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

Updated June 12, 2021
Published May 25, 2018

Education Research Alliance NOLA.org

What Gets Measured Gets Done: Principles for Performance Measurement in School Accountability Systems and How States Can Meet Them

June 12, 2021

Douglas N. Harris
Lihan Liu

Abstract: We outline five principles of school performance measurement for accountability purposes. Then we use these principles to evaluate current state accountability policies. Third, we provide empirical evidence on the implications of two of the principles: focusing on student outcomes most closely associated with adult outcomes and using value-added measures that focus performance measures on what schools can control. Using statewide student-level data from Louisiana, we specifically show how school performance ratings change when we align accountability with these two principles, by adding college entry to high school performance measures and by switching from outcome levels-only to a mix of levels and value-added. Fourth, we simulate the effects of alternative accountability metrics on actual school performance. We find that current policies violate some of the principles and this has consequences for 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.

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] While the focus on test scores has continued in law and in practice, the *Every Student Succeeds Act* (ESSA) of 2015 eliminated the 100 percent proficient goal and gave states more flexibility over many of the policy 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, which are the core of any accountability system.

One contribution of this study is to show that the choice of performance measures for accountability, such as those in NCLB and ESSA, can be guided by general principles. Our work builds on a rich tradition of prior work that has focused more on principles for student assessment as well as the consequences of the use of those assessments for accountability purposes (Linn, 2001; Wiliam, 2010; Koretz, 2017). Here, we focus on school performance measures, which generally include student assessments, but involve a broader set of measures and issues.

In Section II, we identify five principles based on research from a variety of disciplines (economics, philosophy, psychology, and sociology). School performance measures for accountability should: (1) focus on core educational objectives related to outcomes in adulthood; (2) focus on what educators can control; and be (3) valid and reliable; (4) inexpensive; and (5) simple and transparent.

Polikoff, McEachin, and Wrabel (2014) is perhaps most similar to our work. These authors also apply general measurement principles—construct validity, validity, reliability, and transparency—to the school performance measures used in state waivers from NCLB. The last three measurement principles in their list align with principles (3) and (5) above. Others, too, have borrowed principles from the measurement and assessment literature to evaluate school performance measures (Linn, Baker, & Betebenner, 2002; Figlio and Loeb, 2011). However, our other principles are different. First, we add the cost of creating the measures because these are becoming a non-trivial (Hart et al., 2015). Second, construct validity is essentially agnostic about what construct is of interest. We are not agnostic and argue for a focus on factors that are oriented toward long-term outcomes when students become adults.

In addition to outlining the principles, we provide a theoretical and scholarly basis for them that goes beyond references to the prior measurement literature. Since we are primarily interested in the role of these measures in a specific use—accountability—we can be specific about how accountability policies are likely to fail if the principles are violated. Our outline also includes a richer discussion of the different aspects of the various principles, including several corollary principles that we have not seen discussed in prior studies.

A second general contribution is applying these principles to the accountability systems that states have adopted more recently under ESSA. In Section III, we describe the measures used by each state as of 2018, three years after ESSA. Our analysis shows that four decades into the standards and accountability movement, performance measures are still focused on a narrow set of measures and avoid some that are more closely related to students' long-term life outcomes, a violation of principle (1). Only small steps have been made toward growth and value-added measures, a violation of principle (2).

Using data from Louisiana, Section IV shows what would happen if states changed their policies to align more closely with some of the principles. Regarding the first two principles, we ask, how much do performance measures and school ratings change 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? If changing the measures to align with the principles does not change the measures themselves (e.g., the rank order of schools), then this might mean that the principles are not important in practice. We find some noteworthy differences between how individual schools are rated under current accountability measures versus the alternatives we propose.

We address the cost aspect of the measurements by examining the degree to which school performance measures change when we use more measures to create performance indices. Our analysis shows that there are diminishing returns to performance measures because the measures are positively correlated with one another. This further implies that the benefits of additional, richer measures might not be worthwhile.

Next, we calculate reliability coefficients of the alternative metrics discussed above. As others have recognized, the shift to value-added can create problems with reliability (Kane & Staiger, 2002; Chay, McEwan, & Urquiola, 2005). With regard to validity, Angrist et al. (2017) find that although statistically significant, the magnitude of bias for conventional school value-added estimates is modest. There is also a much larger research base with regard to teacher value-added, which we interpret as coming to the same conclusion (Kane & Staiger, 2008; Chetty et al., 2014; Goldhaber & Chaplin, 2015; Rothstein, 2017).

In short, we make four main contributions: first, outlining the larger number of principles of performance measurement in greater depth and attention to their trade-offs (section II); second, showing how state ESSA plans align with some of these principles (section III); third, proposing ways to better align school accountability with at least some of the principles (sections IV-VI); and fourth, showing the potential practical implications of these solutions through simulations of closure and takeover policies (section VII). The potential of adding medium-term outcomes have also been explored by McEachin and Polikoff (2012), but we take this further by simulating the effects on student outcomes.

II. Five Principles for School Performance Measures

In this section, we list the principles of measuring school performance for accountability based on scholarship from the academic disciplines, elaborate on the reasoning behind them, and highlight their interconnections.

1. Performance measures should focus on the core educational objectives that society expects schools to accomplish, particularly to produce short- and medium-term

outcomes that are likely to improve long-term life outcomes for students. Brighouse et al. (2015), for example, emphasize that "educational goods" are those positive aspects of schooling that contribute to human "flourishing" and that "contribute to their future income and health" (p.5).[2] 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.

A key challenge to achieving this objective is that if *all* of the relevant objectives are not measured, then schools will shift attention from unmeasured objectives to measured ones (Carnoy, Loeb and Smith, 2001; Figlio and Loeb, 2011). This is where the paper's title comes from: "what gets measured gets done." Some educational objectives, even if they may be of similar importance, are more measurable than others and this can distort educator behavior in unintended ways.

The focus here on being "predictive" should not be interpreted to mean that accountability should exclude measures that are also of immediate, short-term interest. For example, parents value the safety of their children, perhaps above all else. Brighouse et al. (2015) also discuss educational values that are independent of these adulthood-focused factors. They write "Some goods may be available only in childhood. Purposeless play, naïve curiosity, unreserved joy and carefreeness are the most obvious examples." While it is possible that these are unrelated to adult outcomes, it seems quite plausible that they are all predictive of life outcomes.[3]

Since schools have multiple objectives, it will generally be necessary to include multiple measures. Moreover, to identify schools for rewards and intervention, states generally combine these multiple measures through indices. The weights attached to each

measure, whether assigned implicitly or explicitly, should also reflect their relative importance in predicting long-term outcomes.

(2) Performance measures should focus on what educators can control. This is a restatement of Harris's (2011) "cardinal rule of accountability." It is rooted, first, in the economic idea that education, in economics terms, is jointly produced by families, educators, and communities, meaning that many of the key contributors to student outcomes are outside educators' control. This creates a problem because students vary considerably in their family and community situations and, as a result, they start school with different levels on most outcomes (including test scores). Focusing on outcome levels therefore means that some schools are 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.[4] Value-added measures address this problem by accounting for students' prior achievement, which, in turn accounts for the history of students' family and community resources and their contributions to student learning, so that schools' contributions can be more clearly identified (Kane & Staiger, 2002).

This is not to say that value-added measures should completely replace measures based on outcome status/levels. The main reasons for using outcome levels are practical. First, we cannot measure student outcomes in grades K-3, which means, at present, we can really only start measuring value-added in grade 4.[5] Combining value-added with status measures therefore ensures that schools are accountable for achievement in all grades, not

just in those grades where value-added measures are available. An additional valid argument against value-added measures arises under principle 5, which address the simplicity and transparency of school performance measures.

Other reasons for using status measures are less persuasive. For example, one might argue that we cannot rely on value-added measures because policymakers wish to track schools' progress toward meeting objectives, such as 100 percent proficiency, which cannot be done with value-added measures. But an important distinction has to be made between *holding schools accountable* for results, which is of primary interest here, versus *measuring progress toward system-level goals*. School performance measures for accountability should be designed to encourage progress toward goals, but the performance measures should not be confused with the goals themselves.[6] Similarly, it has been argued that status measures allow us to hold all students to the "same standard." However, standards are a type of goal. We can say that we want every student to be proficient, for example, and use value-added to measure how much schools help students reach that standard.

