Technical Report

WHAT GETS MEASURED GETS DONE: MULTIPLE MEASURES, VALUE-ADDED, AND THE NEXT GENERATION OF ACCOUNTABILITY UNDER ESSA

EDUCATION RESEARCH ALLIANCE FOR NEW ORLEANS

Douglas N. Harris, Tulane University
Lihan Liu, Tulane University

May 25, 2018

EducationResearchAllianceNOLA.org

What Gets Measured Gets Done: Multiple Measures, Value-Added, and the Next Generation of Accountability under ESSA

Douglas N. Harris
Lihan Liu

May 25, 2018

Abstract: We identify five general principles for measuring school performance and show that, even in the age of ESSA, state governments are still mainly holding schools accountable based on student achievement levels and therefore apparently violating two principles: (a) that the mission of schools is to prepare students for long-term life outcomes; and (b) that accountability will be most effective when holding schools accountable for that which they can control. Moreover, we show that the violation of these principles has consequences for schools. Using statewide student-level data from Louisiana, we show how sensitive performance measures are to alternative performance measures that more closely follow the above two principles (albeit with some potential trade-offs with other principles). Adding college entry to a high school performance measure that is equally weighted between test scores and high school graduation, for example, leads about one-quarter of schools to switch performance categories. Switching from test score levels-only to an equal weight between levels and value-added has a somewhat smaller effect. We also carry out simulations that quantify the improvement in student outcomes from switching from levels-only to an equally weighted combination of levels and value-added across all student outcomes (not just test scores), under a state intervention policy. The results suggest that there is room for improvement in the next generation of school performance measurement and school accountability, improvements that would help raise actual student outcomes.

Acknowledgements: This study was conducted at the Education Research Alliance for New Orleans at Tulane University. The authors wish to thank the organization's funders: the John and Laura Arnold Foundation, William T. Grant Foundation, the Smith Richardson Foundation, the Spencer Foundation and, at Tulane, the Department of Economics, Murphy Institute and School of Liberal Arts. We particularly, thank Nathan Barrett, Morgan Polikoff, Sara Slaughter, and participants in the 2018 annual meeting of the Association for Education Finance and Policy. For their outstanding research assistance on the state ESSA plan analysis, we thank Catherine Balfe, Molly Kalat, and Natalie Philips.

Author Information: Douglas N. Harris (corresponding author) is Professor of Economics, Schleider Foundation Chair in Public Education, and Director of the Education Research Alliance for New Orleans at Tulane University (dharri5@tulane.edu). Lihan Liu is a Senior Research Fellow at ERA-New Orleans (lihanliu24@gmail.com).

I. Introduction

Accountability for student outcomes represents arguably the most important education policy trend of the past quarter-century. Many states instituted such plans during the 1990s, and these had some impact on student outcomes (Carnoy & Loeb, 2003). In 2001, Congress required test-based accountability by passing President George W. Bush's signature proposal, *No Child Left Behind* (NCLB). Among other things, the law subjected schools in all states to a gradually intensifying cascade of interventions in schools not meeting Adequate Yearly Progress toward the goal of 100 percent proficiency (Jennings & Renter, 2006; Dee & Jacob, 2011).[1] More recently, while the focus on test scores has continued in law and in practice, the *Every Student Succeeds Act* (ESSA) has eliminated the 100 percent proficient goal and given states more flexibility over many of the details, including the types of school performance measures they use (Klein, 2016). The present study is designed to help state policymakers better understand the trade-offs involved when choosing different types of school performance measures and therefore to help them design accountability systems to better meet their own objectives.

One contribution of the present study is to show that the design of performance measures for accountability, such as those in NCLB and ESSA, can be informed and guided by general principles. Research from psychology and management (DeNisi & Pritchard, 2006), as well as political discourse about accountability, suggests that a variety of principles are relevant. While it would be difficult to develop an exhaustive list, we propose five principles as a starting point. The construct of performance is defined as: (1) outcomes that are predictive of students' long-term life outcomes; and (2) what educators can control. Moreover, the measures of these constructs should be: (3) valid and reliable; (4) inexpensive; and (5) simple and intuitive.

[1] Strictly speaking, states could avoid these requirements by forgoing federal Title I funding, but no state was willing to sacrifice these funds.

In addition to explaining the underlying rationale for these principles, we attempt to highlight the trade-offs involved, both in theory and in practice. For example, one way to follow principle 1 is to hold schools accountable for college entry, but depending on how this is implemented, doing so may come at the expense of principle 2. Likewise, shifting toward value-added measures would help satisfy principle 2, but perhaps at the expense of principle 5 related to simplicity (Harris, 2011).

Whether any given trade-off is worthwhile depends in part on how much performance measures actually change. Do performance measures and school ratings change much when we increase the weight given to measures, such as college entry, that predict long-term life success and/or when we hold schools accountable for what they can control using value-added measures? The answer is not obvious. The degree to which adding any measure changes performance measures depends, as we show, on the correlation between the old and the new measures. The correlations among some common measures are relatively high, and this keeps the performance measures from changing dramatically. We also show that school performance measures are subject to diminishing returns; that is, policymakers can keep adding measures, but each additional measure has less practical impact on actual performance ratings.

These potential changes in today's school performance measures, to the degree they reflect improved validity, could lead to better actual school performance through at least three pathways: (1) high-stakes accountability may lead governments to intervene in different schools; (2) improved performance measures send more accurate signals to school leaders about their performance and the success of their practices and school improvement initiatives; and (3) parents, armed with better information, are more likely to choose schools (directly or indirectly through housing choices) that are higher-performing. While there is conflicting evidence about just how much families pay attention to common measures like test scores and graduation when choosing schools, research consistently shows that families do respond to this information to some degree, both in traditional

district systems where school choice is driven by housing choices (e.g., Black, 1999; Figlio & Lucas, 2004) and in newer choice-based systems (Harris & Larsen, 2016; Glazerman & Dotter, 2017). In short, mismeasurement of school performance has real consequences for the actual school performance that students experience and therefore for student outcomes.

We quantify the practical impact of the first of these mechanisms through a policy simulation. Suppose that policymakers close or take over the bottom five percent of schools for each of four consecutive years (as occurred recently in New Orleans) and that we wish to compare an accountability system that includes only test score and high school graduation levels (equally weighted at 1/2) with one that has four measures: test score levels, test score value-added, high school graduation levels, and high school graduation value-added (equally weighted 1/4 each). Moving to this new composite measure would increase actual school performance in the bottom 20% of schools by 0.4 percentiles and increase those schools' graduation rates by 0.4 percentage points. While these could be interpreted as small effects for such an aggressive policy, it is important to recognize that this policy has essentially zero economic cost. The effects also ignore the ways that better performance measures could improve student outcomes through other mechanisms (school choice and internal school improvement), effects that are more difficult to simulate.

We make four main contributions: (a) outlining the principles of performance measurement and their trade-offs; (b) showing that state ESSA plans violate some of these principles; (c) proposing ways to better align school accountability with at least some of the principles (adding college entry and value-added, and applying value-added to measures beyond test scores); and (d) showing the potential practical implications of these solutions through simulations of closure and takeover policies. Some elements of part (c) have also been explored by McEachin and Polikoff (2012), but we take this further is simulating the effects on student outcomes.

In Section II, we describe the principles of performance measurement in more detail. This is followed by a brief summary of our statewide, student-level, longitudinal

data (Section III). Section IV presents results comparing test levels to high school graduation and college entry levels. Section V briefly describes the various value-added models we estimate. Section VI shows the correlations between test score value-added and test levels. Finally, to test the practical implications of this, we simulate in section VII the effects of a state intervention policy giving more weight to value-added. Section VIII concludes.

II. Principles of School Performance Measurement

II.A. Five Principles

In this section, we list the principles, elaborate on the reasoning behind them, highlight their interconnections, and introduce some possible ways to improve alignment between actual accountability and the principles.

1. Performance measures should focus on what society expects schools to accomplish, particularly to produce outcomes that are predictive of students' long-term life outcomes. What specific long-term outcomes are most important to society is a matter of philosophy, but there seems to be little debate that schools should focus on preparation for adulthood, broadly defined.[2]

2. Performance measures should focus on what educators can control. This is closely related to the "cardinal rule" of accountability (Harris, 2011) and is rooted in the idea that holding people accountable for factors outside their control reduces response and may reduce motivation (DeNisi & Pritchard, 2006).[3] The principle has two important implications: (a) that we have to isolate the contributions to student outcomes made by

[2] This principle also rests on the assumption that accountability-induced improvement in student outcomes that are predictive of long-term life success holds the best prospect for increasing actual long-term student outcomes. This may not be completely true as attaching stakes to any measure leads to distortions in outcomes (Campbell's Law), e.g., high-stakes math tests predict long-term outcomes less well than low-stakes tests. However, this problem arises with all measures to which stakes are attached, not just those that are strongly predictive, so the assumption is plausible.

