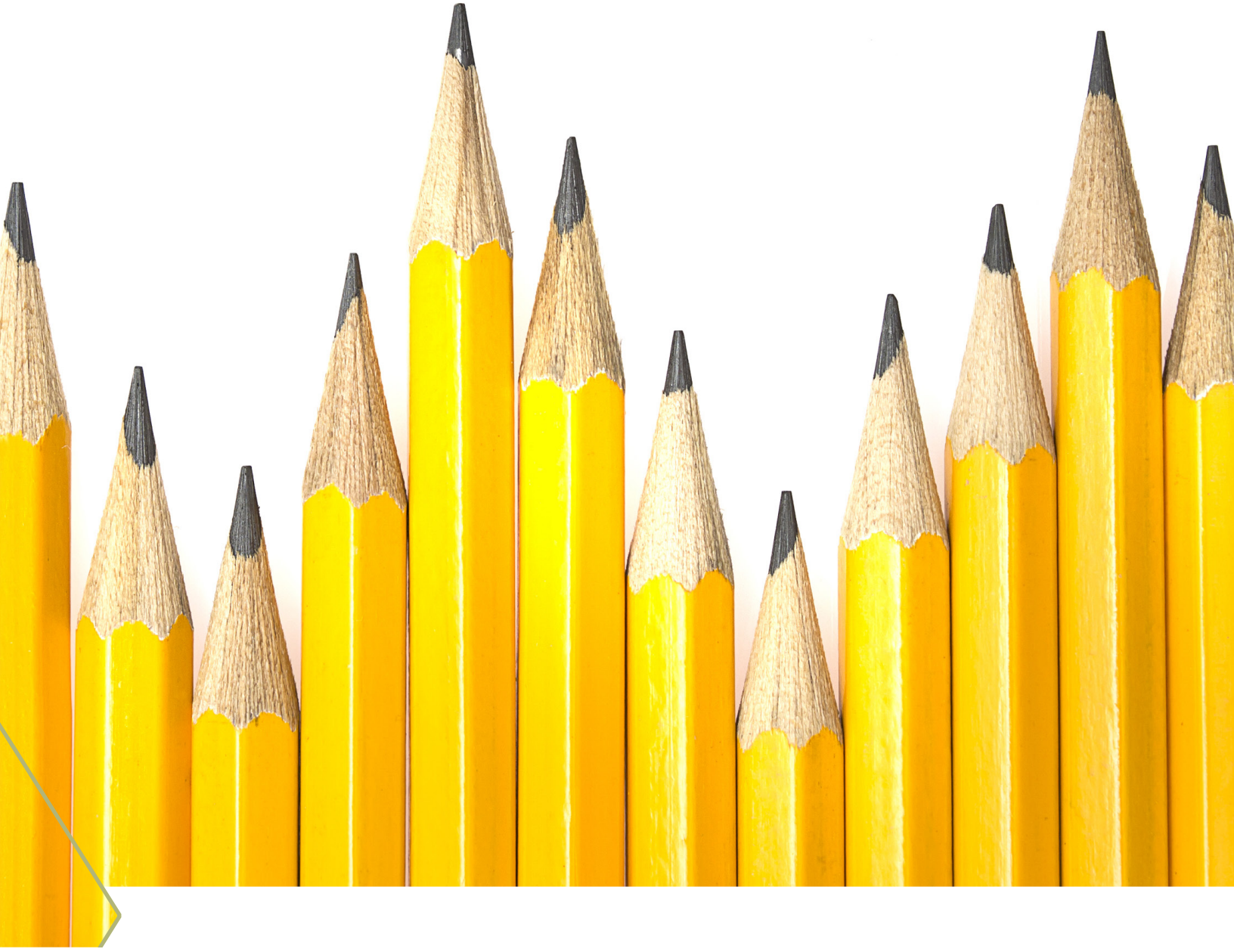
OCTOBER 2016

Grading Schools

How States Should Define "School Quality" Under the Every Student Succeeds Act



Chad Aldeman

BELLWETHER
EDUCATION PARTNERS
IDEAS | PEOPLE | RESULTS

Table of Contents

Introduction

The way we evaluate school quality in this country is outdated. We've been too rigid and formulaic, and we too often identify "low-performing" schools only by where students end up, rather than how much they progress in a year's time. We've ignored students in certain grades and subjects, and we've been willfully blind to whether students are truly prepared for college or careers when they graduate high school.

The window is now open for a new conversation about how to improve our nation's schools. Confronted with a strong desire for new ways to measure school quality and for more nuanced responses to the results, Congress replaced the 13-year-old No Child Left Behind Act with a new law, the Every Student Succeeds Act (ESSA). ESSA requires states to design their own accountability systems, but it gives states much more flexibility in how those systems are built and what consequences happen as a result.

States can and must use ESSA as an opportunity to do better.

Those new opportunities also carry new risks, and there's a danger in letting the pendulum swing too far the other way. If states fail to use their newfound flexibility wisely, they may not provide sufficient urgency for schools to improve, especially for the disadvantaged students who rely on public schools the most and who have historically been under-served by them. States are currently contemplating a wide array of measures that are more *elaborate* than anything under NCLB, but those may create new problems. Or, if states simply adopt "NCLB-lite" accountability systems, school quality may continue to be defined largely by low-level measures just as they were under NCLB. States can and must use ESSA as an opportunity to do better, not just less or different.