What happens when we hold schools accountable for things outside their control by relying on status measures? First, educators may misjudge their performance. When they try a new initiative and the status measures do not improve, they may judge that it is a failure even when it succeeds in improving student outcomes. Also, from a psychological standpoint, the perception that they might be held accountable for outside factors may reduce educator motivation (DeNisi & Pritchard, 2006; Gneezy et al., 2011).[7] Further, since status accountability measures place pressure on low-performing schools, teachers might wish to avoid misplaced blame by avoiding high-need schools. In contrast, with

value-added, teachers would be rewarded for helping students learn, regardless of their initial achievement levels.

Parents might also respond to measures in unintended ways. To see why, note that: (a) due to current and historical socio-economic inequality, students from families with higher incomes tend to start school with higher academic readiness (Lee and Burkham, 2002); and (b) higher-income families can more easily afford high-quality schools, so that the school choice process will tend to reinforce pre-existing educational segregation and inequality.[8] This tendency is compounded by peer effects (Sacerdote, 2010); when high-scoring students are initially concentrated in particular schools, their initial advantages are compounded.

Value-added measures could attenuate this self-reinforcing cycle of educational inequality. If school performance measures focused more on what schools contribute to learning, then high-income families might be less inclined to segregate on test levels, lessening the role of peer effects. Value-added measures by themselves could only have a small effect on these larger forces, of course, but even a small effect might be meaningful.

Corollary to Principle (2). Performance measures should be based on information that is proximal in time to the educator behavior that may have affected student outcomes. Time is a key factor for holding educators accountable for what they can control for two reasons. First, the longer we wait to measure outcomes, the more likely it is that the causes of those outcomes can be attributed to factors other than the educators being held accountable (e.g., changes in family and community conditions noted above). Second, "holding schools accountable" means holding specific educators accountable, but educator turnover means that the educators in each school are changing over time. If we hold

educators accountable now for the outcomes produced by actions years earlier, then we would be holding one group of educators accountable for the actions of another group—clearly outside their control.

(3) Performance measures should be valid and reliable. We mean this in the usual statistical sense that measures should capture the construct on average (validity) and that there is limited random error (reliability). These are basic principles of any form of measurement (Linn, 2001; Kane & Staiger, 2002; Polikoff, McEachin, and Wrabel, 2014; Koretz, 2017).[9] Either form of error places greater distance between educator efforts and how those efforts are reflected in the performance measures, which is the core problem of accountability.

Corollary to Principle (3). It should also be difficult to distort or manipulate measures. The manipulation or distortion of measures in accountability is a specific type of validity problem, related to Campbell's Law (1979): "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." The specific concern is that educators might be able to look better on accountability metrics without improving the underlying construct the measure is intended to capture (Clotfelter & Ladd, 1996). Teaching to the test is one obvious example (Koretz, 2017) as is cheating (Jacob and Levitt, 2003). High school graduation rates can also be manipulated (Harris, et al., 2020).

(4) Performance measures should be inexpensive. This is a simple matter of resource allocation. The resources devoted to performance measurement cannot be used for other aspects of the educational enterprise (Levin & McEwan, 2000). With standardized

tests, the key costs include the time of students and teachers, which take away from instructional time and other non-educational opportunities. By one estimate, students and teachers are spending 20-25 hours per year taking state standardized tests (Hart et al., 2015).[10] This is in line with a survey and analysis of a sample of urban districts suggesting that 1.7 percent of all instructional time is spent on state and district standardized tests (Teoh, et al., 2014). Other measures, such as attendance, also require student and teacher time (i.e., calling out the attendance roll at the beginning of each class).

This principle applies only to the marginal costs, however. Costs that would be incurred without accountability are not relevant here and some of the seemingly costly measures might be necessary independent of accountability systems. Schools took attendance well before accountability systems included this measure.

(5) Performance measures should be simple and transparent. This principle, suggested by some scholars of school accountability (Figlio & Ladd, 2011; Deming and Figlio, 2016) as well as general personnel performance assessment (e.g., Bowman, 1999), is partly rooted in evidence that people are constrained in their ability to process vast quantities of information; decision-making suffers under heavy cognitive loads (e.g., Sweller, 1994; Deck & Jahedi, 2015).[11]

Even when individual measures are simple, the aggregated metrics of accountability systems might not be. Since federal accountability requires identifying schools for improvement interventions, it is necessary to combine performances in some way, usually into an index measure where multiple measures are weighted. It is sometimes unclear in these indices how much weight is being given to specific measures.[12]

Discussion. One overarching theme of these principles is that it should be feasible for educators to respond productively to the measures used in accountability, in a way that improves student outcomes. This applies especially to principles 2, 3, and 5. It is easiest to respond to incentives productively when they are within the control of actors, subject to little random error, and/or easy to understand.

Meeting all five principles is difficult because there are trade-offs among them in practice. The measures predictive of students' long-term outcomes may not be most within educator control (principles (1) and (2)). For example, initial college entry, which is strongly predictive of life outcomes (Wolfe & Haveman, 2002; Hout, 2012), is somewhat less within the control of schools than high school graduation. Also, adjusting measures to be within the control of educators (principle (2)) may make the measures more complicated and costly (principles (4) and (5)). Using outcomes that arise in the future (principle (1)), such as employment may also reduce validity and reliability because it can be harder to track these outcomes over long periods of time for individual students (principles (3)). Finally, taking steps to reduce manipulation of the measures (principle (3)) will generally increase the costs (Figlio & Loeb, 2011).

We considered adding the additional principle that societal objectives might also pertain to the *distribution* of student outcomes along two dimensions: (a) the distribution of outcomes across student groups; and/or (b) whether any given group of students reaches some specific performance standard. Federal law, for example, has focused on both of these distributional considerations, requiring that students reach proficiency and holding schools accountable for racial, income, and disability subgroups. We decided to exclude this principle, first, because it is not sufficiently widely held. There was considerable

debate about the use of sub-groups in NCLB, reflecting the long-standing arguments about the relative importance of excellence and equity (Noguera, 2001). Also, without a clear definition of the specific desired distribution of outcomes it would be difficult to determine whether a given set of school performance measures are consistent with the principle.

Given these complications, our aims in this section are modest. We have simply tried to outline and justify a set of principles to help guide policy decisions, principles that are broadly accepted and rooted in educational and social science scholarship. We leave extensions of this proposed list to future research and instead proceed by showing how some of these principles can be applied to actual policy.

III. Applying the Principles to State ESSA Plans

It is not obvious that the current state and federal accountability systems optimally balance the five principles. Table 1 summarizes the core elements of each state policy as of 2018.[13] In what follows, we summarize our assessment of school performance measures according to the five principles.

Evaluating States on Principle (1). 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 outcomes, particularly years of education, are predictive of students' long-term outcomes (even after controlling for test scores).[14] High school graduation is now required as a performance measure at the high school level, and there is clear evidence that high school graduation also meaningfully affects life outcomes (e.g., Levin et al., 2007).[15]

In theory, the passage of the federal ESSA law freed up states to use alternative measures more in line with this principle. Our analysis shows that this has occurred, but only to a limited extent. In reviewing states' ESSA plans, we find that the most common measure states added was students' school attendance (Table 1).[16] This usage is rooted partly in correlational research linking attendance to long-term outcomes (Rumberger, 1987; Halfors et al., 2002; Harlow, 2003; Balfanz, Herzog, and MacIver, 2007; MacIver, 2011; Bauer et al., 2018) and attendance is easy to add because it is already widely measured. Still, it is noteworthy that the strongest predictor of life outcomes—college enrollment[17]—is still largely omitted. Only four states mention college outcomes in their ESSA plans (Connecticut, Michigan, Nebraska and Vermont).

Evaluating States on Principle (2). While value-added measures are gradually gaining acceptance, our analysis shows that only 30 states are using measures that include the words "growth" or "value-added." Twenty-two of these states, however, are planning to use "growth-to-target" or "growth-to-proficiency."[18] 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 measures are 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. Nevertheless, states have relied on growth-to-target, perhaps because they align with the intuitive notion of being "on track" or because of a mistaken belief that it addresses the problem with test levels, or because they believe it represents a sort of compromise between levels and growth. Also, of the four states that are clearly using value-added (and not a mix of this with growth-to-proficiency), only one is weighing value-added more than 40 percent.

While the recent debate, and the discussion above, has been about applying value-added adjustments to test scores, this method can also be applied to high school graduation, attendance, college entry, and other measures.[19] Jackson et al. (2014), for example, find that school contributions to these and other outcomes vary considerably. Value-added need not be synonymous with standardized test scores.

We focus mainly on the first two principles in this section due to space constraints and the fact that these have received much less attention than the others. These first two principles also the focus of the empirical analysis that follows.

IV. Data and Performance Indices

Most of the data used in the empirical 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).