[3] DeNisi & Pritchard (2006) write, "Performance management and performance appraisal systems that strengthen the perceived connection between actions and results will be associated with a higher level of performance improvement" (p.265). A corollary to this principle is that we should hold schools *partly* accountable for factors that are *partly* within their control (Harris, 2011).

educators from those made by students themselves, their families, communities, and others; and (b) that student outcomes need to be proximal in time to the time educator actions occur. If we are trying to measure educators' performance in time *t*, but the outcomes arise in period *t+x*, then *x* should be as small as possible. The most obvious reason for this is that the educators in a school in time *t* might not be working in the given school many years in the future.

These first two principles are the focus of the present study and help define the construct of school performance. These are not the only principles that matter, however:

3. Performance measures should be valid and reliable. We mean this in the sense that the measures capture the construct on average and that there is limited random error. Math tests should measure the relevant math skills with a high degree of precision, for example. As we discuss later, any form of error in performance measures can be problematic in accountability.[4]

4. Performance measures should be inexpensive. This is a matter of simple cost-benefit analysis. Resources devoted to performance measurement cannot be used for other aspects of the educational enterprise.

5. Performance measures should be simple to understand. This principle, suggested by some scholars of school accountability (Deming and Figlio, 2016) as well as general personnel performance assessment (e.g., Bowman, 1999), is partly rooted in evidence that people have limited cognitive ability and are therefore constrained in their ability to process vast quantities of information and therefore that decisionmaking suffers under heavy cognitive loads (e.g., Sweller, 1994; Deck & Jahedi, 2015). However, note that this principle 5 is really about what happens when providing more information for a given set of decisions (e.g., improving educational practice in schools), which themselves may be

[4] Note that since we are focused on school-level performance, and averaging outcomes across large numbers of students, reliability of the testing instrument may not be a great concern. As we discuss later with value-added, however, this is not the only immediate source of random error.

inherently complex. While we could not find direct evidence on this point in the context of personnel performance appraisal, there is some prior evidence that providing more information can lead to worse consumer decisions (Keller & Staelin, 1987; Ariely, 2000).

As noted earlier, there are trade-offs among these principles in practice. The measures most within educator control may not be predictive of students' long-term outcomes (principles 1 and 2). Adjusting measures to be within the control of educators may make the measures more complicated and costly (principles 2, 4 and 5). Using outcomes that arise in the future, such as earnings may also increase measurement error because it can be harder to track these outcomes over long periods of time (principles 1 and 3).

To be clear, the above list of principles focuses on creating performance *measures* for purposes of accountability. It therefore omits principles regarding the design of the performance *incentives* that might be attached to those measures, which can range from providing public information to dismissing schools principals to closing schools. Also omitted from the list are principles of formative performance assessment, a related and equally important topic that is beyond the scope of the present analysis.

II.B. The Principles and State ESSA Plans

It is not obvious that the current state and federal accountability systems optimally balance these principles. Throughout a quarter-century of state and federal expansion and changes in test-based accountability, state and federal policies have violated principle 1 by focusing narrowly on student test scores even though other academic and school-age outcomes, particularly years of education are predictive of students' long-term outcomes (even after controlling for test scores).[5] High school graduation is now required as a performance measure at the high school level, and there is clear evidence that high school

[5] Test score levels are correlated with individual career success (Kuncel, Hezlett, & Ones, 2004) and macroeconomic growth (Hanushek & Woessman, 2012), but so are other outcomes.

graduation also meaningfully affects life outcomes (e.g., Levin et al., 2007).[6] Test scores and high school graduation also have the advantage of being perceived as primary goals (or proxies for goals) of schooling and, for that reason, they are widely measured. Equally strong cases can be made for school attendance (Finn, 1989; Halfors et al., 2002; Harlow, 2003; Rumberger, 1987) and college entry (Kane & Rouse, 1995; Goldin & Katz, 2008; Heckman, Humphries, & Veramendi, 2016), both of which strongly predict life success even after controlling for other outcomes.[7]

Principle 2 has also been consistently violated in modern accountability systems. The problem is that students start school with different levels on most outcomes (including test scores), implying that some schools would be expected to improve student outcomes far more than other schools (e.g., Kane & Staiger, 2002; Weiss, 2008; Harris, 2011; McEachin & Polikoff, 2012). This creates a "starting gate inequality" that rewards schools through higher performance measures because they serve more advantaged students and, perversely, punishes schools that serve students most in need.[8]

In this study, we consider ways of measuring school performance that are more in line with these principles. First, states can collect data on measures that have potentially higher predictive validity with regard to students' long-term success and include these as part of school performance measures. Second, states can shift some of the focus from outcomes levels to *value-added*, i.e., take into account students' predicted outcomes when

[6] High school graduation was added as a federal requirement years after NCLB passed, not because this was an important outcome, but to dissuade high schools from raising their scores by pushing out students who had low scores (Swanson, 2003).

[7] It is somewhat difficult to say whether test scores are more or less predictive because they are on different scales than college and other outcome measures. This can be addressed through measures of explanatory power such as the R^2, but we are not aware of such evidence for the measures that are commonly considered for accountability purposes. A partial exception is Murnane, Willett, & Levy (1995) who find that the return from a one standard deviation increase in (low-stakes) test scores is 1.3 percent per year and the return to a year of schooling is 3.7 percent per year (these estimates are from the same model). From Baird and Pane (2018), a one standard deviation increase in test scores annually, during the middle or early years of high school, is roughly 0.25 "years of schooling." Setting aside the problems with translating test gains in this manner (Baird & Pane, 2018), these results imply that the return to a year of schooling (3.7 percent) is slightly lower than the return to an equivalent change test scores (1.3/0.25=5.2 percent), but still substantial.

[8] The term "starting gate inequality" is borrowed from a report by Lee and Burkham (2002) on the related topic of inequalities at the start of kindergarten.

judging actual outcomes (Kane & Steiger, 2002; Harris, 2011). This has the effect of largely eliminating the starting gate inequality and focusing on what schools actually contribute (e.g., Chetty, Friedman & Rockoff, 2014).

In theory, the passage of the federal ESSA law freed up states to pursue these and/or other solutions. Our analysis shows that this has occurred, but only to a limited extent.[9] In reviewing states' ESSA plans, we find that the most common measure states are adding to their performance metrics is school attendance (Table 1).[10] While there is correlational research linking attendance to long-term outcomes (Finn, 1989; Halfors et al., 2002; Harlow, 2003; Rumberger, 1987), and attendance is easy to add because it is already widely measured, it is noteworthy that the strongest predictor of life outcomes—college enrollment[11]—is still largely omitted. Only three states mention college outcomes in their ESSA plans (Connecticut, Rhode Island, and Vermont). Table 1 shows which states are planning to include which measures.

While value-added measures are gradually gaining acceptance, our analysis shows that only 34 states are using value-added-like measures that address the starting gate inequality problem. Ten of these states, however, are also planning to use "growth-to-target" or "growth-to-proficiency." As with the original test levels in NCLB, researchers have pointed out that this alternative approach is quite different from measuring value-added (Weiss, 2008; Weiss & May, 2012). Growth-to-target yields results very similar to proficiency itself because students who are not on track are also those with low test score levels, recreating the problem of using levels alone (Harris, 2011). This fact, not being widely recognized, has led states to rely on growth-to-target, perhaps in the mistaken belief that it addresses the problem with test levels, or that is represents a sort of compromise

[9] For additional analyses of state ESSA plans see, for example: Martin, Sargrad, & Batel (2016) and Aldeman et al. (2017).
[10] The table also shows that progress for English language learners is a very common metric, though this represents achievement for a particular subgroup. It is not really a different outcome.
[11] This is generally measured in research as years of education. See Wolfe and Haveman (2002) for an extensive review of the positive individual and social (external) benefits of years of education.

between levels and growth. Also, note that, of the 24 states that are clearly using value-added (and not a mix of this with growth-to-proficiency), only 10 are weighing value-added more than 40 percent.

A related issue is that, with the additional measures states are considering, such as attendance, there has not been much consideration to applying value-added. Fundamentally, value-added calculations represent differences between actual and predicted outcomes, and essentially all student outcomes are somewhat predictable from past outcomes and student background. While the recent debate, and the discussion above, has been about using value-added with test scores, this can also be applied to high school graduation, attendance, college entry, and other measures.[12]

When we broaden our view to include all five principles, it is clear that there are trade-offs and that no measure is perfect. For example, at the extreme, we could hold schools accountable directly for students' adult outcomes, such as voter participation, employment, and incarceration, but these would not be proximal to teacher behavior. The longer we have to wait to observe outcomes, the more likely it is that these outcomes are outside the control of *current* educators. That is, the cardinal rule of accountability is not just about the role of non-school factors affecting outcomes, but about the timing of those outcomes and whether they can be measured while educators are still in the schools being held accountable.

In what follows, we measure the degree to which violating these principles is practically important by examining policy options that attempt to address these weaknesses. As we will see, college entry represents a type of middle ground—an outcome that narrowly occurs after K-12 schooling but which is still clearly within the control of

[12] The application of value-added to dichotomous measures is often called "risk-adjustment" and is applied increasingly in health care accountability (DesHarnais et al.,1998).

schools and predictive of long-term life outcomes.[13] We also address the reliability issues with value-added by averaging across years.