States should start by rethinking the purpose and goals of their accountability systems and the proper state role in the education system. To begin with, school rating systems should not merely be an act of punishment. Instead, they should provide clear evidence to parents and the general public about what's happening in schools, while also providing clear signals to school leaders and teachers about how to improve.

States are the best-positioned entity to create accountability systems that accomplish these goals. States collect much of the data needed to compare schools and districts against each other on quantifiable metrics. As the authorizing entity for local school districts, states also have the authority and legal responsibility to set minimum standards of quality. But after that, we may be asking too much of states to expect they are capable of actually helping schools improve. Instead, local communities will have to make the tough decisions on how to improve their schools. That leaves states in the role of defining what it means to be a "low-performing school" and then providing those schools resources and guidance on how to turn around.

This paper begins with a discussion of accountability, why it matters, and for whom. It then turns to specific design choices states must consider as they create their own school rating systems. In brief, it argues that states should create school rating systems that are simple, clear, and fair for schools. There are a range of indicators states could employ to accomplish these objectives, but states must be vigilant against selecting measures that hold schools accountable for things the schools themselves can't control or that subtly encourage schools to change their behavior in unproductive ways. These principles hold true for schools at all grade levels, but as one concrete example, the paper concludes with suggestions for a model accountability system for elementary and middle schools. It would use test scores, cut in new ways, as an initial "flag" on low performance, but then rely on holistic, on-site school reviews from professionally trained inspectors as a way to both investigate school quality and suggest ways to improve.

What Is Accountability and Why Does It Matter?

In order for states to receive their share of federal Title I dollars, ESSA requires them to create their own accountability systems. The law imposes additional requirements on what measures must go into those systems and how they can be combined, but before state leaders begin making decisions on the specific measures to include and how to combine those measures, they should spend some time thinking about what they want their accountability system to accomplish.

There are lots of potential purposes for accountability. At one end, accountability is simply about transparency—providing information to the general public that would otherwise not be available. At the other end are consequences based on those results, such as rewards for high performers and sanctions against low performers. Somewhere in between transparency and sanctions is the notion that accountability can act as a tool for improvement through goal-setting, performance benchmarking, and re-evaluation. Accountability systems are the state's best tool to signal what it values and how schools should be working to improve. Not all stakeholders will value each of these purposes equally, but states must balance how their systems, or different parts of the same overall system, meet each of these broader purposes.[1]

Accountability systems are the state's best tool to signal what it values and how schools should be working to improve.

Beyond the purposes of accountability systems, there's the question of who the systems are designed for. Ultimately, *school* accountability systems are about schools and the people in them, so states should start with those people in mind. That means teachers and school leaders must value and trust the information coming out of accountability systems. The entire continuous improvement premise hinges on teachers and school leaders acting in response to the information coming out of accountability systems.

Parents are another important end-user. Parents clamor for information about schools—in 2015, the website GreatSchools.org estimated that 50 million families accessed their information on schools—so states must design systems with parents in mind as well.[2]

Each of these groups might need different levels of information delivered at different times. Teachers and school leaders need information they can act on, ideally delivered as soon as possible to inform instruction. Parents also want timely, understandable information about their own children, but they're not likely to act on it as quickly. If they're researching schools, parents want to be able to find data on students who are similar to their own children. In this case, the data need to be tailored based on specific groups of students, but there's less urgency, so the data can be synthesized over longer periods of time.

Finally, public schools are funded by public dollars, and the quality of local schools can affect everything from property values to business decisions. So taxpayers have a right to information about the schools that they're helping to pay for, not to mention an interest in ensuring those schools are as good as they can be.

ESSA compels states to address all facets of accountability.

ESSA compels states to address all facets of accountability. It requires states to create transparent school "report cards" that collect a wide array of information about schools. Some states have signaled their preference to stop here, with a "dashboard" approach where they merely release data in a single place. But parents and taxpayers need a clear, easy way to differentiate among schools. Parents make high-stakes decisions about where to live and where to send their kids to school, and a data dashboard simply throws information at them rather than trying to guide them through those choices. No parent should have to manually sort through data on schools, clicking through websites and eyeballing each individual data point, in order to get a sense of school performance.

That's why the law also asks states to make meaning of the data through simple and clear accountability systems. Each state must create a "statewide accountability system" with a "state-determined methodology" to "meaningfully differentiate" among schools and identify, at a minimum, schools that need "comprehensive" or "targeted" support. States must identify for comprehensive support schools in the bottom 5 five percent of all schools, any high school with a graduation rate below 67 percent, and schools with persistently large achievement gaps. Districts with comprehensive support schools must draft a plan to improve student outcomes. Additionally, states must also identify another category of schools called "targeted support" schools, which have large achievement gaps and need more tailored interventions. Upon identification, targeted support schools must craft their own improvement plan.

In other words, Congress didn't merely stop at transparency, and states shouldn't either. In addition to publicly reporting raw data, states should also identify low-performing schools and give them extra support to improve.

The research on school rating systems suggests they can help focus schools' attention on the students who need the most help, and that those efforts lead to improved short- and long-term outcomes.