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. 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 a high school 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 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 (including both 2-year and 4-year college) and zero otherwise.[20] 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 released an earlier version of this analysis (Harris & Liu, 2018) and this was followed by a similar analysis (Deutsch, Johnson, and Gill, 2020) using these same data. The latter study is similar in focusing on the role of value-added adjustments and extending their use to high school graduation, college, and other outcomes. The main overall difference is that our study focuses on the specific uses of performance measures in accountability. Therefore, we exclude college graduation and labor market earnings, which are included in their study, because these measures are too far in the future to be impractical for use in school accountability. Also, reflecting their different purpose,

Deutsch, Johnson, and Gill (2020) do not simulate the effects on school performance ratings.

V. Measuring School Quality

V.A. Value-added Estimates

In this paper, we estimated a variety of value-added models that are now standard in the research literature:

$$A_{ist} = f(A_{i,lag}) + \beta X_{ist} + \theta_{st} + \varepsilon_{ist} \quad (1)$$

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 the data section above).

The benefit of value-added measures is to separate school's contribution to students' outcomes from other factors such as prior achievement and demographic characteristics. We extend this idea beyond test scores to high school graduation and college entry, but the same logic applies to essentially any binary student outcome. While graduation can only occur once, and therefore lacks a lagged value for individual students, we can calculate an expected outcome for each student and then compare the actual to the predicted outcome as a measure of school performance. For high school graduation and college entry as performance outcome, we estimate the following model,

$$O_{ist} = f(A_{is8}) + \beta X_{is8} + \delta_{st} + \omega_{ist} \quad (2)$$

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. The estimates measure to what extent does the accrual graduation/college entry of a high school exceeds the average level from students with similar 8th grade test sores and background characteristics.

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.

V.B. Validity and Reliability of Value-Added Measures

In keeping with principle (3), we examined the validity and reliability of value-added measures. Following the method proposed by Kane and Staiger (2002), we find test score/graduation/college entry value-added measures are quite reliable (with reliability coefficients 80%, 81% and 76%, respectively).[23] The coefficients imply that, for instance,

persistent factors account for about 81 percent of the total variance in the graduation value-added measure. We also found that using value-added measures (vs. level measures) does not come at the great cost of losing reliability. Test score, graduation and college entry levels have higher reliability coefficients (93%, 86% and 87%, respectively) but still comparable to those from value-added measures (as shown above).

Although studies have shown that teacher and school value-added measures are generally valid (Kane and Staiger, 2008, Kane et al., 2013, Chetty et al., 2014, Deming 2014), no studies have examined the validity of high school graduation and college entry value-added measures. These measures would very likely be less valid than the standard test score value-added measures for reasons described in section V.A.[24] The appendix provides related evidence on this point, showing how value-added is less correlated with poverty than outcome levels.

VI. Empirical Analysis of Changes in School Ratings

In this section, we provide empirical evidence on the practical implications of changing school performance measures in the ways suggested by the principles. We start by showing the effect of switching from levels to value-added in a simple accountability framework where test scores are the only student outcome. Next, we discuss a more complex model where we add medium-term outcomes and shift from levels to value-added at the same time. Finally, we consider the costs and benefits of adding measures.

VI.A. Switching from Test Levels-Only Toward the Mix of Levels and Value-Added

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 shares in Louisiana, which, like most states, has a relatively small share of schools (eight percent) with the lowest rating of F. If more schools were in this lowest performance category, then the effect of changing performance measures would be greater.[25]

The upper-left cell of Table 2A provides the percentage of elementary/middle schools that receive F grades using both levels and a mix of 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 shift more toward value-added. For example, 75.8% of schools maintain the same letter grade while only 0.1% change by two letter grades. This is consistent with the high correlation (+0.85) between test levels and its value-added as shown in Figure 1A.

At the high school level, Figure 1B shows that the correlation between test levels and test value-added is weaker (+0.68). This is probably mostly because the testing regime in high school, where we control for 8th grade scores in the value-added model, is different from that used in middle school, where we control for the scores in the previous grade in the value-added model. Table 2B shows the transition matrix. High schools use both test scores and graduation rates as accountability measures therefore the accountability measures switch from levels-only of test scores and graduation rate to the levels/value-added mix, including four measures: test levels, test value-added, high school graduation

level, and graduation value-added (each equally weighted). 67.5 percent of schools remain in the same category while only 0.4 percent change by two letter grades.

VI.B. Adding Medium-Term Student Outcome Measures

With ESSA, states must consider more than just test scores when evaluating schools. This section examines the effect of adding college enrollment levels (*College*), which are predictive of students' long-term life outcome and within the control of K-12 schools, on top of high school test scores (*Test*) and graduation rates (*HSGrad*). Louisiana data shows that 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).[26] 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 college entry, but test scores are generally not a prerequisite to the other outcomes.

The transition matrices in Table 3 shows results comparing performance measures with test scores and high school graduation (0.5 weight for each of the two measures) with a measure that adds college entry (0.33 weight for each of the three). Adding the third measure, 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 adding a measure means changing all the weights as well. Figure 2 shows, for example, that when adding a third measure to a two-measure performance index, with correlations among the three measures as +0.6,

which is similar to the high school correlations we reported earlier, then 70-80% of schools stay in the same category.

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 (value-added) with Table 3 (multiple levels-based measures), 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.

VI.C. The Costs and Benefits of Additional Performance Measures

Principle (3) in our framework focuses on the cost of the measures. While there are multiple types of costs (student and teacher time, etc.), we focus here on how adding additional (costly) measures adds information to school performance ratings.

We focus specifically on the question: How does the composite index change when keep adding measures? The short answer is that this depends on the correlations of the various measures. 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 less impact on the composite performance measure than one that has a lower correlation.

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 outcome levels are positively correlated, so that there are likely to be diminishing marginal returns to information. 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 2 uses a simulated data set to show visually that there are diminishing marginal returns.[27] 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.

VII. Policy Simulations

To provide a sense of the potential impact of alternative school performance measures on students and schools, we carry out a policy simulation focusing on one mechanism through which better performance measures could improve schools. To be clear, this is not meant to promote such a policy, but only to illustrate how changes in measurement could influence actual student outcomes.

Specifically, we simulate a policy of taking over the bottom five percent of schools every year, for each of four years, and replacing closed schools with new ones. A similar

policy (without explicit percentages of schools being closed/taken over) has been implemented in one Louisiana city, New Orleans, in recent years. This policy is also similar to a fully implemented version of NCLB or ESSA, both of which emphasized state intervention in low-performing schools. To make the simulation as realistic as possible, we use the actual data from New Orleans schools.

The first step in the simulation is to rank New Orleans elementary/middle schools using school average test scores (across Math and ELA). Next, we identify the bottom five percent of schools and replace them with new schools which have a quality equals to the city median in that year and calculate the average school quality. We assumed the new schools have the median quality because this is how the policy played out in New Orleans (Harris & Liu, 2016).[28] Also, when closing schools, students are dispersed to new schools or other existing schools. We assume the student enrollment share of new school equals to the replaced school. This closure and takeover process is repeated for four consecutive years and obtain the average school quality for all four years, assuming average test scores for schools without closure and takeover are fixed over the four years. The simulation is very similar for high schools.

Additional details and a table of results are provided in Appendix C. In one simulation (Table C1), we examine the effect on the bottom 20 percent of schools (those directly affected by the policy) of adding college entry to accountability measure, switching from an equal mixture (weights=0.5) of test score and graduation rates to a mixture of all three outcomes with equal weights=0.33 (all status measures without value-added adjustments). With college entry added into the accountability measure, the improvement of college value-added of +0.020 school-level s.d. is not surprising. Making

decisions based on any specific measure will tend to increase that measure when that measure is used to make school closure/takeover decisions—what gets measured gets done.

In a second simulation (Table C1), we consider the effects on the bottom 20 percent of schools of switching from an equal mixture of test score and graduation rates to a system with all three outcomes and mixing levels and value-added, so that there are now six equally weighted (weight=0.166) measures: test score levels, test score value-added, high school graduation levels, high school graduation value-added, college entry levels, and college entry value-added. School value-added improves in all three dimensions (+0.034 for graduation rate school-level s.d., +0.032 school-level s.d. for test score, followed by +0.022 school-level s.d. for college entry). Based on prior research, these increases in value-added are likely to translate into improved student outcomes (CITES).

This section estimates and illustrates one mechanism through which performance measures affect actual student outcomes. While these effects might seem small, note that they reflect only one of the mechanisms through which school accountability measures might affect actual student outcomes. In addition to our focus on taking over low-performing schools, these measures can also affect internal school improvement and parental school choices (e.g., Glazerman & Dotter, 2017).