III. Data and Performance Indices

Most of the data used in the analysis were provided by the Louisiana Department of Education (LDOE) and include a panel of student-level data that tracks enrollment and achievement in all Louisiana publicly funded schools. The student-level data also provide other information about race, gender, grade level, free or reduced priced lunch status, special education status, and English language learner status. While performance measures such as test scores, high school graduation, and college entry are from 2010 to 2014 school years, we use prior years to obtain lagged test scores (8th grades test scores for high school analysis) for the estimation of equation (3).

State standardized tests (LEAP and iLEAP) are given in the spring to all students enrolled in grades 3-8. High school student, during the years in this analysis, were required to pass the Graduate Exit Exam (GEE) or End-of-Course tests (EOC) in order to graduate from high school.[14] All test scores are standardized by test, year, grade, and subject (math, English language arts (ELA), science, and social studies for grade 3-8, and math, ELA, and science for high school) within Louisiana to have a statewide mean of 0 and standard deviation (s.d.) of one.

We created the graduation indicator based on students' last exit codes. Students are coded as a "graduate" if they either exit or complete some type of degree or credential. The most common type of completion by far is graduation with a regular diploma, but we also

[13] In elementary schools, the equivalent would be to use performance in the first grade of middle school.
[14] The state of Louisiana was transitioning its testing system during the years of our analyses. Therefore, we use whichever tests are available in each year. Specifically, we use EOC math scores through 2011 to 2014 spring years, use GEE ELA scores for 2011 spring year and EOC ELA scores through 2012 to 2014 spring years, and use GEE science scores for 2011 and 2012 spring years and EOC science scores for 2013 and 2014 spring years.

include GED, certificate of achievement, or other forms of completion as these are included in Louisiana's accountability system.

Data on enrollment in college came from the National Student Clearinghouse (NSC). College entry is coded as one if students are found enrolled in any college and zero otherwise. The college data are only available for high school graduates. We assume that all non-graduates do not attend college. We restrict the high school analysis to schools with actual enrollment per grade more than 15 students. This is mostly to exclude alternative schools, which have different objectives and are often treated differently in accountability policies.

We use these data to create different types of performance measures. States usually create composite indices of the following general form:

$$P_s = \alpha_1 X_1 + \cdots + \alpha_N X_N \quad (1)$$

where P_s is the composite performance measure of school s. For simplicity, we assume that all the components of the composite measure $X_1, \ldots, X_N$ are on the same scale so that the weights sum to one ($\sum_n \alpha_n = 1$). If policymakers are attempting to maximize some index of students' long-term outcomes Y then these weights should be proportional to their contributions to Y.

States are required to give schools into performance ratings based on specific standards for the performance index.[15] We use Louisiana's method of assigning letter grades (A-F without E) for the performance ratings.

IV. The Effect of Adding Additional Outcomes

IV.A. Correlations and Diminishing Returns

[15] As explained in the notes to Table 1, some states do not use performance indices and instead use a series of decision rules, from which weights can sometimes be inferred.

We start by examining the issue of multiple measures/outcomes. Federal law requires that states use test scores for all schools. At the high school level, we also examine the effect of adding high school graduation to performance measures that already include test scores, as the federal government began requiring in 2003 (Swanson, 2003). We also go a step further and add college entry (Harris, 2015). Our analysis in this section focuses on high schools because our elementary/middle school data include only one measure that is commonly considered for accountability.

A key policy question is: What is the (marginal) contribution of adding measure X_N conditional on the vector of already included measures $\{X_1, \ldots, X_{N-1}\}$? The short answer is that this depends on the covariance matrix of the various measures, including both its dimension (the number of measures already included) and the magnitude of the covariance between each already included outcome and the new outcome. Adding a new outcome that is correlated with already included ones will generally add less information than one that is weakly correlated with the already included outcomes. By "adds less information," we mean it has a greater impact on the composite performance measure (and Y) than one that has a higher correlation.

As a general rule, student outcomes tend to be highly correlated with one another, though they are less so with intermediate outcomes. For example, the correlation between school-level attendance and achievement proficiency rates in Ohio is in the range of +0.54-0.78 depending on the grade (Roby, 2004). Student satisfaction and student achievement levels, on the other hand, are less correlated, in the range of +0.10-0.20 (Ostroff, 1992).[16] The weaker correlation with student satisfaction is most likely due to the fact that the two outcomes reflect quite different types of constructs (an opinion versus a skill). Survey measures also tend to be highly correlated with one another. (We do not unfortunately have

[16] The correlations in the Ostroff (1992) study are less comparable to those discuss elsewhere in this study because they do not average test scores across subjects. Averaging would likely increase the correlations somewhat.

access to either attendance or survey data in the analysis that follows, but, as we explain later, this does not affect the analysis much, as the point of the analysis is that the effect of adding any measure is driven by the correlations, not the specific measures).

The importance of the correlation between any two measures also depends on how many other measures are included. As a general rule, and as the prior and subsequent evidence shows, essentially all student outcomes are positively correlated, so that there are likely to be diminishing marginal returns to information.[17] The intuition behind this is easiest to see at the extremes: if we add another measure that is perfectly correlated with an already included measure, then there is no new information.

Figure 1 uses a simulated data set to show visually that there are diminishing marginal returns.[18] The y-axis shows the percentage of schools that receive the same performance rating when an additional measure is added. (Being at the very top of the y-axis therefore means that adding the measure has no practical impact.) The specific shape of the diminishing marginal returns depends on the correlation. When the correlations among all the potential measures are all +0.9, the percentage of schools receiving the same performance rating flattens out after the fourth measure is added.

While not obvious from the figures, it is important to point out that whenever a new measure is added, the weights on all the already included measures have to change. For this reason, even adding a measure that is perfectly correlated with another measure, though it would not add new information per se, would still change the performance index by re-weighting the components.

IV.B. Analysis of Correlations in Louisiana Data

[17] One can construct negative correlations of course (e.g., between dropout and test scores), but what we mean here is that measures for which more is better are positively correlated (dropout can be replaced with graduation, which is the positive version of dropout).

[18] For the simulation, we created multiple data sets of 1,000 (school) observations where the correlation among all the measures is exactly the same (and as specified in the legend at 0.3/0.5/0.7/0.9). Regardless of the number of measures (or correlations), all the measures are equally weighted (i.e., the weights always sum to unity).

Our analysis of the Louisiana data focuses on high school test scores (Test), high school graduation (HSGrad), and college enrollment levels (College). The correlations among these measures are all positive, but range in magnitude: Test-HSGrad (ρ =+0.57), Test-College (ρ =+0.62), and HSGrad-College (ρ =+0.73). These suggest that high school graduation and college outcomes levels are more closely linked with each other than either is with test scores. This may be because high school graduation is a prerequisite to attending college. Students can have low test scores and still graduate, but they cannot easily go to college without graduating high school. High school graduation and college entry are also proximal in time—they usually occur within months of each other, whereas the tests are taken in grades 9 and 10.[19]

The correlations are not, by themselves, very informative about the potential practical impact of adding measures on schools' performance ratings. Also, such correlations may be stronger or weaker at the extremes of the distribution, which are especially relevant in accountability systems that tend to punish very low performance and reward very high performance. Therefore, next, we report transition matrices that show how performance ratings would change, assuming that the share of schools with each of the ratings is held constant. We use the percentages in Louisiana, which, like most states, has a relatively small share of schools (eight percent) with the lowest rating of F. When these percentages are higher, the share of schools with the lowest and highest performance ratings will also tend to be higher, making the performance ratings more sensitive to small changes in performance measures.[20]

---

[19] We also estimated the correlations between the various measures when switching from levels to value-added. As expected, the correlations drop, mostly because of increased measurement error that comes with value-added adjustments. The Test-HSGrad and Test-College value-added correlations are cut more than in half with value-added (to +0.14 and +0.21, respectively). The HSGrad-College value-added correlation remains high, however, at +0.60. The relatively small drop in the HSGrad-College correlation suggests that the risk-adjustments for student demographics are quite similar for high school graduation and college entry (and these may be biased as well). These results are available upon request.

[20] This is because, if schools are normally distributed, a larger share of schools will be near the threshold, as the threshold moves closer to the mean.

The transition matrices in Table 2a show that when adding graduation to test scores, 62.1 percent of schools end up with the same performance rating, and only 2.5 percent change more than one letter grade. While high school graduation is an interesting outcome in and of itself, this is also a noteworthy case of adding a measure that has a moderate correlation with test scores (ρ =+0.57). The percent with the same rating in Table 2a is almost exactly as predicted by the simulation in Figure 1.

Next, we add college entry. In addition to predicting life outcomes, this measure is within the control of K-12 schools (Harris, 2015). High schools are responsible for preparing students for the academic demands of college and assist students in applying to colleges and for college financial aid. For the same reasons, using college *graduation* as a high school performance measure is questionable because the more time that has passed since students have left high school, the more likely it is that factors outside of high school control will drive student outcomes and bias school performance measures, violating principle 2.