Accountability systems are a state's best tool to accomplish all these objectives. The research on school rating systems suggests they can help focus schools' attention on the students who need the most help, and that those efforts lead to improved short- and long-term outcomes. For example, researchers Martin Carnoy and Susanna Loeb of Stanford University created an index to measure the strength of states' pre-NCLB accountability systems. They measured each state's use of high-stakes testing to reward or sanction schools and found that math scores rose more quickly for states that had stronger accountability systems between 1996 and 2000.[3]

Other researchers have exploited the differences among states to identify the influence of various types of accountability systems. Between 1993 and 2002, 43 states adopted some form of accountability system. Fourteen required schools and districts only to report their performance information ("report-card states"). Another 29 states crafted "consequential" accountability systems that also included sanctions for poor performance. Researchers Eric Hanushek and Margaret Raymond used fourth- and eighth-grade NAEP math data to compare student performance growth across states by type of accountability system (none, report card, or consequential). After controlling for key variables, including parental education, race/ethnicity, poverty and state spending on education, they found that the consequential accountability systems implemented during the 1990s had a stronger positive impact on student math performance than data reporting alone.[4]

Researchers have also found positive effects behind the mere act of notifying schools in need of improvement that they faced the potential of sanctions. For example, Thomas Ahn of the University of Kentucky and Jacob Vigdor of Duke University analyzed the impact of NCLB's accountability sanctions on school performance in North Carolina. They found that the largest improvements came from schools on the cusp of being exposed to sanctions for low performance."[5] In this case, the threat of imminent consequences was a catalyst for schools to improve. For those schools that failed to make AYP for multiple years, Ahn and Vigdor found that the threat of the "ultimate penalty"—implementation of a restructuring plan—also had a strong positive impact on test scores. Studies on accountability systems in Florida and New York City have reached similar conclusions.[6]

Perhaps more importantly, a study out of Texas found that accountability systems can have long-term benefits. In Texas, schools at risk of being identified as "Low Performing" implemented reforms that boosted the outcomes of low-performing students.[7] Those benefits included short-term gains on test scores that translated into higher college attendance rates and higher early-career earnings. In other words, well-designed accountability systems can lead schools to change their practices in ways that improve long-term student outcomes. There are very few tools at state policymakers' disposal that can boast similar outcomes.

How Should States Select Accountability Measures?

In order to maximize the potential of accountability systems, states must design those systems carefully. At their most basic, they should provide signals to schools about what society values and prioritizes. But poorly designed systems can be "gamed" in ways that have little relevance to those larger goals. Instead of focusing on higher-order skills and annual progress, schools have been encouraged to focus on lower-level skills and pushing all students through to a diploma, regardless of what they learn. The trick, then, is to design accountability systems in which schools are competing on measures that truly matter.

States should design accountability systems that are simple, clear, and fair.

In considering which measures to select, states should begin with three overarching principles. An accountability system should be:

- **Simple:** The system should be simple enough that parents can easily understand how it works, and it should be no more complicated than necessary to accomplish all of its goals.
- **Clear:** The system should provide clear signals about which schools need to improve and in what ways they need to improve to boost student outcomes.
- **Fair:** Each school should be held accountable for what it can control, and not just the type of students it enrolls.

States may be able to design accountability systems that meet these goals through a combination of measures, but ideally each measure included in the state's accountability system would meet each of these goals individually. Some potential systems or measures might meet some of these goals but not others. For example, NCLB was relatively *simple*.

After states set performance goals, any school that failed to meet its proficiency targets for two consecutive years was identified as "in need of improvement." This rule also sent a *clear* signal to schools to focus on getting all students above the proficiency benchmark. But NCLB wasn't an inherently *fair* system, because it treated all schools the same regardless of how far they fell short of the standard and how much their students progressed over the course of the year.

In addition to the state's own priorities for its accountability system, ESSA also imposes rules on the type, quality, and combination of the measures selected. On type, ESSA says that each school must be rated on at least four factors. For elementary and middle schools, those factors must include 1. Achievement rates, as measured by proficiency on the state's annual assessments; 2. Some other "valid and reliable" academic indicator; 3. Progress in achieving English language proficiency; and 4. At least one indicator of school quality or success. High schools are treated similarly, except that instead of the required other academic indicator, they must be evaluated based on graduation rates.

States can add measures beyond these four, but any measure included in state accountability systems must meet ESSA quality rules. All measures must be reported separately for all students and for each subgroup of students (economically disadvantaged students, students from major racial and ethnic groups, children with disabilities, and English learners). Additionally, the indicator(s) must "allow for meaningful differentiation" across schools and be valid, reliable, and comparable statewide.

Table 1 below looks at a variety of potential accountability indicators for elementary schools and attempts to illustrate the trade-offs for including them. For each potential indicator, it asks whether it meets the requirements under ESSA for differentiation and disaggregation and whether it meets the principles of simplicity, clarity, and fairness.

An individual school should not be held accountable for things over which it has no control.