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 and show them empirically.

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 in most states, will likely help make the performance measures more predictive of students' adult outcomes (Finn, 1989; Halfors et al., 2002; Harlow, 2003; Rumberger, 1987), though the weights attached to attendance are small and, because of Campbells' Law, this measure is likely to suffer from validity issues now that it is high-stakes.

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, from the simulations, we see that switching from outcome levels to an even mix of levels and value-added would increase student achievement.

Future empirical research is needed on 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' adult outcomes. Our evidence also informs the interpretation of such research; in particular, adding more measures, on top of those included in this study, would likely have minimal effect on school performance indices or therefore accountability-related decisions.

Additional empirical work is also warranted regarding how parents and educators respond to different types of measures. While there is empirical evidence that people are more responsive to transparent information in general, it is important to know how these

measures are perceived, how people make sense of them, and how the measures change their actions. All of these considerations, in addition to the principles we habve outlined, are important to understanding the “optimal mix” of performance measures.

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.

Balfanz, R., Herzog, L., and MacIver, D.J. (2007). Preventing Student Disengagement and Keeping Students on the Graduation Path in Urban Middle-Grades Schools: Early Identification and Effective Interventions. *Educational Psychologist* 42 (4): 223–35.

Bauer, L., Liu, P., Schanzenbach, D.W., and Shambaugh, J. (2018). *Reducing Chronic Absenteeism under the Every Student Succeeds Act*. Washington, DC: Brookings Institution, Hamilton Project.

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

Brighouse, H., Ladd, H., Loeb, S. & Swift, A. (2015). Educational goods and values: A framework for decision makers. *Theory and Research in Education* 14(1): 3-25.

Campbell, D.T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning* 2(1): 67–90.

Carnoy, M., Loeb, S., Smith, T., 2001. *Do higher scores in Texas make for better high school outcomes?* Consortium for Policy Research in Education, CPRE Research Report no. RR-047.

Chay, K.Y., McEwan, P.J., & Urquiola, M. (2005). The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools. *American Economic Review* 95(4): 1237-1258.

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.

Clotfelter, C. & Ladd, H.F. (1996). Recognizing and Rewarding Success in Public Schools (pp. 23-63). In *Holding Schools Accountable* (ed). Helen F. Ladd. Washington, DC: The Brookings Institution.

Data Quality Campaign (2019). *Growth Data: It Matters, and It's Complicated*. Washington, DC.

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, David J. (2014). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review* 104(5): 406–411.

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.

Deutsch, J., Johnson, M. & Gill, B. (2020). *The Promotion Power Impacts of Louisiana High Schools*. Princeton, NJ: Mathematica.

Dewey, J. (1897). My Pedagogic Creed. *School Journal* 54: 77-80.

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. & Loeb, S. (2011). School Accountability. In Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, editor: *Handbooks in Economics*, Vol. 3, The Netherlands: North-Holland, pp. 383-421.

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.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives* 25(4): 191–210.

Goldhaber, D., Chaplin, D. (2015). Assessing the "Rothstein falsification test." Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness* 8(1), 8–34.

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.

Hanushek, E.A., Raymond, M. (2003). Lessons about the design of state accountability systems. In: Peterson, P.E., West, M.R. (Eds.), *No Child Left Behind?: The Politics and Practice of School Accountability*. Brookings Institution, pp. 127–151.

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., Liu, L., Barrett, N., Li, R. (2020). *Is the Rise of High School Graduation Rates Real? High-Stakes School Accountability and Strategic Behavior*. (EdWorkingPaper: 20-210). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/5jh6-0526.

Harris, D. & Taylor, L. (2008) *The Resource Costs of Standards, Assessments, and Accountability*. Final Report to the National Research Council. National Research Council: Washington, DC.

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., Spurgeon, L. *Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis*. Washington, DC: Council of the Great City Schools.

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.

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.

Hout, M. (2012). Social and Economic Returns to College Education. *Annual Review of Sociology* 38: 379–400.

Jacob, B.A. and Levitt, S.D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics* 118 (3): 843–77.

Jackson, C. K., Johnson, R., & Persico, C. (2014). *The effect of school finance reforms on the distribution of spending, academic achievement, and adult outcomes* (No. 20118). National Bureau of Economic Research.

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

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

Kane, T. J., and D. Staiger (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER Working Paper No. 14607*. Cambridge, MA: National Bureau of Economic Research, 2008.

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." *MET Project Research Paper*, 2013.

Kaput, K. (2018). Briefing Memo: Summaries of States' ESSA School Quality/Student Success Measures. *Education Evolving*.

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.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher* 35(7): 3-12.

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.

Levin, H. M., & McEwan, P. J. (2000). *Cost-Effectiveness analysis: Methods and Applications*. Sage Publications.

Linn, R.L., Baker, E., & Betebenner, D. (2002). Accountability Systems: Implications of Requirements of the No Child Left behind Act of 2001. *Educational Researcher* 31(6): 3-16.

Linn, R. (2001). Validation of the uses and interpretations of results of state assessment and accountability systems, in *Large-Scale Assessment Programs for All Student: Development, Implementation, and Analysis*, eds. Gerald Tindal and Thomas M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates, pp.27-48.

MacIver, M.A. (2011). The Challenge of Improving Urban High School Graduation Outcomes: Findings from a Randomized Study of Dropout Prevent Efforts. *Journal of Education for Students Placed at Risk* 16 (3): 167–84.

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.

Noguera, P. (2001). Racial politics and the elusive quest for excellence and equity in education. *Education and Urban Society* 34(1): 18-41.

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

Polikoff, M., McEachin, A.J., & Wrabel, S.L. (2014). The waive of the Future? School accountability in the waiver era. *Educational Researcher* 43(1): 45-54.

Roby, D.E. (2004) Research on school attendance and student achievement: A study of Ohio schools. *Educational Research Quarterly* 28(1): 3-14.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review* 107(6): 1656-84.

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

Sacerdote, B. (2010). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland.

Schneider, M. & Buckley, J. (2002). What Do Parents Want From Schools? Evidence From the Internet. *Educational Evaluation and Policy Analysis* 24(2): 133-144.

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.

Teoh, M., Coggins, C., Guan, C., & Hiler, T. (2014). *The Student and the Stopwatch. How Much Time Do American Students Spend on Testing?* TeachPlus. Downloaded May 28, 2021 from: http://www.hunt-institute.org/wp-content/uploads/2016/02/The-Student-and-the-Stopwatch-How-Much-Time-do-American-Students-Spend-on-Testing.pdf.

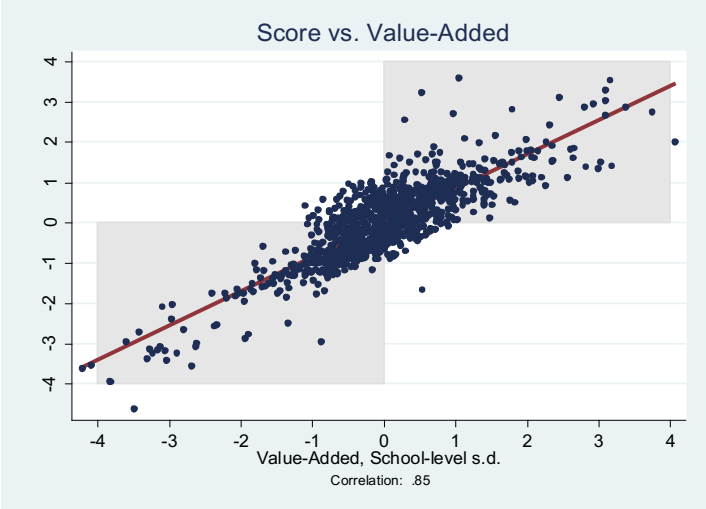
Weiss, M.J. (2008). *Examining the Measures Used in the Federal Growth Model Pilot Program*. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, D.C., March 3, 2008.

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.

Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*.107-122.