Table 2b shows results comparing performance measures with test scores and high school graduation (1/2 weight for each of the two measures) with a measure that adds college entry (1/3 weight for each of the three). As shown in Table 2b, the effect of adding the third measure is nearly identical to adding the second one, with 71.4 percent of schools receiving the same grade (though no schools change by more than two letter grades). The size of this change from adding college entry may seem surprising given the higher correlation between high school graduation and college entry, and diminishing returns to information, but recall that the weights are also changing.

Given the weaker correlations reported above between, for example, survey-based measures and students' main academic outcomes, the estimates in this section suggest a higher degree of stability than we would expect from students' or parents' qualitative assessments of school performance, but we leave this for future research.

V. Value-Added Estimation

In the remainder of the study, we focus on the second issue with school performance measures: whether levels are adjusted using value-added methods. This section provides a brief introduction to value-added methods, which is followed by sections comparing levels versus value-added measures.

We estimate a variety of value-added models that are now standard in the research literature:

$$A_{ist} = f(A_{i,lag}) + \beta X_{ist} + \theta_{st} + \varepsilon_{ist} \quad (2)$$

where A_{ist} represents student achievement for student *i* in school *s* at time *t*, X_{ist} is a vector of student/family characteristics, and θ_{st} represents value-added of the test-taking school in year *t*. ε_{ist} is a random error term. For grades 4-8, the lagged test scores are the scores in the previous school year. For high schools, the lagged scores are the 8th grade LEAP test scores while A_{ist} is either the GEE or EOC exam depending on the year (see above).

We also estimate value-added-like measures for high school graduation and college entry. While graduation can only occur once, and therefore lacks a lagged value for individual students, the logic of value-added models is to calculate an expected outcome for each student and then compare the actual to the predicted outcome as a measure of school performance. We can therefore apply the same approach to high school graduation as well as college entry and estimate the following model,

$$O_{ist} = f(A_{is8}) + \beta X_{is8} + \delta_{st} + \omega_{ist} \quad (3)$$

where O_{ist} represents graduation or college entry indicators, A_{is8} represents student achievements in 8th grade, X_{is8} is the same vector of student/family characteristics as in equation (1), except focused on students' 8th grade information, and ω_{ist} is a random error term. δ_{st} represents school value-added. We apply a post-estimation shrinkage adjustment similar to that employed by Herrmann, Walsh, and Isenberg (2016). In some specifications, we also add a vector of school-level-aggregated version of the variable in X_{ist}. In these

cases, we also apply the two-step procedure recommended by Ehlert et al. (2014).[21] Note that we use fairly simple models here (e.g., OLS) because this is what would likely be used in state policy.

Many states use, and researchers advocate for, value-added measures that average across years in order to reduce their inherent statistical unreliability (e.g., Harris, 2011, 2015). We follow suit and average across four years.[22] While individual schools are affected by averaging, the overall patterns are not sensitive to averaging over time. It is the shift to value-added, and accounting for prior achievement, that leads school performance ratings to change.

VI. The Effect of Switching from Levels-Only Toward Value-Added

Below, we report the results from the estimation of equations (2) and (3) with only student-level covariates, applying shrinkage adjustments, and (usually) averaging over four years. We also carry out robustness checks involving the addition of school-level covariates and data averaged across multiple years.

VI.A. Correlation between Elementary/Middle School Test Score Levels and Value-Added

[21] Ehlert et al. (2014) argue that the standard value-added model cannot separately identify the school fixed effects and the effects of school-aggregated student characteristics, thus potentially under-correcting for the influence of school-level characteristics. A two-step model corrects for it. The first step partials out the influence of lagged test scores, student characteristics, and schooling environment controls. The second step regresses residuals from the first step on school dummies to yield school value-added measures.

[22] For consistency, we also average the outcome levels across four years. However, this has almost no influence on the results given the much higher reliability of test levels.

From the above analysis, we hypothesize that the influence of adding more weight to value-added versus outcome levels should depend substantially on how correlated the two measures are. That is, from a statistical standpoint, adding a value-added measure has the same effect on the performance measures as adding, for example, college entry, as long as the covariance matrix is the same.

We therefore start by plotting test levels and value-added measures for all schools in Louisiana. The x-axis represents the value-added measure. The y-axis represents the *school-level* standard deviations of the levels measure. (While using the school-level standard deviation is unusual it allows us to put all the measures on the same unit of measure[23] across the test score, graduation, and college analyses. See later discussion of Table 5.) Each figure has four quadrants: the lower-left and upper-right quadrants place schools in the same half of the distribution regardless of which measure is used. The schools in the upper-left quadrant are unfairly rewarded with high ratings but low value-added while the schools in the lower-right are unfairly punished with low ratings but high value-added. The scatterplot in Figure 2 illustrates this relationship visually. There is a clear positive relationship between test levels and value-added for elementary/middle schools. The linear projection highlights the slope of that relationship. The correlation is +0.85.

The scatterplots reflect a complete shift from levels to value-added. However, this approach has two disadvantages. First, researchers who have studied value-added do not advocate for switching entirely to value-added.[24] Second, it makes it difficult to compare the results for value-added with those in Section IV where we added high school

[23] We recognize that this does not allow us to put test scores, graduation, and college entry on the "same scale." By standardizing all of them to the school s.d., however, a one-unit move along the vertical axis moves schools the same amount in the school-level distribution. It is a relative measure.

[24] One reason for mixing the two includes the fact that testing does not start until 3rd grade, so a focus on value-added would create a perverse incentive of reducing 3rd grade scores in a way that increases subsequent growth. Also, to the extent that families use school performance measures to make schooling choices, it is rational for them to look partly at test levels as signals of peer influences. See (Harris, 2011) for more on this issue.

graduation with a 1/2 weight to high school graduation and 1/2 to test scores, and adding college entry with a 1/3 weight to college entry, test scores, and high school graduation. The transition matrices in Table 3 avoid these problems, showing a shift from levels-only to a 50-50 mix of levels and value-added.

The upper-left cell of Table 3 provides the percentage of elementary/middle schools that receive F grades using both levels and value-added, and the remaining diagonals do the same for the other letter grades. The off-diagonals show the schools that are affected by the switch to value-added. For example, 75.8% of schools maintain the same letter grade while only 0.1% change by two letter grades.

VI.B. Correlation between High School Test Scores and Value-Added

At the high school level, the correlation between test levels and test value-added drops to +0.68, and the relationship is visually looser in the scatterplot (Figure 3). This also translates into a drop in the number of schools with the same grade, to 67.1% (Table 4a). This is probably mostly because the testing regime in high school is different from that used in middle school where the 8th grade scores, used as a covariate in the value-added model, comes from.

VI.C. Correlation between Levels and Value-Added for High School Graduation and College Entry

Part of the contribution of this article is highlighting the potential of applying value-added to outcomes beyond test scores. We extend here the idea to high school graduation and college entry, but the same logic applies to essentially any student outcome.

The scatterplots comparing levels and value-added for high school graduation and college entry are shown in Figures 4 and 5, respectively. The correlation between levels and value-added increases to +0.90 for high school graduation and +0.80 for college entry. The slopes of the relationships between levels and value-added are nearly identical across the three outcomes (test scores, high school graduation, and college entry). As with value-

added to test scores, these results are averaged across four years, though averaging has limited impact on the results.

Tables 4b and 4c show the equivalent results with transition matrices. Switching from high school graduation levels to value-added, 77.9 percent of schools remain in the same category. The equivalent number is 69.3 percent when switching from college entry levels to college entry value-added yields.

In most respects, the results here are unsurprising, but they are still enlightening given how little attention has been paid to broadening the application of value-added. It is noteworthy that switching to value-added for a given set of student outcomes seems to have almost as large an impact on school performance ratings as changing the outcome measures themselves. Comparing Tables 2a-2b (multiple measures) with Tables 4a-4c (value-added), in particular, we see that performance categories are affected in quite similar ways when adding value-added, which adjusts a given set of outcomes, as by adding high school graduation or college entry, which are entirely different student outcomes.

VII. Policy Simulations for Switching from Levels to Value-Added

A key difference between adding value-added and adding additional student outcomes like college entry is that the former is more a matter of statistics and the latter is more about values. That is, whether we add college entry depends in part on what long-term outcomes policymakers value most (college is entry is related to almost all of them). With value-added, in contrast, policymakers have already picked the outcomes (e.g., test scores), and the question is whether they support the principle that performance measures should focus on what educators can control. Research suggests that value-added measures are less biased measures of contributions to student outcomes, at least with regard to

student test scores (Chetty et al., 2014; Guarino et al., 2015).[25] Further, since the measures themselves have consequences for schools, adding value-added measures to school performance indices could improve school performance. The remainder of this section attempts to quantify that impact.

VII.A. How Does the Use of Value-Added Matter with Single Measures?

To provide a sense of the potential impact, we carry out a policy simulation focusing on one mechanism through which better performance measures could improve schools. Specifically, we simulate a policy of taking over the bottom five percent of schools every year, for each of four years, and replacing that school with another. A similar policy (without explicit percentages) has been implemented in New Orleans in recent years (Bross, Harris, & Liu, 2016). This policy is also similar to a fully implemented version of NCLB or ESSA, both of which emphasized state intervention in low-performing schools. The idea is also reinforced by evidence that test-based accountability seems to have a greater impact on low-performing students (Deming & Figlio, 2016).