As the table makes clear, few measures satisfy all the goals. Some indicators might provide useful context, but aren't necessarily fair measures of school quality, or might lead schools to respond in unhelpful ways. For example, while it might be important as a policy matter to address inequities in resources across schools and districts, a *school* accountability system may be a poor place to reflect that priority. For sure, a school's resources—everything from teacher salaries to curriculum to non-academic support programs—will affect the quality of education it's able to deliver, but schools have no power to tax residents, and things like teacher salaries and teacher placement policies usually are determined at the *district* level. An individual school has no control over these larger resource questions, and so they shouldn't be accountable for them. It might be important to consider how well a given school is performing with a certain level of resources, but it wouldn't make sense, for example, to hold a school principal accountable for something he or she can't change. There are ways to incorporate these components into *district* rating systems, or as contextual pieces of information on school report cards or other holistic assessments of schools, but they don't belong in school rating systems.[8]

Table 1 Considering Potential Accountability Indicators for K–8 Schools

Potential Indicator	Can it meaningfully differentiate among schools?	Can it be disaggregated at the school level for particular subgroups of students?	Is it simple to understand?	Does it provide clear guidance to schools?	Does it treat schools fairly?	Other considerations
Student achievement in reading and math	Yes	Yes	Yes	Yes	Not necessarily	Sole focus on reading and math could lead to curriculum narrowing
Progress toward English language proficiency	Yes	Will only apply to English language learners	Yes	Yes	Yes	May be difficult to define cohorts, timeframes, and what constitutes "sufficient" progress
Student growth	Yes	Yes	Maybe	Maybe	Yes	Many different models of student growth, some of which are quite complicated
Student achievement in other subjects (such as science, social studies, or writing)	Yes	Yes	Yes	Yes	Not necessarily	May alleviate curriculum narrowing, but could require students to take additional assessments
Growth of the bottom 25 percent of students	Yes	No	Yes	Yes	Yes	Intentionally double-counts low-performing students
Successful matriculation to high school	Maybe	Yes	Yes	Yes	Yes	States would need to define what "successful" matriculation means
Access to resources (curriculum, facilities, etc.)	Yes	Not easily	Yes	Yes	No	May be more applicable in a *district* accountability system
Student surveys	Maybe	Yes	Maybe	No	No	Surveys could be gamed, and results may be correlated with student demographics
Teacher satisfaction surveys	Maybe	Not easily	Maybe	No	No	Surveys could be gamed, and results may be correlated with student demographics
Student attendance	Maybe	Yes	Yes	Maybe	Yes	Student-level counts of chronic absenteeism would address many of these issues
Social and emotional learning (aka "grit" or "21st-century skills")	Maybe not[9]	Yes	No	No	Maybe	There's wide disagreement on how to measure these concepts and whether they can be taught
Student discipline, such as suspension rates	Maybe	Yes	Yes	No	Yes	The goal behind tracking this measure—keeping kids in school—may be better accomplished through a measure of attendance
School quality reviews (or "inspections")	Yes	Not easily	Yes	Maybe	Maybe	Time- and people-intensive to administer; dependent on high-quality implementation

Required by ESSA

Not required but could be used as the second academic indicator or as a school quality indicator

Not required but could be used as a school quality indicator

Other entries in the table offer contrasts and point to implementation choices. For example, many states collect information about "average daily attendance." This indicator divides the total number of students in school on a given day by the number who were supposed to be there, and then averages those results over multiple days. But while this number is relatively easy to collect, it's not that educationally meaningful. What really matters is whether individual students are showing up to school, and whether they're chronically absent from school. That's why researchers prefer measures of "chronic absenteeism," which tracks, for each student, how many days of school they miss. Unlike average daily attendance, chronic absenteeism has been linked to a host of short- and long-term academic and non-academic outcomes for students.

In selecting indicators, states should also be cognizant of any correlations or overlap. There's no added value in including two measures if they both essentially measure the same thing. In that case, states should pick the indicator that best meets the goals of simplicity, clarity, and fairness. For example, there may be overlap between measures of "grit" or "stick-to-it-iveness" and chronic absenteeism. After all, showing up every day is in itself a demonstration of the ability to persevere. States should look for potential cases of overlap like this and then opt for the indicator that best meets the goals of an accountability system. These policy choices matter in the real world and will affect people's behavior.

ESSA does not set a maximum on how many measures states should use, but the law does require states to use the measures they choose to "establish a system of meaningfully differentiating" all public schools on an annual basis. Those systems must give "substantial weight" to each of the academic indicators, and the academic indicators must be given, in the aggregate, "much greater weight" than measure(s) of school quality or success.

These decisions are complicated, and there are many choices within each option. The following sections offer some suggestions for how states should think about accomplishing the full range of policy goals through an accountability system.

How Can States Design School Rating Systems That Are Simple, Clear, and Fair?

The proposal outlined below is designed as the simplest, clearest, and fairest school rating system that any state could quickly and easily adopt. It would not require new data systems, and it does not depend on a state adopting any particular assessments, signing new contracts, or spending any new money. For simplicity's sake, the model outlined below focuses on elementary and middle schools, but the basic principles and approach could stay the same for high schools.[10]

The system starts with simple test score data—used in new ways that reward schools for improving student performance at all levels—as an initial "flag" on school performance. Those scores would be used to prioritize subsequent actions, starting with high-quality, professional, on-site school reviews. Rather than just focusing on certain tested subjects, the on-site reviews would measure school quality more holistically. Additionally, while traditional accountability systems have looked only at student test results in grades 3–8, on-site reviews would look at the quality of the entire school (see sidebar: Why States Should Include Grades K–2 in Accountability Systems).