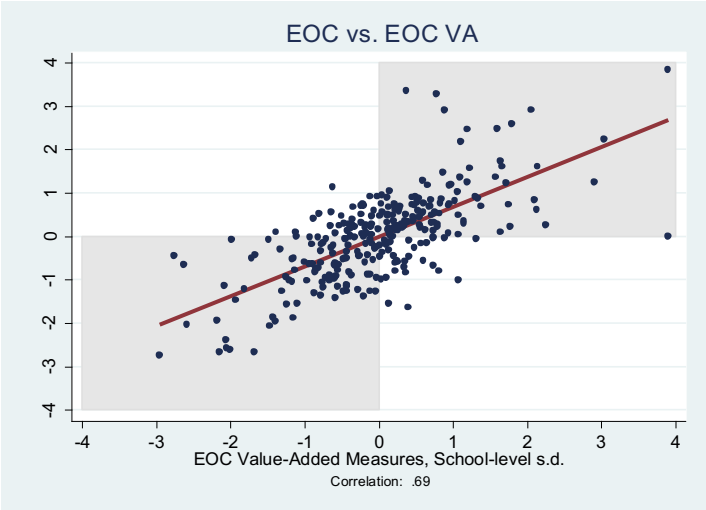
Wolfe, B.L. & Haveman, R.H. (2002) *Social and Nonmarket benefits from education in an advanced economy*. Boston Federal Reserve Conference Series.

Figure 1A
Scatterplots of Levels and Value-Added: *Elem/Middle Test Scores (LA)*



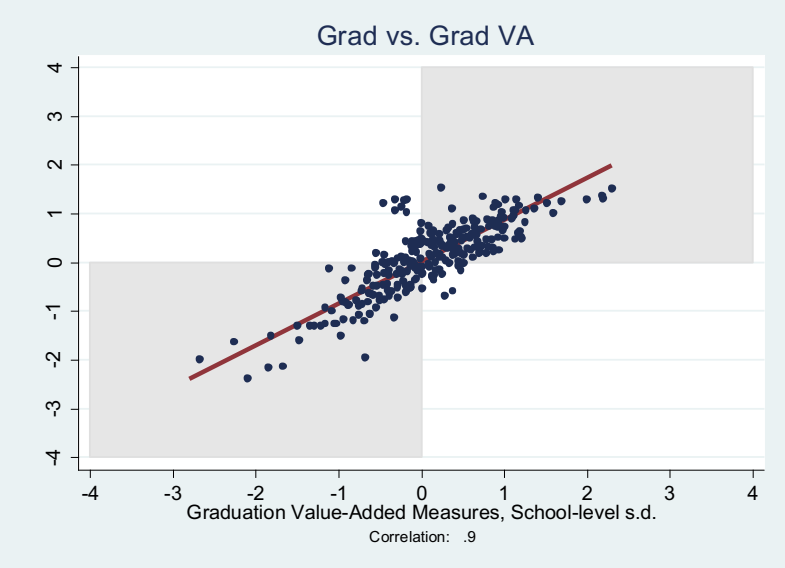
Notes: The scatterplot shows the correlation between four-year averages of school-level test levels (y-axis) and school value-added (x-axis) for Louisiana elementary/middle schools. The correlation is listed at the bottom of the figure.

Figure 1B
Scatterplots of Levels and Value-Added: *High School Test Scores (LA)*



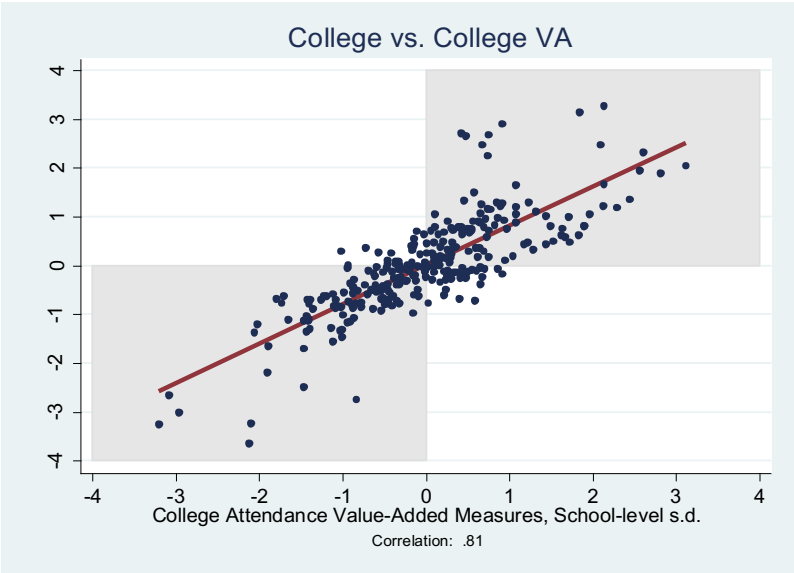
Notes: The scatterplot shows the correlation between a four-year average of school value-added (x-axis) and average test levels for Louisiana high schools. The correlation is listed at the bottom of the figure. While using the school-level standard deviation is unusual it allows us to put all the measures on the same unit of measure across the test score, graduation, and college analyses.

Figure 1C
Scatterplots of Levels and Value-Added: *Graduation (LA)*



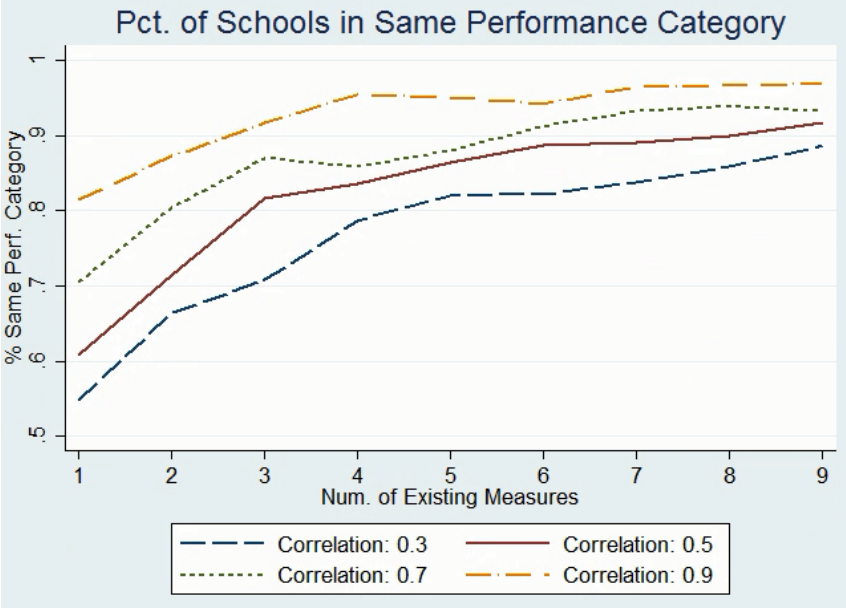
Notes: The scatterplot shows the correlation between a four-year average of graduation value-added (x-axis) and average graduation levels for Louisiana high schools. The correlation is listed at the bottom of the figure. While using the school-level standard deviation is unusual it allows us to put all the measures on the same unit of measure across the test score, graduation, and college analyses.

Figure 1D
Scatterplots of Levels and Value-Added: *College Entry (LA)*



Notes: The scatterplot 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. While using the school-level standard deviation is unusual it allows us to put all the measures on the same unit of measure across the test score, graduation, and college analyses.

Figure 2
Diminishing Marginal Returns to Additional Measures (Simulation)



Notes: This figure uses data simulated based on the assumption of the correlation across measures. See text for additional discussion.

Table 1: Summary of State ESSA Plans

States	Grade K-8			Grade 9-12		
	Growth Measures	VAM Weight	Outcomes (other than Math & Reading test scores and English-language proficiency)	Growth Measures	VAM Weight	Outcomes (other than Math & Reading test scores, English-language proficiency and Graduation Rate)
Alabama	Gain Score		Attendance	Gain Score		Attendance, College & Career Ready
Alaska	Value Table		Attendance	N/A		Attendance
Arizona	SGP, GTS		Attendance, Acceleration/readiness	N/A		College & Career Ready
Arkansas	VAM	50%	School Climate, Sci., Reading	VAM	35%	GPA, Community Service, Computer Science Course, College & Career Ready
California	N/A		Attendance, Suspension	N/A		Suspensions, College & Career Ready
Colorado	SGP		Sci., Attendance	SGP		Sci., Dropout
Connecticut	GTS		Sci., Attendance, Physical Fitness, 9th On-track	N/A		Sci., Attendance, Physical Fitness, 9th On-track, Art Access, College & Career Ready, Postsecondary Entrance
Delaware	Other		Attendance, Sci., S.S.	Other		Attendance, Sci., S.S., 9th On-track, College & Career Ready
Florida	Value Table		Sci., S.S.	Value Table		Sci., S.S., College & Career Ready
Georgia	SGP		Attendance, Sci., S.S., Literacy, Beyond the Core	SGP		Attendance, Sci., S.S., Literacy, College & Career Ready
Hawaii	SGP		Attendance	N/A		Attendance
Idaho	GTS		School Climate	N/A		College & Career Ready
Illinois	Other		Attendance, School Climate	N/A		Attendance, School Climate, 9th On-track, College & Career Ready
Indiana	Value Table, SGP, GTS		Attendance	N/A		College & Career Ready
Iowa	SGP		School Climate	SGP		School Climate, Post-Secondary Readiness
Kansas	N/A		Pct. of Low Performing	N/A		Pct. of Low Performing
Kentucky	Value Table, GTS		Sci., S.S., Writing, School Climate	N/A		Sci., S.S., Writing, School Climate, Transition Readiness