The transition matrices discussed above provide some indication of what would happen. The top of Table 3, as discussed above, shows that some of the eight percent of schools that are in the F category based on test levels (sufficient for state intervention in Louisiana) are in the B and D categories under value-added, and would therefore not be subject to intervention. Instead, some of the schools initially receiving higher ratings would experience intervention. This same logic can be seen in Figures 2-5 where the observations in the upper-left box are those that have higher-than-average test scores, but low value-added. These schools might benefit from intervention, but the accountability system instead rewards them for their high outcomes levels.

In any study of closure and takeover, it is important to consider the quality of the replacement schools. In New Orleans, for example, the replacement schools from the city's

[25] The studies cited here focus on teacher value-added. There is far less evidence on school value-added, but see: Kane and Staiger (2002) and Ladd and Walsh (2002).

charter authorization process tended to be near the average. So, one option would be to assume that the replacement schools have value-added at the 50th percentile (as of the year the policy starts).[26] We chose to replace schools with the median because this is how the policy played out in New Orleans (Bross, Harris, & Liu, 2016). This is also more reasonable that it might seem because a more common policy is to close schools entirely and move students to other existing schools. That is, it might not be realistic for a takeover school to be immediately as effective as the districtwide median, but it may be realistic to assume that students move to schools with average value-added for the district. Naturally, lowering the choice of replacement school attenuates the effect of adding value-added, though perhaps to a lesser extent than expected given that the same assumption about the replacement school's performance percentile has to be made when using both levels-only and the levels/value-added mix.

The simulation shows that taking over the bottom five percent of elementary/middle schools on the basis of a 50-50 mixture of test score levels and value-added, as opposed to the more standard levels-only measure, would increase the average performance of the original bottom 20 percent of schools by 0.015 student-level standard deviations (see Table 5 Panel A). The number might seem modest, but note that this change is accomplished only by changing the measurement and represents only a single year of achievement growth, setting aside the potential accumulation of gains across years. (Note that we focus on the bottom 20 percent of schools because these are the only schools directly affected by the policy.)

Table 5 Panel B provides simulations of the same policy for high schools. To simplify matters, with three different outcome measures, we start by showing, in the first three columns, how value-added would change when switching from levels-only to the

[26] In some earlier versions of this work, we also considered that the distribution of school value-added may not be smooth and therefore we ran the simulation using the 45th, 50th and 55th percentile and average the three together. However, this had essentially no impact, and this is why we report only the results based on the median.

levels/value-added mix under an accountability system that only uses each outcome measure separately. The first column, for example, tells us how value-added to high school test scores in the bottom quintile would be affected by the switch from levels to levels/value-added mix if test scores were the only outcome measure. The answer is similar to what we found with elementary school test scores as achievement levels improve by +0.029 student-level s.d. The effects on high school graduation and college entry are +0.3 and +0.9 percentage points, respectively. When we switch to school-level s.d., so that we can compare the test score results to high school graduation and college entry, the effects appear largest for test scores (+0.076 school s.d.), followed by college entry (+0.070 school-level s.d.) and high school graduation (+0.026 school-level s.d.). Panel C of Table 5 shows the effect of switching from single-year to four-year averages, but these changes are small and work in no clear direction.

VII.B. How Does the Use of Value-Added Matter with Multiple Measures?

As noted earlier, it is unrealistic to have performance measures with only a single student outcome like test scores, especially in the era of ESSA that requires multiple measures at all grade levels. The fourth column in Table 5 Panel B applies an equal weight to test scores and graduation rates. Since each outcome is included as both levels and value-added (also equally weighted), this means the fourth column represents four measures: test levels, test value-added, high school graduation level, and graduation value-added (each equally weighted).

The effect of switching from levels-only of test scores and graduation to a levels/value-added mix reduces the effect on average value-added to test scores (from 0.029 to 0.013), as expected given that the weight attached to the test score value-added is lower when graduation is also part of the mix. Interestingly, however, value-added to high school graduation actually *increases* slightly when test scores are part of the mix (compare columns 2 and 4). This is possible for two reasons. First, the high correlation between high school graduation levels and graduation value-added (0.9) means that giving more weight

to value-added, when graduation is the only student outcome, does not have an especially large effect (column 2). The second reason is that, as emphasized earlier, the effect of adding any one measure depends on which measures we start with and on the entire covariance matrix. In the case of column 4, switching to value-added means switching to value-added for both graduation and test scores. Moreover, it turns out that test value-added is more highly correlated with graduation value-added (the correlation is 0.29) than the latter is with test levels (the correlation is 0.26). This is not the case with the other measures and has the effect of counteracting the more intuitive effect of reducing the weight on graduation value-added.[27] When switching to a system with all three outcomes and mixing levels and value-added, so that there are now six equally weighted measures,[28] the impact on each outcome either remains steady or continues to decline.

This section estimates and illustrates one of the increasingly common mechanisms through which performance measures affect actual student outcomes. While these effects might seem small, note that they reflect only one of the three mechanisms, excluding internal school improvement and school choice processes, and these effects come at essentially no cost.

VIII. Conclusion

In this study, we propose five principles for creating school performance measures. While most of these have been considered individually in prior research, the first two principles—the focus on predicting long-term life success and on what educators can control—are not often considered. We also discuss some of the trade-offs among the principles.

[27] In contrast, when we compare columns 1 and 4, the effect on test score value-added drops; this is because the correlation between graduation levels and test score value-added (0.38) is higher than the correlation between graduation value-added and test score value-added (0.29).
[28] The six measures are: test score levels, test score value-added, high school graduation levels, high school graduation value-added, college entry levels, and college entry value-added.

Our analysis shows that there is still considerable room for progress in how we measure school performance. State ESSA plans are still mostly inconsistent with the first two principles. The addition of school attendance, being planned by most states, will likely help make the performance measures more predictive of students' long-term life outcomes (Finn, 1989; Halfors et al., 2002; Harlow, 2003; Rumberger, 1987), though the weights attached to this new measure are small.

We examine two ways to better align performance measures with the first two principles that most states are not relying on. First, we find that adding college entry to high school performance measures would change the performance categories of about one-quarter of schools. Second, switching from outcome levels to an even mix of levels and value-added would increase student achievement, among the bottom quintile of schools, by about 0.4 percentiles and high school graduation and college entry by about 0.4 percentage points.

While the size of these effects might seem small, it is important to note, first, that the simulation is only designed to capture effects through the opening and closing of schools. Schools in the higher quintiles are also likely to respond given that families seem to prefer schools with stronger academic outcomes (e.g., Glazerman & Dotter, 2016). This is especially true among higher-income families whose children tend to attend high-performing schools (Harris & Larsen, 2017). Perhaps the most important point is that the switch to value-added is essentially costless in an economic sense. It does not require any new data collection, only statistical adjustments in the existing ones. That is, switching to value-added measures can, by itself, improve average school performance.

Future research should do more to examine the correlations among a wider variety of measures that are being considered for accountability and to estimate the degree to which the measures predict students' long-term life outcomes. This evidence will help move the field and practice toward an "optimal mix" of performance measures that accounts for all the various principles. As we show, all these measures are correlated with

one another, so changing the mix of outcomes, especially when this means adding many outcomes, may not have much practical impact.

As with everything else in education policy, policy design and implementation matter in accountability. These results may inform states as they move forward with their renewed flexibility under ESSA. School accountability, and the performance measures they rest on, have the potential to facilitate school improvement, but also to undermine it. All of the various principles of performance measurement, and their practical consequences, need to be considered.

References

Aldeman, C., Hyslop, A., Marchitello, M., Schiess, J.O., & Pennington, K. (2017). *An Independent Review of ESSA State Plans*. Washington, DC: Bellwether Education Partners.

American Education Research Association (2014). *Standards for Educational & Psychological Testing*.

Ariely, D. (2000). Controlling the Information Flow: Effects on Consumers' Decision Making and Preferences. *Journal of Consumer Research* 27(2): 233–248.

Baird, M.D. & Pane, J.F. (2018). *Translating Standardized Effects of Education Programs into More Interpretable Metrics*. Santa Monica: RAND.

Black, S.E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *Quarterly Journal of Economics* 114(2): 577–99.

Bowman, J.S. (1999). Performance Appraisal: Verisimilitude trumps veracity. *Public Personnel Management* 28(4): 557-576.

Brighouse, H., Ladd, H., Loeb, S. & Swift, A. (2016). *Educational Goods: Values, Evidence, and Decision-Making*. Chicago University of Chicago Press.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9): 2593-2632.

Deck, C. & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review* 78: 97-119.

Dee, T.S. & Jacob, B. (2011). The Impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management* 30(3), 418-446,

Deming, D. & Figlio, D. (2016). Accountability in US Education: Applying Lessons from K–12 Experience to Higher Education. *Journal of Economic Perspectives* 30(3): 33–56.