Those on-site reviews would help guide schools in need of improvement in selecting interventions. In fact, while ESSA requires states to identify low-performing schools, it does not prescribe what the interventions in those schools should be. States have approval authority over district improvement plans for comprehensive support schools, and states

ESSA relies heavily on local will to change behavior.

must monitor and review local implementation of those plans, but ESSA relies heavily on local will to change behavior. States may add their own requirements on local interventions, but they would be acting on their own authority. Absent any such provision, it's even more important that states set up accountability systems that clearly identify areas for improvement.

In other words, ESSA's main burden for states relates to the act of identifying schools that need to improve. In this new environment, school quality reviews offer advantages over a statistical-based approach to identifying low-performing schools. On-site school inspections would provide fine-grained, actionable feedback for school leaders, give them extra capacity and justification for addressing any existing weaknesses, and help them set priorities for how to improve.

Sidebar

Why States Should Include Grades K–2 in Accountability Systems

ESSA gives states the opportunity to rethink their school accountability systems. With fewer federal rules governing those systems, states have the opportunity to start from scratch and think creatively about what those systems might look like. In doing so, they should strive to include all students in those rating systems.

NCLB required states to hold schools accountable for student achievement in grades 3–8, but that ignored the 11.2 million students in kindergarten through second grade.[i]

That policy reflected technical challenges inherent in accurately assessing learning outcomes for young children, as well as concerns about the developmental appropriateness of using standardized assessments for young children. There's a large body of scientific research on the importance of being able to read by the end of third grade, however, meaning that by the time most current school accountability systems register children's achievement and growth, students who score below grade level are already far behind—and it will be hard for them to catch up. Achievement gaps start at very young ages, and the gaps among third-graders also reflect gaps among second-graders, and so on.

Perhaps just as importantly, focusing accountability systems only on grades 3–8 creates a bad set of incentives. Schools may be inclined to shuffle their best teachers or other resources into those grades and away from earlier grades. That's counter-productive from a long-term perspective, but it may have made sense for an individual principal under pressure from existing state accountability systems.

With the expanded range of measures in ESSA, states now have a chance to go beyond federal requirements and incorporate measures of child outcomes and classroom quality that reflect how well schools are serving children in the early elementary years without imposing developmentally inappropriate assessments on them. [ii]

Continued on next page

Incorporating younger students into accountability systems does not mean states should merely project existing systems onto younger students, but there are places where measures could be replicated. For example, there's widespread interest in including attendance measures in state accountability systems. But school attendance matters as much, if not more, for younger students, so states are losing an opportunity if they fail to include students in grades K–2. Even on measures where there aren't perfect parallels, states may be able to adapt certain measures to fit younger populations. If states decided to incorporate measures of parental satisfaction, for example, they could potentially modify the surveys to include parents of all ages of students.

Most crucially, school quality reviews would provide a comprehensive analysis of school quality in all classrooms—including in K–2 classrooms. Evaluators with expertise in early childhood and elementary education could use evidence-based classroom observation protocols that reflect what research tells us quality teaching in the early grades looks like. And they would also analyze data on young children's achievement and progress. Early childhood and elementary teachers use a variety of formative and authentic assessments to monitor young children's progress—doing so is in fact a hallmark of developmentally appropriate practice in the early years. Qualified observers could analyze this data to understand whether young children are on track to reach grade-level standards by third grade, how much progress they are making, and how teachers are using information from assessments to inform and differentiate instruction. This careful analysis could inform judgments about the quality of education a school provides in the early grades, as well as recommendations for improvement. In extreme cases, reviewers could even recommend a school be identified for additional support based on the quality of its early grades programs.

The absence of federal requirements here is an opportunity. States should be smart about how to incorporate younger children into school rating systems, but states would be missing an opening if they continued to ignore these students entirely.

i U.S. Department of Education, National Center for Education Statistics, Statistics of Public Elementary and Secondary School Systems, 1980–81; Common Core of Data (CCD), "State Nonfiscal Survey of Public Elementary/Secondary Education," 1985-86 through 2011–12; and National Elementary and Secondary Enrollment Projection Model, 1972 through 2023. (This table was prepared January 2014.)

ii For more on how this could be accomplished, see Elliot Regenstein, Maia Connors, and Rio Romero-Jurado, "Valuing the Early Years in State Accountability Systems Under the Every Student Succeeds Act," the Ounce of Prevention Fund, February 2016, http://www.theounce.org/pubs/policy-pubs/Policy-Convo-05-Valuing-The-Early-Years-final.pdf

Incorporating Student Achievement

ESSA requires states to consider student achievement, and this system starts with a relatively simple performance index. Each school and district would receive a predetermined number of points based on where students fall on a performance spectrum. Higher performance levels would be worth additional points, and all schools would have an incentive to help all students reach higher and higher levels of achievement. It places significant weight on the proficiency benchmark—proficiency is, after all, a benchmark for future success in college and careers—but unlike NCLB, the sole focus of the system won't be on one narrow band of performance. NCLB required, and ESSA will continue to require, all states to define at least three performance levels, but this system works best with more frequent, smaller gradations (see sample weighting below). Many states already sort students into four or five levels.