Louisiana	VAM, GTS	25%	Sci., S.S., Dropout	VAM, GTS	13%	Sci., S.S., College & Career Ready
Maine	Value Table		Attendance	N/A		Attendance
Maryland	SGP		Attendance, School Climate, Well-rounded Edu.	N/A		Attendance, School Climate, Well-rounded Edu, 9th On-track, College & Career Ready
Massachusetts	SGP		Attendance	SGP		Attendance, Sci., Dropout, 9th On-track, Advanced Coursework
Michigan	SGP, GTS		Attendance, Arts/Physical Edu., Librarian/Media Specialists	SGP, GTS		Attendance, Advanced Coursework, Postsecondary Entrance
Minnesota	Value Table		Attendance	N/A		Attendance
Mississippi	Value Table		Sci.	Value Table		Sci., S.S., College & Career Ready
Missouri	VAM	30-37.5%	Attendance	N/A	30-37.5%	Attendance
Montana	Other		Attendance, Sci., School Climate	N/A		Attendance, Sci., School Climate, College & Career Ready
Nebraska	Value Table		Attendance, Sci., School Climate, Dropout, Transitions	N/A		Attendance, Sci., School Climate, Dropout, Transitions, Educator Effectiveness, College & Career Ready, Postsecondary Entrance
Nevada	SGP, GTS		Attendance, Sci., High School Readiness, Academic Plans	N/A		Attendance, Sci., Academic Plans, 9th & 10th On-track, College & Career Ready
New Jersey	SGP		Attendance	N/A		Attendance
New Hamp.	SGP		Ach. of Low Performing, Equity,	N/A		College & Career Ready
New Mexico	SGP, VAM	40%	Attendance, School Climate, Sci.	SGP, VAM	30%	Attendance, School Climate, Sci., College & Career Ready
New York	SGP		Attendance	N/A		Attendance, College & Career Ready
North Carolina	VAM	20%	Sci.,	VAM	20%	Sci., College & Career Ready
North Dakota	Gain Score		School Climate	Gain Score		School Climate, College & Career Ready
Ohio	VAM	29%	Attendance, Sci., S.S., Literacy Improvement	VAM	23%	Attendance, Sci., S.S., Gap Closing, College & Career Ready
Oklahoma	Value Table		Attendance	N/A		Attendance, College & Career Ready
Oregon	SGP		Attendance	N/A		Attendance, 9th On-track, High School Completion Rates
Pennsylvania	VAM	-	Attendance	VAM	-	Attendance, Career Readiness
Rhode Island	SGP		Attendance, Suspension, Sci., Exceeds Expectation	SGP		Attendance, Suspension, Sci., Exceeds Expectation, College & Career Ready

South Carolina	VAM		Sci., S.S., School Climate, College & Career Ready	N/A		Sci., S.S., History, School Climate, College & Career Ready
South Dakota	SGP, GTS		Attendance	N/A		Attendance, College & Career Ready
Tennessee	VAM, Value Table	35%	Attendance	VAM, Value Table	25%	Attendance, College & Career Ready
Texas	Gain Score		Sci., S.S., Writing	N/A		Sci., S.S., Writing, College & Career Ready
Utah	SGP, GTS		Growth of Low-performing, Sci.	SGP, GTS		Growth of Low-performing, Sci., College & Career Ready
Vermont	SGP		Sci., Physical Edu.	N/A		Sci., Physical Edu., College & Career Ready, Career and College Outcomes
Virginia	Value Table		Attendance, Dropout	N/A		Attendance, Dropout
Washington	SGP		Attendance	N/A		Attendance, 9th On-track, College & Career Ready
West Virginia	Value Table		Attendance, Suspension	N/A		Attendance, Suspension, On-track to Grad.
Wisconsin	SGP		Attendance	N/A		Attendance
Wyoming	SGP		Growth of Low Performing	SGP		College & Career Ready

Source: Original state ESSA plans approved by US Department of Education, state ESSA plan summary reports by (Kaput, 2018) and (Data Quality Campaign, 2019).

Notes: The "Growth Measures" column lists the approach used to create the growth measure in the state accountability index. VAM= Value-added, SGP = Student Growth Percentile, GTS= Growth to Standard, Other= approaches other than listed ones, "N/A" = no growth measure is used. Among these approaches, only VAM applies the true value-added model. See appendix for detailed description of each approach. The "VAM Weight" column lists weights for value-added measures only. It is blank if no value-added measure is used. A dash (-) indicates that the information is unclear. The outcomes other than test scores and graduation rates are sometimes vaguely worded, we use the words from the ESSA plan with minor modification (e.g., many states refer to "chronic absenteeism," which we reduce to "attendance").

Table 2A:
Transition Matrix: Levels-Only versus Half Levels and Half Value-Added
(Elementary/Middle Schools, Test Score Only)

	Letter Grade of Half Levels and Half Value-Added					
Letter Grade of Levels	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%

Note: This transition matrix shows how performance ratings would change in Louisiana, assuming that the share of schools with each of the ratings is held constant. For example, the upper-left cell provides the percentage of elementary/middle schools that receive F grades using both levels and a mix of 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 a mix of levels and value-added.

Table 2B:
Transition Matrix: Levels-Only versus Half Levels and Half Value-Added
(High Schools, Test Score and Graduation Rate)

	Letter Grade of Half Levels and Half Value-Added					
Letter Grade of Levels	F	D	C	B	A	Total
F	5.7%	2.1%	0.0%	0.0%	0.0%	7.9%
D	2.1%	12.9%	3.6%	0.0%	0.0%	18.6%
C	0.0%	3.6%	17.9%	5.7%	0.4%	27.5%
B	0.0%	0.0%	6.1%	18.2%	4.3%	28.6%
A	0.0%	0.0%	0.0%	4.6%	12.9%	17.5%
Total	7.9%	18.6%	27.5%	28.6%	17.5%	100.0%

Note: This transition matrix shows results from analogous analysis to Table 2A for high schools. The only difference is that high schools use both test scores and graduation rates as accountability measures in the era of ESSA therefore the accountability measures switch from levels-only of test scores and graduation rate to the levels/value-added mix, including four measures: test levels, test value-added, high school graduation level, and graduation value-added (each equally weighted).

Table 3:
Transition Matrix: Adding College Entry (Levels-Only for All Measures)

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%

Note: This transition matrix shows how high school performance ratings would change in Louisiana, when college entry rate is added into accountability measure on top of test scores and graduation rate. For example, the upper-left cell provides the percentage of high schools that always receive F grades both before and after accountability measure change, and the remaining diagonals do the same for the other letter grades. The off diagonals show the schools that switch letter grades due to the addition of college entry.

Online Appendix

What Gets Measured Gets Done: Principles for Performance Measurement in Accountability Systems and How States Can Meet Them