DeNisi, A.S. & Pritchard, R.D. (2006). Performance Appraisal, Performance Management and Improving Individual Performance: A Motivational Framework *Management and Organization Review* 2(2): 253–277.

DesHarnais, S.I., Chesney, J.D., Wroblewski, R.T., Fleming, S.T., & McMahon, L.F. (1998). The Risk-Adjusted Mortality Index: A New Measure of Hospital Performance. *Medical Care* 26(12): 1129-1148

Ehlert, M., Koedel, C., Parsons, E. and Podgursky, M.J., 2014. The sensitivity of value-added estimates to specification adjustments: Evidence from school-and teacher-level models in Missouri. *Statistics and Public Policy*, *1*(1), pp.19-27.

Figlio, D.N. & Lucas, M.E. (2004). What's in a Grade? School Report Cards and the Housing Market. *American Economic Review* 94(3): 591-604.

Finn, J.D. (1989). Withdrawing from school. *Review of Educational Research* 59,117-142.

Glazerman, S. & Dotter, D. (2017). Market Signals: Evidence on the Determinants and Consequences of School Choice From a Citywide Lottery. *Educational Evaluation and Policy Analysis* 39(4): 593-619.

Goldin, C. & Katz, L. (2008). *The Race Between Education and Technology*. Cambridge, MA: Harvard University Press.

Guarino, C.M., Reckase, M.D., & Wooldridge, J.M. (2015). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy* 10(1): 117-156.

Hallfors, D., Vevea, J. L., Iritani, B., Cho, H., Khatapoush, S. & Saxe, L. (2002). Truancy, grade point average, and sexual activity: A meta-analysis of risk indicators for youth substance use. *Journal of School Health* 72: 205-211.

Hanushek, E.A. & Woessman, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth* 17:267–321.

Harlow, C. (2003). *Education and correctional populations*. Bureau of Justice Statistics Special Report. Washington, DC: U.S. Department of Justice.

Harris, D.N. (2011). *Value-Added Measures in Education*. Cambridge, MA: Harvard Education Press.

Harris, D.N. (2015). *Recommendations to Improve the Louisiana System of Accountability for Teachers, Leaders, Schools, and Districts: Second Report to Louisiana Accountability Commission*.

Harris D.N. & Larsen, M. (2017). *How Schools Do Families Want (and Why)?* New Orleans, LA: Education Research Alliance for New Orleans.

Hart, O., and B. Holmstrom (1987): The Theory of Contracts, in T.F. Bewley (ed.), *Advances in Economic Theory: Fifth World Congress of the Econometric Society*, 71-155, Cambridge University Press: Cambridge UK.

Heckman, J.J., Humphries, J.E., & Veramendi, G. (2016). Returns to education: The causal effects of education on earnings, health, and smoking. *NBER Working Paper 22291*. Cambridge, MA: National Bureau of Economic Research.

Heiner, R.A. (1983). The origin of predictable behavior. *American Economic Review* 73, 560–595.

Herrmann, M., Walsh, E. and Isenberg, E., 2016. Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy*, *3*(1), pp.1-10.

Jennings, J. & Rentner, D.S. (2006). Ten Big Effects of No Child Left Behind. *Phi Delta Kappan* 88(2): 110-113.

Kane, T., & Rouse, C. (1995). Labor-Market Returns to Two- and Four-Year College. *American Economic Review* 85(3): 600-614.

Kane, T. & Staiger, D. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives* 16(4): 91-114.

Keller, K.L. & Staelin, R. (1987). Effects of Quality and Quantity of Information on Decision Effectiveness. *Journal of Consumer Research* 14(2): 200–213.

Klein, A. (2016). Issues A-Z: The Every Student Succeeds Act: An ESSA Overview. *Education Week*. Retrieved April 2, 2018 from http://www.edweek.org/ew/issues/every-student-succeeds-act/.

Koretz, D. (2017). *The Testing Charade*. University of Chicago Press.

Kuncel, N.R., Hezlett, S.A., & Ones, D.S. (2004) Academic performance, career potential, creativity, and job performance: can one construct predict them all? *Journal of Personality and Social Psychology* 86(1):148-161.

Ladd, H.F. & Walsh, R.P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review* 21: 1–17.

Lee, V. and Burkham, D. (2002). *Inequalities at the Starting Gate*. Washington, DC: Economic Policy Institute.

Levin, H. M., Belfield, C. Muennig, P.A., & Rouse C. (2007). *The Costs and Benefits of an Excellent Education for All of America's Children*. Columbia University Academic Commons.

Martin, C., Sargrad, S., & Batel, S. (2016). *Making the Grade: A 50-State Analysis of School Accountability Systems*. Washington, DC: Center for American Progress.

McEachin, A. & Polikoff, M. (2012). We are the 5 Percent: Which Schools would be Held Accountable Under a Proposed Revision of the Elementary and Secondary Schools Act. *Educational Researcher* 41(7): 243-251.

Murnane, R.J., Willett, J.B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* 77(2): 251-266.

Ostroff, C. The relationship between satisfaction, attitudes, and performance: An organizational level analysis. *Journal of Applied Psychology* 77(6): 963-974.

Roby, D.E. (2004) Research on School Attendance and Student Achievement: A Study of Ohio Schools. *Educational Research Quarterly* 28(1): 3-14.

Rumberger, R.W. (1987). High school dropouts: a review of issues and evidence. *Review of Educational Research*, 57: 101-121.

Schmidt, W.H., Wang, H.C. & McKnight, C.C. (2005). Curriculum coherence: an examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies* 37(5): 525-559.

Swanson, C.B. (2003). *Keeping Count and Losing Count: Calculating Graduation Rates for all Students Under NCLB Accountability*. Washington D.C.: The Urban Institute Education Policy Center.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction* 4(4): 295-312.

Tversky, A. & Shafir, E. (1992) Choice under conflict: The dynamics of deferred decision. *Psychological Science* 3, 358-361.

Weiss, M.J. (2008). *Examining the Measures Used in the Federal Growth Model Pilot Program*. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, D.C., March 3, 2008.

Weiss, M. J. and May, H. (2012). A Policy Analysis of the Federal Growth Model Pilot Program's Measures of School Performance: The Florida Case. *Education Finance and Policy* 7(1): 44–73.

Wolfe, B.L. & Haveman, R.H. (2002) *Social and Nonmarket benefits from education in an advanced economy*. Boston Federal Reserve Conference Series.

Figure 1
Diminishing Marginal Returns to Additional Measures (Simulation)

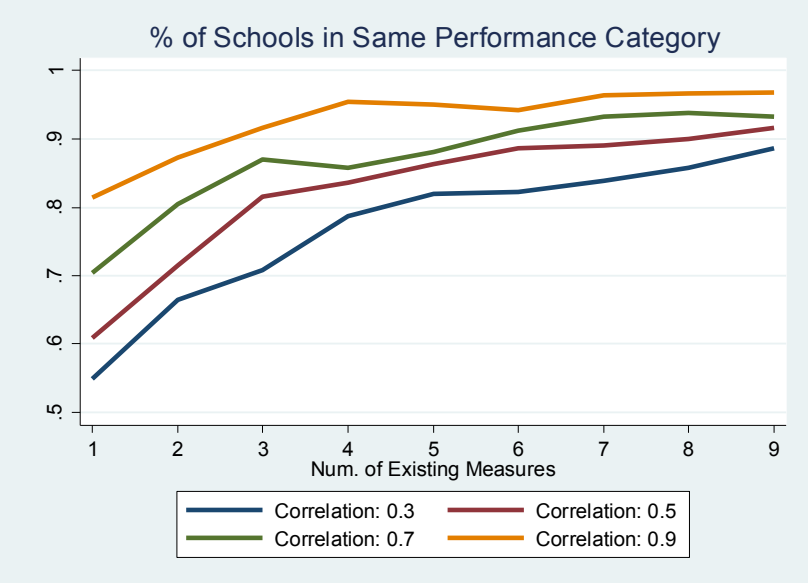
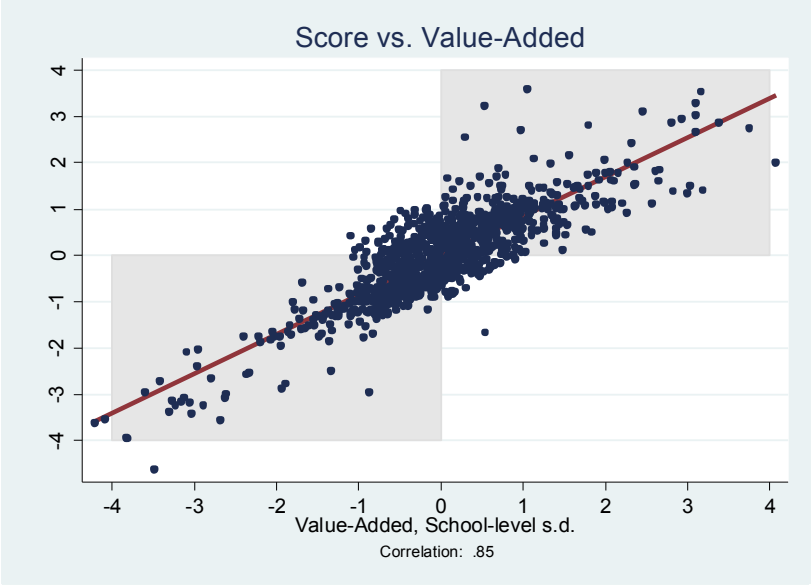
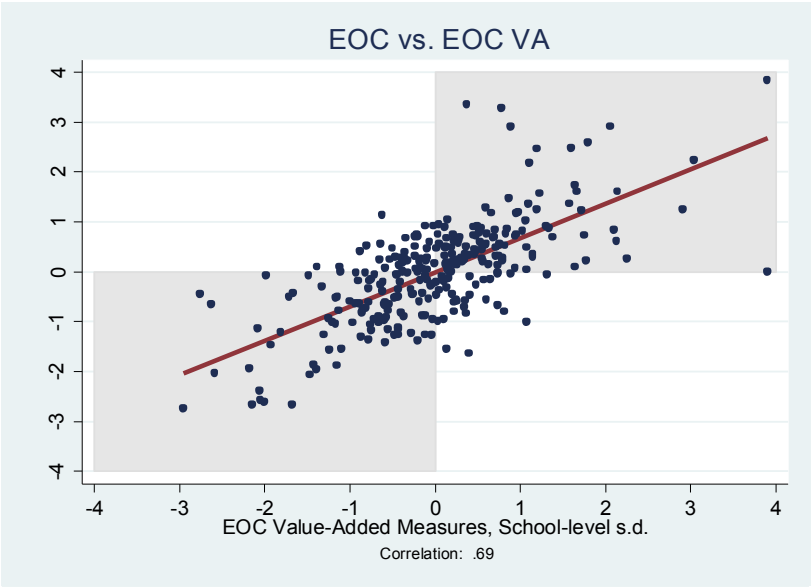