Sample Performance Index for States with Five Performance Levels

0 points	**Level 1** (Far Below Basic)
15 points	**Level 2** (Below Basic)
35 points	**Level 3** (Basic)
70 points	**Level 4** (Proficient)
100 points	**Level 5** (Advanced)

This sort of point system is simple and clear, and it provides an incentive for all schools to care about moving all students along the achievement spectrum. What's more, it could be used at any grade level for any assessment, provided the state defined performance levels in similar ways.

A model system for elementary schools would weight any test result the same, regardless of the grade level or subject in which it was earned. That is, a third-grade math score would be worth the same amount as a fourth-grade English Language Arts score or a fifth-grade science score. If states tested in additional subjects, those could be folded in and given equal weight. (The same weights could also apply to high schools.) This is partly for simplicity's sake, but it also sends a message to schools that no particular year of school or subject is more important than any other. Schooling is continuous, and accountability systems should be as well.

In building this sort of performance index, states should use three years of data to increase year-to-year rating stability.[11] To be clear, using additional years of data would make it harder for any particular school to show progress in the short term. But that is a trade-off worth making for greater predictability and stability, because one-year results tend to be more susceptible to misleading, random fluctuations anyway, especially for small schools.[12]

Extending the timeframe would also offer greater predictability for schools and allow them to take the school improvement process more seriously. Schools would no longer have to play a waiting game until late summer or fall, when the prior year's test results were finalized. When schools aren't informed of their status until close to when the school year begins, they have little time to develop and implement improvement strategies. The three-year timeline also aligns with ESSA's requirement that all comprehensive support schools continue to implement improvement plans for at least three years regardless of any interim progress, with the expectation that short-term gains must be proven over at least three years in order to demonstrate real progress.

Using Growth as the "Other" Academic Indicator

ESSA requires states to include one other academic indicator in school rating systems. It also suggests, but does not require, that measure to be student growth, but in order to design fair accountability systems, states should follow Congress' informal recommendation. In general, disadvantaged students tend to have worse educational outcomes than their less-disadvantaged peers, and school accountability systems that fail to account for progress will, to a large extent, simply reflect incoming student advantages and disadvantages. Student growth is a fairer way to look at school quality, because it measures how much students progress under the school's care.

When NCLB was passed, states were not equipped to measure the year-to-year progress of individual students. At the time, only 13 states and the District of Columbia administered annual tests in math and only 11 states plus D.C. administered annual tests in reading.[13] Today, thanks to NCLB's annual testing requirements—which will continue under ESSA—all states have the data and capacity to measure how much students learn from year to year.

Student growth is a fairer way to look at school quality, because it measures how much students progress under the school's care.

Shifting the accountability emphasis away from a strict focus on achievement and toward a growth mindset will change the incentives for schools. Holding schools accountable for meeting one predetermined proficiency benchmark, as NCLB did and as ESSA will do if states opt not to include growth, encourages schools to focus all their efforts on students just on the cusp of the target. Adding growth into accountability systems would change that calculus, because being accountable for the growth of *all* students would force schools to pay attention to *all* kids. It would also create the need for better assessments to accurately measure student progress at all performance levels, and, thus, rely less on cheap, fill-in-the-blank tests.

There are several different ways to measure student growth. Some use complex statistical models to "control" for student characteristics and predict student scores in future years. By comparing a student's actual versus predicted score, the student earns a growth score that can be attributed to his or her teacher or school. While these models have significant appeal, they also have significant downsides. They rely on complex regressions that aren't easily understandable for parents or teachers, and they often require external vendors to run the numbers. In turn, some external vendors often have their own proprietary growth models and they're unable or unwilling to share the details of those systems with teachers or the general public.

Another growth model, called "Student Growth Percentiles," compares students against their peers. If a student scored at the 40th percentile in Year 1, this model looks at how the student scored one year later compared to all other students who also scored at the 40th percentile. If they outpace their peers, the student is considered to have made above-average growth. However, this model has to be run *after* Year 2. In the midst of Year 2, a student or teacher has no idea what it will take to make above-average growth; they're competing blind against all their peers statewide. While the percentages may be simple to understand—my child made more progress than 50 percent of kids like him—there was no way to know beforehand what that might entail. Moreover, because students are being compared to each other rather than to any sort of objective standard, students could be making strong growth while still not making sufficient progress toward proficiency.

There's another way to calculate growth that does include an objective, predetermined target for all students. Like other growth models, it would shift away from the emphasis under NCLB, which focused exclusively on students right at the cusp of proficiency—the so-called "bubble" kids. Called a "transition matrix," it builds on the Performance Index outlined above and gives students points based on whether they advance through the various performance thresholds. A federal evaluation found that a transition matrix approach identified slightly fewer students as meeting growth targets than other, more complicated growth models, but all the models identified broadly similar groups of students as making sufficient progress.[14]

While transition matrices may not be quite as precise as more complicated growth models, they offer several other advantages. As a starting bid, they expect all students to make at least one year's progress for every year they attend school. Moreover, any state could implement a transition matrix without any external support, and the calculations could be implemented on any state test. Most importantly, a transition matrix provides a clear, predetermined goal for *each student*. School leaders and teachers would know exactly where an individual student began the year and where they need to be at the end of the year to receive growth points.

