893.6	883.6	98.9%	10.0	1.1%