Figure 2
Scatterplots of Levels and Value-Added: *Elem/Middle Test Scores (LA)*



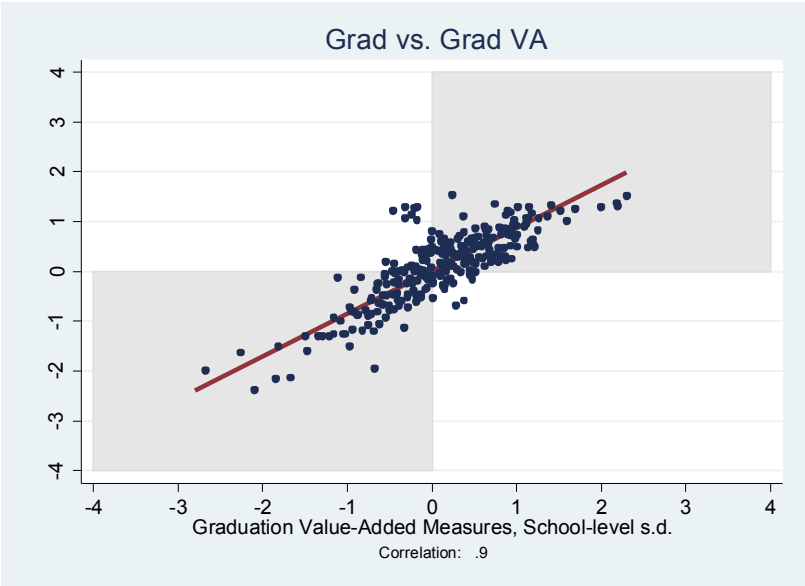
Notes: The scatterplot in Figure 2 shows the correlation between four-year averages of school-level test levels (y-axis) and school value-added (x-axis) for Louisiana elementary/middle schools. The correlation is listed at the bottom of the figure.

Figure 3
Scatterplots of Levels and Value-Added: *High School Test Scores (LA)*



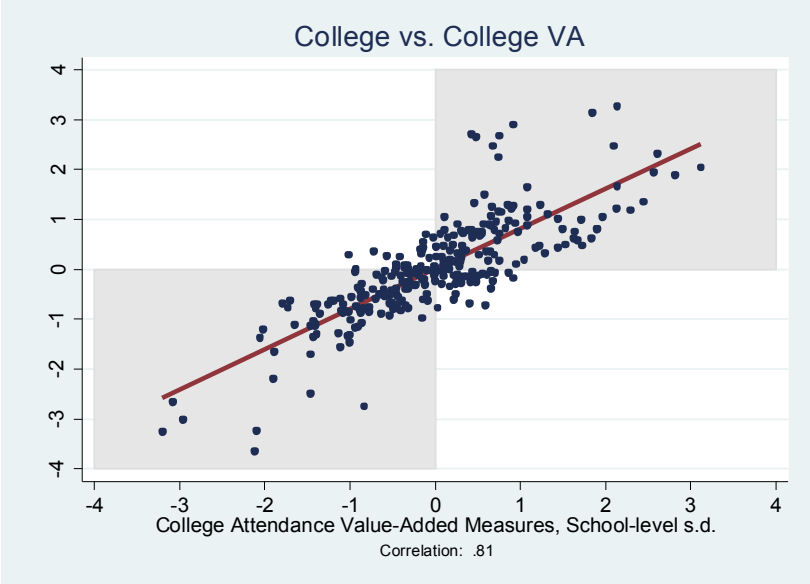
Notes: The scatterplot in Figure 3 shows the correlation between a four-year average of school value-added (x-axis) and average test levels for Louisiana high schools. The correlation is listed at the bottom of the figure.

Figure 4
Scatterplots of Levels and Value-Added: *Graduation (LA)*



Notes: The scatterplot in Figure 4 shows the correlation between a four-year average of graduation value-added (x-axis) and average graduation levels for Louisiana high schools. The correlation is listed at the bottom of the figure.

Figure 5
Scatterplots of Levels and Value-Added: *College Entry (LA)*



Notes: The scatterplot in Figure 5 shows the correlation between a four-year average of college entry value-added (x-axis) and average college entry levels for Louisiana high schools. The correlation is listed at the bottom of the figure.

Table 1: Summary of State ESSA Plans

	VAM or SGP	Growth to Target	Test Score Level Weight (%)	Growth Weight (%)	Outcomes (other than math and reading test scores)	Test Score Level Weight (%)	Growth Weight (%)	Grad Level Weight (%)	Outcomes (other than math and reading test scores and graduation)
Alabama	-	-	40%	-	ELL, attendance	20%	25%	30%	ELL, attendance, college & career ready
Alaska	-	-	36%	-	ELL, attendance	40%	-	30%	ELL, attendance, freshmen on track
Arizona	SGP	Y	30%	50%	ELL, acceleration/readiness	30%	20%	20%	ELL, college & career ready
Arkansas	SGP	-	35%	50%	ELL	35%	35%	15%	-
California	-	Y	-	-	ELL, attendance, suspension	-	-	-	ELL, suspensions
Colorado	SGP	Y-ELL	23%	40%	ELL, science ach., attendance	20%	20%	15%	Science, dropout
Connecticut	-	Y	31%	41%	Attendance, HS grad on-track, phys. fitness	52%	28%	15%	Attendance, college entry
Delaware	-	Y	30%	40%	ELL	40%	-	15%	ELL
Florida	VAM	-	25%	51%	Science achievement	20%	40%	10%	Science & SS, acceleration
Georgia	SGP	Y	30%	31%	ELL, Beyond the Core, Science & SS	20%	27%	15%	ELL, science & SS
Hawaii	SGP	Y-ELL	40%	40%	ELL, attendance	30%	-	50%	ELL, attendance
Idaho	-	Y	36%	36%	ELL, student satisfaction survey	45%	-	23%	ELL progress, college & career readiness
Illinois	SGP	Y	20%	50%	ELL, attendance, school climate	20%	-	50%	ELL, attendance, climate surveys
Indiana	SGP	Y	43%	43%	ELL, attendance	15%	15%	30%	ELL, college & career readiness
Iowa	SGP	Y	28%	47%	Test participation, conditions for learning	20%	40%	15%	ELL, test participation, learning conditions
Kansas	-	-	-	-	ELL	-	-	-	ELL
Kentucky	-	Y	20%	25%	Ach. gap, transition readiness, opport./access	15%	-	10%	Gap closing, transition readiness
Louisiana	VAM	-	50%	25%	ELL measure	21%	-	42%	ELL
Maine	VAM	-	42%	38%	ELL, attendance	40%	-	40%	ELL, attendance
Maryland	SGP	Y-ELL	20%	25%	ELL, attendance, school climate	30%	-	15%	Attendance, college readiness
Massachusetts	SGP	-	40%	25%	ELL achievement	33%	20%	6%	ELL
Michigan	SGP	Y	-	-	-	-	-	-	-
Minnesota	-	-	-	-	Attendance	-	-	-	Attendance
Mississippi	-	-	34%	29%	ELL	15%	-	20%	ELL
Missouri	-	Y	40%	30%	ELL, attendance	-	-		
Montana	-	Y	25%	30%	ELL, attendance	30%	-	25%	ELL
Nebraska	-	Y	-	-	-	-	-		
Nevada	SGP	Y	25%	20%	ELL, attendance, school climate	25%	-	30%	ELL
New Jersey	SGP	-	30%	40%	ELL, attendance	30%	-	40%	ELL