More complex growth models can't make these claims. They use post-hoc statistical adjustments to determine growth calculations for students in comparison to other students like them. While these comparisons may be useful for certain purposes, they don't offer the front-end transparency, simplicity, or clarity that a transition matrix offers.

As with the Performance Index, states could choose how to allocate points in a Growth Index, but both indices would be based on the same performance levels. In the Growth Index proposed in Table 2 below, schools would only earn points when students rose to a new performance level or, in the case of students already scoring at the advanced level, maintained their performance level. This sort of Growth Index would encourage schools to diagnose the performance levels of all students at the beginning of the year and shoot for gains by the end of the year.

Table 2 Sample Growth Index

		Year 2				
		Level 1	Level 2	Level 3	Level 4 (Proficient)	Level 5 (Advanced)
Year 1	Level 1	0	70	80	90	100
	Level 2	0	0	70	80	90
	Level 3	0	0	0	70	80
	Level 4 (Proficient)	0	0	0	0	70
	Level 5 (Advanced)	0	0	0	0	70

States could adjust these weights, but the key concepts are that 1) there are potentially more points available for significantly raising the performance of lower-performing students; 2) the performance levels are clearly articulated at the beginning of the school year; 3) there's a balance between having too few performance levels, where schools might focus on only one group of students, and too many, which might not send a sufficiently clear signal to schools and teachers; 4) growth scores are worth at least as many points as the raw performance index[15]; and 5) schools should have an external incentive—beyond merely their own good intentions—to care about growth for all students.

Some states may even be tempted to focus accountability *only* on student growth. That would be a mistake. First, ESSA clearly requires states to measure and report proficiency rates. Second, growth-only systems require students to take a test that wouldn't count for accountability purposes (states would have to test all third-graders but their third-grade scores wouldn't count). That also has the effect of lowering sample sizes. Third, growth-only systems do not include mobile students who cross state lines. Depending on the state and year, five to nine percent of students can't be matched even one year later, and the mobile students falling into that category tend to be significantly lower-performing.[16]

Creating an Overall Index and Incorporating Subgroup Results

Each school would receive an Overall Index score based on an equally weighted average of the Performance and Growth Indices (0–100 scale). Next, the state would *flag* the bottom five percent of schools on the Overall Index as "comprehensive support" schools. The verb "flag" here is important, because the Overall Index is merely one step in the state's identification process, and it alone would not satisfy ESSA's requirements to incorporate subgroup results, progress toward English language proficiency, and at least one indicator of school quality or success. The next section will articulate a method to arrive at final summative ratings that reflect a broader array of measures than just reading and math scores.

Before turning to school quality indicators and final summative ratings, the next step would address ESSA's requirement to identify for targeted support schools with persistently large achievement gaps. As a check to ensure that no school ignores subgroups of students, states would calculate Overall Index scores for each subgroup of sufficient size, and any school with a subgroup performing at or below the average comprehensive support school would be initially flagged as a "targeted support" school. For example, any school where black students performed comparably to the lowest-performing schools in the state would be flagged for targeted support. The state would then repeat the process for all other subgroups. This would create a set of clear, transparent results for each subgroup.

States would repeat the process based on the state's English Language Proficiency exam. Because not all schools will have a sufficient sample size of students taking this test, it belongs as a separate check against the system rather than included as a measure in the

original Overall Index (English language learners are also a separate group included in the process outlined above). This method would also uniquely identify schools struggling to support student language acquisition, which would not be true if it were merely one component within a broader weighting system.

To be clear, this approach assumes that identifying schools for improvement is an important lever at the state's disposal. That's intentional, because there are positive effects associated with the mere act of notifying schools that they need to improve. That's especially true for accountability systems bearing consequences for schools, but it's even true in systems relying purely on information and transparency. For example, researchers Eric Hanushek and Margaret Raymond found that schools alter their practices in response to accountability pressure, albeit sometimes in unintended and unfortunate ways.[17] The state's role, then, is to ensure its accountability system creates the right set of incentives for schools to respond in productive ways. The next section turns to how states can best facilitate that process.

The state's role is to ensure its accountability system creates the right set of incentives for schools.

Incorporating Other Measures of School Success Into Final School Ratings

After using test scores as an initial "flag," no school's rating would be final until they completed a formal, on-site school quality review. The process outlined above would help the state prioritize formal, on-site inspections of school quality. Schools would be given no more than a few days' notice of their on-site visit, but all comprehensive support schools would receive one in Year 1 under ESSA (the 2017–18 school year), all targeted support schools would receive one no later than Year 2, and the remainder of schools by the end of Year 3. This would ensure that all schools received a formal review over a three-year period, but the state would prioritize attention to comprehensive and targeted support schools.