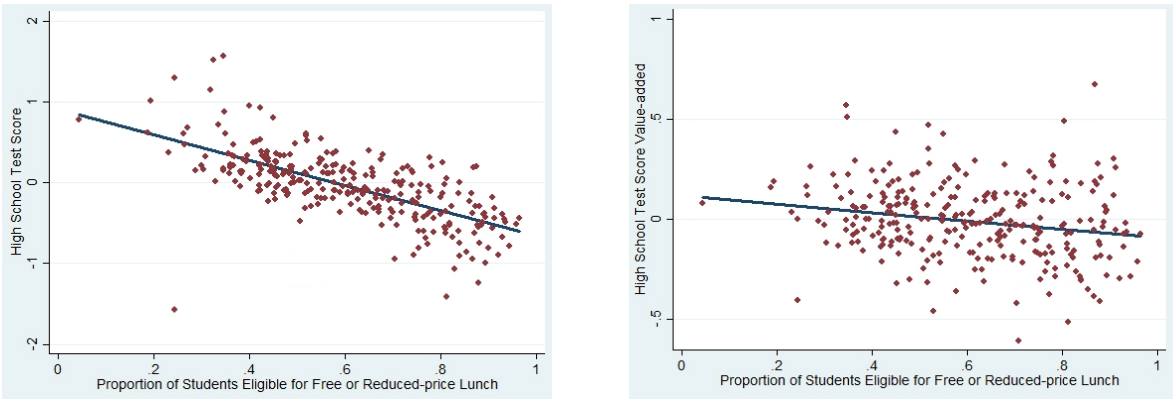
Douglas N. Harris
Lihan Liu

Appendix A: How Poverty Correlates with Test Levels and VA

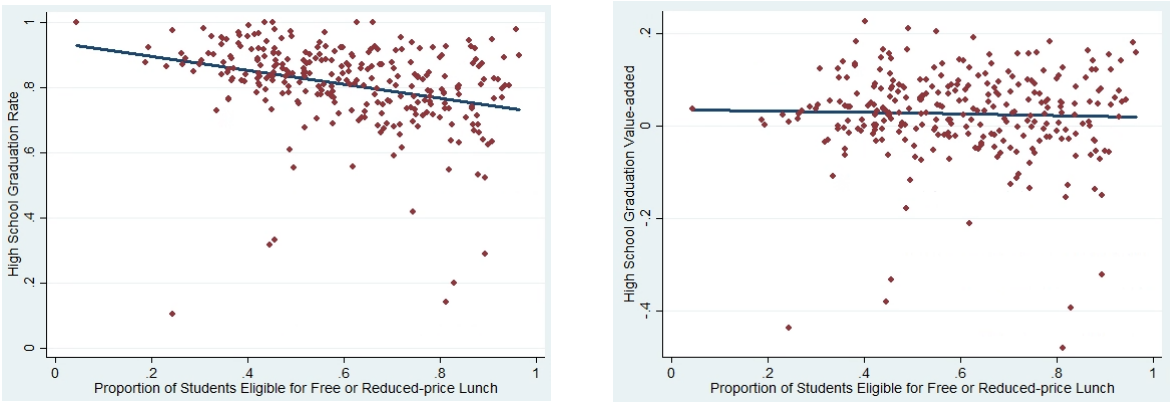
Value-added to high school graduation and college entry are very likely to better reflect schools' actual contributions to these outcomes. The left panels of Figure A1 show that school average levels of test score, graduation and college entry are all negatively correlated with school poverty, which is measured by the percentage of students eligible for free or reduced-price lunch. After controlling for 8th grade student characteristics, graduation and college entry value-added measures are not correlated with school poverty. Although test score value-added is significantly negatively correlated with school poverty, the slope becomes flatter comparing to test score level. This implies that value-added measure effectively isolating the impacts of concentration of advantaged or disadvantaged students on school outcomes.

Figure A1: Correlation between Levels/Value-Added and School Poverty

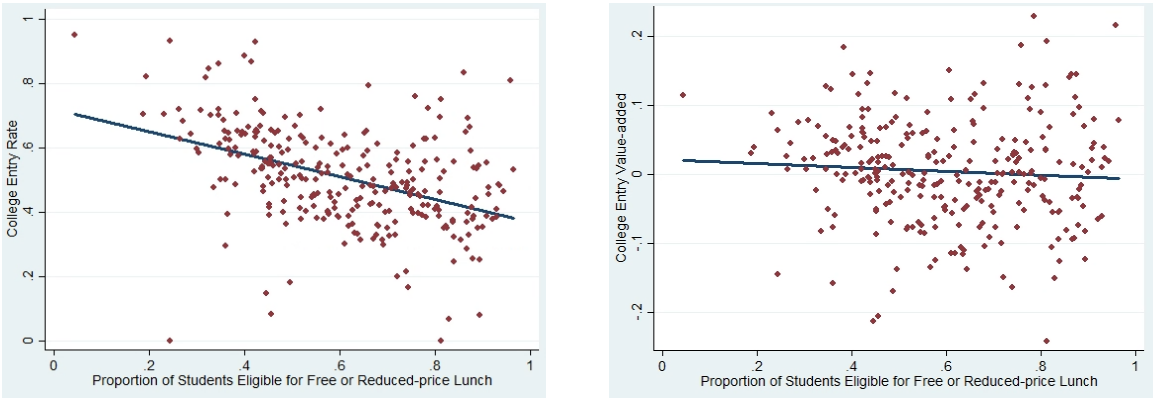
Panel A: High School Test Score



Panel B: High School Graduation



Panel C: College Entry



Notes: Each dot represents a Louisiana high school. The straight blue lines are the linear fitted lines. For high school graduation and college entry, the slope of the regression line on the left is negative and statistically significant; the slope of the regression on the right is not significantly different from zero. The fitted line for test score value-added is flatter than the one for test score level, although both slopes are statistically negative.

Appendix B: Growth Measures in State ESSA Plans

According to approved ESSA plans (summarized in Table 1), all but two states have committed to measuring student growth as an indicator of academic achievement in accountability systems for elementary and middle schools, and 20 states will do so for high schools. Some states are using a sophisticated analysis of multiple data points that evaluate the impact of schools on student learning, while others are using simpler measures of change in student assessment year to year. As shown in Table B1, different approaches answer different questions and tell different stories about what is happening in schools and classrooms.

Table B1:Measures of Growth Used in State Accountability Systems

Type	Value Added	Student Growth Percentile	Value Table	Gain Score	Growth to Standard
These measures use individual student performance data to demonstrate . . .	The impact of adults in a school on student achievement.	How schools served students with the same academic starting point.	Student progress.	Student progress.	A student's distance from grade-level learning goals.
They are calculated by . . .	Using advanced statistics to analyze data about Joey's past performance, and sometimes other characteristics that would affect his score, such as income or English language learner status, to predict how Joey will perform on the assessment. Joey's actual growth score is compared to his expected growth score, and the result is attributed his school.	Using data about Joey's past performance, to group Joey with students across the state who got the same or similar score on the same test in the same grade. Joey is then assigned a percentile or rank—between 1 and 99—based on how his current year performance compares to that of his academic peers.	Using a series of performance levels developed by the state that are based on a range of scores (e.g., 1–12 points, 13–24 points, etc.). Joey's test score this year is placed in a performance level and compared to where his score fell last year. These measures note whether Joey moved between levels year to year.	Looking at the change in Joey's test score on a comparable assessment from last year to this year. An additional layer of analysis is applied to the test scores to make this measure possible; they are translated into what is called a "scale score," which allows for comparison (for example, if students take different versions of a test).	Comparing Joey's performance this year to a long-term learning goal. Assuming Joey will improve at the same rate every year, this type of measure estimates whether Joey is on track to achieve that goal within a given timeframe.
The resulting data tells us . . .	Joey's school helped him improve more than other schools helped similar students.	Joey is in the 70th percentile; compared to a group of academically similar peers, he did better than 70 percent of them.	Joey moved from below basic to basic based on the state's cut scores.	Joey scored 50 points higher than last year.	As a 4th grader, Joey is 100 points away from proficiency and is on track to be proficient by the end of the next two years.

Source: Summary table on page 4 of (Kaput, 2018).

Appendix C: Simulations

Switching from Levels to a Mix of Levels and Value-Added

This section provides more detail on the simulations discussed in the text. The results are summarized in Table C1.

Elementary/Middle Schools

To examine the effect of accountability measure changes, we conduct the simulation described in the main text under an alternative accountability measure that is an even (50-50) mixture of test score levels and value-added, as opposed to the more standard test score level-only measure. The transition matrix discussed above provide some indication of what would happen. As shown by the top of Table 2A, 1.2 percent of schools that were in the F category based on test levels (sufficient for state intervention in Louisiana) end up in the B to D categories under the new accountability measure and would therefore not be subject to intervention.

Table C1 Panel A shows that the average school quality (measured as test value-added) of the original bottom 20 percent of schools (we focus on the bottom 20 percent of schools because these are the only schools directly affected by the policy) increases by 0.015 student-level standard deviations due to the accountability measure change. 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.

High Schools

The high school analysis applies nearly the same simulation procedure as above except using different accountability measures. As noted earlier, high schools use both test

scores and graduation rates as accountability measures in the era of ESSA, the initial accountability measure includes two measures: test scores and graduation rates with equal weights. The counterfactual accountability measure switches from levels-only to the levels/value-added mix, including four measures: test levels, test value-added, high school graduation level, and graduation value-added (each equally weighted).

The top of Table 2B shows that 2.1 percent of schools that are in the F category based on test score and graduation rate (sufficient for state intervention in Louisiana) end up in the B to D categories after the accountability measure change and would therefore not be subject to intervention.