New Hamp.	SGP	-	-	-	-	-	-	-	-
New Mexico	VAM	-	33%	42%	ELL	25%	30%	13%	ELL
New York	SGP	-	-	-	-	-	-	-	-
North Carolina	VAM	-	80%	20%	-	20%	-	-	ELL
North Dakota	-	Y	30%	30%	ELL, climate, engagement	25%	21%	16%	GED Completion
Ohio	VAM	-	20%	20%	Gap closing, literacy improvement	20%	20%	15%	Attendance, discipline
Oklahoma	VAM	Y-ELL	35%	30%	ELL, attendance	45%	10%	10%	ELL
Oregon	SGP		22%	44%	ELL, attendance	-	0%	-	ELL
Pennsylvania	VAM	Y-ELL	-	-	ELL, attendance, career readiness	-	-	-	ELL
Rhode Island	SGP	Y-ELL	-	-	ELL	-	-	-	ELL
South Carolina	VAM	Y-ELL	40%	40%	ELL, science & SS, learning environ.	28%	0%	30%	ELL
South Dakota	SGP	Y-ELL	40%	40%	ELL	40%	-	13%	College & career ready
Tennessee	VAM	-	45%	35%	ELL, attendance	30%	25%	5%	Graduation ready (similar to college/career)
Texas	VAM	-	-	-	ELL	-	-	-	ELL
Utah	SGP	Y	33%	25%	ELL	25%	-	33%	ELL
Vermont	SGP	Y-ELL	90%	-	ELL	50%	-	20%	ELL
Virginia	VAM	Y-ELL	-	-	ELL	-	-	-	ELL
Washington	SGP	-	40%	50%	ELL, attendance	30%	0%	50%	ELL
West Virginia	-	Y	71%	-	Attendance, behavior	78%	-	-	Attendance, behavior
Wisconsin	SGP	Y	40%	40%	ELL, attendance	40%	-	40%	ELL
Wyoming	SGP	Y	25%	25%	ELL, equity	20%	20%	20%	ELL

Notes: Source: Authors' analysis of state ESSA plans with additional internet searches and corroboration with other public summaries. SGP = Student Growth Percentile, a form of value-added. In the Growth to Target column, Y means that the ESSA proposal uses these or very similar words. Y-ELL means the term is used only in reference to English Language Learners. In many states, the nature of the growth/value-added measure was unclear and we carried out additional searches. In other states where the weights are missing, such as New York, the application of the performance measures is through a set of decision rules rather than a composite index with weights. While this is a reasonable approach, it is difficult to characterize the weight attached to the various measures. The outcomes other than test scores and graduation rates are sometimes vaguely worded, but we use the words from the ESSA plan with minor modification (e.g., many states refer to "chronic absenteeism," which we reduce to "attendance"). A dash (-) indicates that the information is missing or unclear, which is common in these proposals. See other notes for specific states:

Table 2a:
Transition Matrix: Average Scores versus Average Scores and Graduation

Letter Grade of Average Scores	Letter Grade of Average Scores & Grad					
	F	D	C	B	A	Total
F	5.4%	2.5%	0.0%	0.0%	0.0%	7.9%
D	1.8%	10.4%	5.7%	0.7%	0.0%	18.6%
C	0.0%	4.6%	16.4%	6.4%	0.0%	27.5%
B	0.7%	0.7%	5.4%	17.1%	4.6%	28.6%
A	0.0%	0.4%	0.0%	4.3%	12.9%	17.5%
Total	7.9%	18.6%	27.5%	28.6%	17.5%	100.0%

Table 2b:
Transition Matrix: Average Scores and Graduation versus Average Scores, Graduation, and College

Letter Grade of Average Scores & Grad	Letter Grade of Average Scores & Grad & College					
	F	D	C	B	A	Total
F	6.4%	1.4%	0.0%	0.0%	0.0%	7.9%
D	1.4%	13.9%	3.2%	0.0%	0.0%	18.6%
C	0.0%	3.2%	18.6%	5.7%	0.0%	27.5%
B	0.0%	0.0%	5.7%	18.9%	3.9%	28.6%
A	0.0%	0.0%	0.0%	3.9%	13.6%	17.5%
Total	7.9%	18.6%	27.5%	28.6%	17.5%	100.0%

Table 3:
Transition Matrix: Elementary/Middle School Test Levels-Only versus Half Levels and Half Value-Added

Letter Grade of Levels	Letter Grade of Half Levels and Half Value-Added					
	F	D	C	B	A	Total
F	6.8%	1.2%	0.0%	0.0%	0.0%	8.0%
D	1.2%	14.4%	3.0%	0.0%	0.0%	18.5%
C	0.0%	3.0%	20.1%	4.4%	0.1%	27.6%
B	0.0%	0.0%	4.5%	20.6%	3.4%	28.5%
A	0.0%	0.0%	0.0%	3.5%	13.9%	17.4%
Total	8.0%	18.5%	27.6%	28.5%	17.4%	100.0%

Table 4a:
Transition Matrix: High School Test Levels versus Half Levels and Half Value-Added

Letter Grade of Levels	Letter Grade of Half Levels and Half Value-Added					
	F	D	C	B	A	Total
F	5.4%	2.5%	0.0%	0.0%	0.0%	7.9%
D	1.8%	12.1%	4.6%	0.0%	0.0%	18.6%
C	0.7%	3.9%	17.1%	5.0%	0.7%	27.5%
B	0.0%	0.0%	5.7%	19.3%	3.6%	28.6%
A	0.0%	0.0%	0.0%	4.3%	13.2%	17.5%
Total	7.9%	18.6%	27.5%	28.6%	17.5%	100.0%

Table 4b:
Transition Matrix: High School Graduation Levels versus Half Levels and Half Value-Added

Letter Grade of Levels	Letter Grade of Half Levels and Half Value-Added					
	F	D	C	B	A	Total
F	7.9%	0.0%	0.0%	0.0%	0.0%	7.9%
D	0.0%	15.7%	2.9%	0.0%	0.0%	18.6%
C	0.0%	2.9%	20.4%	4.3%	0.0%	27.5%
B	0.0%	0.0%	4.3%	20.4%	3.9%	28.6%
A	0.0%	0.0%	0.0%	3.9%	13.6%	17.5%
Total	7.9%	18.6%	27.5%	28.6%	17.5%	100.0%

Table 4c:
Transition Matrix: Levels versus Half Levels and Half Value-Added College

Letter Grade of Levels	Letter Grade of Half Levels and Half Value-Added					
	F	D	C	B	A	Total
F	6.1%	1.8%	0.0%	0.0%	0.0%	7.9%
D	1.8%	13.2%	3.6%	0.0%	0.0%	18.6%
C	0.0%	3.6%	16.8%	6.4%	0.0%	26.8%
B	0.0%	0.0%	7.1%	18.9%	3.2%	29.3%
A	0.0%	0.0%	0.0%	3.2%	14.3%	17.5%
Total	7.9%	18.6%	27.5%	28.6%	17.5%	100.0%

Table 5
Simulated Effects of Switching from Levels to Value-Added in a Policy Regime that Takes Over the Bottom 5% of Schools Annually For Four Years

	School Performance Measures				
Outcome Measures	Test Scores	Grad Rate	College Entry Rate	Tests & Grad Rate	Tests & Grad & College Entry Rate
Panel A: Elem. Schools (1-year)					
Diff in student-level s.d. unit	0.015				
Panel B: High School (1-year)					
Test score (student-level s.d.)	0.031			0.013	0.014
Test score (school-level s.d.)	0.071			0.029	0.032
Grad rate		0.005		0.005	0.004
Grad rate (school-level s.d.)		0.037		0.040	0.034
College entry rate			0.008		0.003
College entry rate (school-level s.d.)			0.055		0.022
Overall Effect (school-level s.d.)				0.035	0.029
Panel C: High School (4-year Avg)					
Test score (student-level s.d.)	0.029			0.013	0.011
Test score (school-level s.d.)	0.076			0.034	0.029
Grad rate		0.003		0.004	0.004
Grad rate (school-level s.d.)		0.026		0.040	0.037
College entry rate			0.009		0.004
College entry rate (school-level s.d.)			0.070		0.033
Overall Effect (school-level s.d.)				0.037	0.033

Notes: Estimates in all the panels are shrunken. The simulation involves taking over the bottom five percent of school for each of four consecutive years, then re-calculating the mean of the bottom quintile, which is most affected by the policy change. Panels B and C use the one-step value-added model, but differ in that the former is based on one-year value-added estimates and the latter average value-added across four years to reduce error. The results are reported in both the usual units (student-level test score deviations and percentage points) and school-level s.d. For reference, the school-level s.d. for test scores, high school graduation, and college entry are: 0.38, 0.11, and 0.13, respectively.