Column (2) of Table C1 Panel B shows the change in average school quality of the original bottom 20 percent of schools. Similar to what we found with elementary school, test scores value-added improves by +0.013 student-level s.d.. High school graduation value-added increases by about +0.5 percentage point. When we switch to school-level s.d., so that we can compare the test score results to high school graduation, the effects appear larger for high school graduation value-added (+0.040 school-level s.d.), followed by test scores value-added (+0.029 school-level s.d.). The overall average effect is +0.035 school-level s.d..

The above results in Panel B are from one single cohort, the 9th graders in 2008. Panel C switches from single cohort to four-cohort averages using the 9th graders between 2008 and 2012. The changes are small and do not operate in any particular direction.

Adding Medium-Term Outcomes

The ESSA plan gives states more flexibility over choice of school performance measures. We conduct the third simulation to examine the effect of adding medium-term

outcome college entry to accountability measure, switching from a 50-50 mixture of test score and graduation rates to a mixture of all three outcomes with equal weights.

Column (3) of Panel B in Table C1 shows the change in average school quality of the original bottom 20 percent of schools. With college entry added into the accountability measure, the improvement of college value-added of +0.020 school-level s.d. is not surprising. Making decisions based on any specific measure will tend to increase that measure when that measure is used to make schooling decisions—what gets measured gets done.

More surprising is that adding college entry has an even larger effect on high school graduation value-added (about +0.034 school-level s.d.). This is possible because college entry is more correlated with graduate rate (+0.74) than with test score (+0.32). Thus, adding college entry is equivalent to give more weight to graduate rate and less weight to test score. This reinforces that the effect of adding any one measure depends on which measures we start with and on the covariance among all the measures. The multiple-cohort effects in Panel C have similar magnitude to the single-cohort effects in Panel B.

Switching from Levels to Levels/value-added Mix and Adding Medium-term Outcomes

How much would average school quality change if switching from levels to levels/value-added mix and adding medium-term outcomes happen at the same time? In Column (4) of Table C1 Panel B, we simulate the case where accountability measure shifts from a 50-50 mixture of test score and graduation rates to a system with all three outcomes and mixing levels and value-added, so that there are now six equally weighted measures

(test score levels, test score value-added, high school graduation levels, high school graduation value-added, college entry levels, and college entry value-added).

School value-added improves in all three dimensions (+0.034 for graduation rate school-level s.d., +0.032 school-level s.d. for test score, followed by +0.022 school-level s.d. for college entry). The improvement is robust when four cohorts are used (+0.037 school-level s.d. for graduation rate, +0.033 school-level s.d. for college entry, followed by +0.029 school-level s.d. for test score).

Comparing columns 4 to 2, the improvement in graduation value-added does not come at the cost of school quality in the other two dimensions. The improvement in graduation value-added drops only slightly (from +0.040 to +0.034 using single cohort, from +0.040 to +0.037 using four cohorts), as expected given that the weight attached to the graduation value-added is lower when college entry is also part of the mix. The change in test score value-added is also small in magnitude and is even positive when single cohort is used.

Table C1: Simulated Effects in a Policy Regime that Takes Over the Bottom 5% of Schools Annually For Four Years

Initial Acct. Measures	(1) Test Score (Levels Only)	(2) Tests & Grad Rate (Levels Only)	(3) Tests & Grad (Levels Only)
Acct. Measures after Policy Change			
	Test Scores (Levels/VA Mix)	Tests & Grad Rate (Levels/VA Mix)	Tests, Grad & College Entry (Levels Only)
Panel A: Elem. Schools (1-year)			
Diff in test value-added (student-level s.d.)	0.015		
Panel B: High School (1-year)			
Diff in test value-add (student-level s.d.)		0.013	0.003
Diff in test value-add (school-level s.d.)		0.029	0.007
Diff in grad value-add		0.005	0.005
Diff in grad value-add (school-level s.d.)		0.040	0.036
Diff in college entry value-add			0.003
Diff in college entry value-add (school-level s.d.)			0.020
Overall Ave. Effect (school-level s.d.)		*0.035*	*0.021*
Panel C: High School (4-year Avg)			
Diff in test value-add (student-level s.d.)		0.013	-0.002
Diff in test value-add (school-level s.d.)		0.034	-0.005
Diff in grad value-add		0.004	0.004
Diff in grad value-add (school-level s.d.)		0.040	0.034
Diff in college entry value-add			0.003
Diff in college entry value-add (school-level s.d.)			0.022
Overall Ave. Effect (school-level s.d.)		*0.037*	*0.017*

Notes: This table reports the change in average school quality due to accountability measure shift. Each column represents one simulation, which involves taking over the bottom five percent of New Orleans schools 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 shrunken estimates from the one-step value-added model as shown in equations 1 and 2 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 comparison across various performance measures. 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. The last row in Panel B and Panel C reports the average effect across all performance outcomes.

Notes

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

[2] Brighouse et al. (2015) list several specific elements of flourishing: economic productivity, personal autonomy, democratic competence, healthy personal relationships, treating others as equals, and personal fulfillment.

[3] This list also echoes John Dewey (1897) who wrote that "I believe that education, therefore, is a process of living and not a preparation for future living."

[4] 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.

[5] It would also be necessary to measure achievement on the first day of kindergarten in order to calculate growth during that first year.

[6] Another potential reason for focusing on outcome levels is to provide information to families as they choose schools for their children. Families do seem responsive to test scores levels when they make schooling choices (e.g., Glazerman & Dotter, 2017). Test scores levels provide signals of peer characteristics, which are clearly important to families (Schnieder & Buckley, 2002). However, from a social welfare standpoint, these are externalities and may be a zero-sum game.

[7] 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).

[8] There are many reasons for this correlation, rooted in current and historical discrimination. Ladson-Billings (2006), for example, refers to this as the "debt and deficit" of education.

[9] 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 the main concern. On the other hand, the differences in true performance between schools may be relatively small, so even a seemingly small reliability problem could have significant consequences (Kane & Staiger, 2002; Harris, 2011).

[10] The costs of creating valid and reliable tests, often reflected in the contracts with testing companies, is an additional cost, but these costs are small on a per-pupil basis (Harris & Taylor, 2008).

[11] There is some prior evidence that providing more information can lead to worse consumer decisions (Keller & Staelin, 1987; Ariely, 2000).

[12] In a simple weighting scheme, there are separate measures combined together with explicit weights with the general form $\sum_i \omega_i M_i$ where $\sum_i \omega_i = 1$. In other cases, performance indices are based on complex protocols where the weights are difficult to discern and/or vary across schools. For example, in Louisiana, schools with high achievement levels automatically receive high "growth" measures. This makes it difficult to generalize about the weight given to growth.

[13] For additional analyses of state ESSA plans see, for example: Martin, Sargrad, & Batel (2016) and Aldeman et al. (2017).

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

[15] 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).

[16] 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.

[17] 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 increased years of education.

[18] Value-added-like measures include value-added measures as well as Student Growth Percentile (growth to target approach). See appendix for detailed descriptions.

[19] 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).

[20] There is some missing data, especially in the two-year college level, during the years of our analysis (Dynarski, Hemelt, & Hyman (2015)), though this is unlikely to affect the general conclusions we draw here. The missing data could reduce reliability, but would only affect validity if missingness happened to be correlated with true school value-added. This is possible, but we note that the college data are coming the colleges and the school data from the schools, which reduces the chance of significant correlation.

[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.

[23] Formally, suppose that measure (y_t) is the sum of a fixed component (α), a persistent component (v_t), which follows an AR(1) process ($v_t = \beta v_{t-1} + u_t$), and an independent identically distributed transitory component (ε_t). Then it is straightforward to show that (1) $\sigma_v^2/\sigma_y^2 = \rho_1/\beta$, where ρ_1 is the correlation of the measure with a one-year lag, σ_v^2 and σ_y^2 are the total variance of v_t and y_t respectively. (2) $\beta = \rho_2/\rho_1$, where ρ_2 is the correlation of the measure with a two-year lag. With four years of data, we estimate β using the average of ρ_2/ρ_1 and ρ_3/ρ_2.

[24] To reiterate, there are two reasons: 1) the timing of effectiveness could not be accurately identified because outcome measures reflect accumulative effects from multiple years; and 2) the standard test score value-added calculation controls for the test scores in the previous year which might be likely more predictive than 8th grade test scores. Factors such as parental education are in most cases out of schools' control and might affect those medium-term outcomes.

[25] This is because, if schools are normally distributed, then increasing the percentage of failing schools would imply a larger share of failing schools near the threshold. The schools near as the threshold are most sensitive to even small changes in performance measures.

[26] We also looked for other evidence on the correlation between academic outcomes and other measures that are sometimes added into performance indices. 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).

[27] 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).

[28] We also 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. The lack of sensitivity is due to the fact 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.