The school quality reviews would be based off the school inspectorate model used in England. As Craig Jerald described it in a 2012 report, "inspectors observe classroom lessons, analyze student work, speak with students and staff members, examine school records, and scrutinize the results of surveys administered to parents and students."[18] Although the interviews provide context, the main focus is on observations of classroom teaching, school leadership, and the school's capacity to improve. New York City, Charlotte-Mecklenburg School District in North Carolina, turnaround schools in Massachusetts, and some charter school authorizers employ similar processes, as do various "school support teams" sponsored by many other states and districts.

School quality reviews would also allow schools to be judged holistically on a wide variety of measures rather than a numeric formula.

School quality reviews would also allow schools to be judged holistically on a wide variety of measures rather than a numeric formula. In Britain, in addition to objective test score data, inspectors assess schools on observed student behaviors and responses on things like

"the extent to which pupils feel safe" and "the extent to which pupils develop workplace and other skills that will contribute to their future economic well-being"; the effectiveness of teaching practices through measures like the "use of assessment" and "the extent to which the curriculum meets pupils' needs"; and the quality of school leadership, through indicators like "the effectiveness of leadership and management in embedding ambition and driving improvement" and "the effectiveness with which the school deploys resources to achieve value for money." The inspections agency develops a comprehensive rubric to measure these things and employs a staff of full-time professionals who are trained to observe them in schools. This approach is radically different than most American accountability systems, and for most states it would represent a much more rigorous way to measure multiple facets of what a school is supposed to do.

Like those in England, the quality reviews would be conducted by professionals trained in rating, evaluating, and providing feedback to schools. States could choose to employ the reviewers directly as an apolitical auditing body, or it could opt to contract out the evaluations to third-party providers. Either way, states would fund the reviews out of their Title I budgets. ESSA allows states to set aside up to seven percent of their federal Title I funds to create a "statewide system of technical assistance and support" for local school districts. As of fiscal year 2017, ESSA authorizes Title I funds of just over $15 billion. Seven percent of that amount would be $1.05 billion, well above the low-cost estimate provided by Craig Jerald in "On Her Majesty's School Inspection Service." With state administrative funds from Title II, federal funds would surpass even Jerald's high-cost estimate. While $1 billion may sound like a lot of money, it would amount to less than 0.2 percent of the $600 billion we currently spend on K-12 public schools. And diverting a slice of federal funds for these purposes means that states or school districts would not be required to allocate any of their own dollars toward the process.

In addition to the raw dollar figures, Jerald estimates that states would need roughly 8,000 trained reviewers to perform the on-site visits, split among 1,000 full-time staff and 7,000 contractors. This is no small lift, particularly if multiple states adopt this approach and are competing to find talented professional reviewers. Their salaries, at least, would be covered in the budget figures above, but states would still need to hire and train a new class of workers. States could start slowly by focusing their first-year efforts on the smaller subset of comprehensive support schools, and states could attempt to recruit reviewers from the ranks of retiring Baby Boomer teachers who might be convinced to lend their expertise to other schools in their state.

Other than comprehensive support schools, which by law must remain as comprehensive support schools for at least three years, all other schools would receive a final summative rating based on their school quality reviews (which factor in test scores and achievement gaps as part of their review) within three years. Their temporary ratings would remain until at least their first formal review. In this way, the reviews would serve as exit criteria for schools to earn their way off state identification lists.

Conclusion and Considerations

After a decade under the No Child Left Behind Act and several more years of large-scale waivers from NCLB, states now have the time and the opportunity to design their own accountability systems. There are reasons to be concerned about how that might go. The easiest thing for states to do will be to simply tweak their existing accountability systems into a sort of NCLB-lite. But we know what will happen under this approach. When states identify "low-performing" schools only by where their students end up, they're much more likely to pinpoint places where students enter school further behind—whether because of weaknesses in their prior schools or lack of access to quality early learning before starting school. Continuing to place too much emphasis on particular grades and subjects will detract from other societal goals for public schools.

States have also struggled to make the leap from identifying schools for improvement to actually helping them to improve. NCLB was not particularly nuanced on this front. It forced states to identify for improvement any school where any subgroup of students was under-performing, but then it applied the same list of interventions regardless of why the school was identified.

States have struggled to make the leap from identifying schools for improvement to actually helping them to improve.

The system outlined above would represent a break from that past. It would still use test scores, which remain one of the cheapest, most reliable ways to determine student performance on a grand scale. But states should use those test scores in smarter ways, make sure they're capturing progress over time, and rely on test results merely as a flag for further investigation. This paper proposes intensive, on-site school quality reviews as a way to accomplish all the other objectives of accountability systems.

While ESSA's requirements fit more neatly with numeric school rating systems, there are reasons to think school quality reviews could (or at least should) pass muster with federal officials. For example, there is evidence that school quality reviews can provide a reasonable amount of differentiation across schools. As of the most recent data from England, 18 percent of primary schools were rated "Outstanding," 67 percent were rated "Good," 14 percent rated "Requires Improvement," and 1 percent were rated "Inadequate."[19] This spread suggests reviewers are able to tell the difference between a great school and one that's merely mediocre, and one that may have a few flaws versus one that needs more wholesale reform. As in England, states would need to periodically evaluate their results to ensure they were sufficiently differentiating across schools.

School quality reviews could also be tailored to investigate how schools were working for particular groups of students. This "disaggregation" would look different under a qualitative system like on-site reviews, but even the British model includes a separate indicator for "outcomes for individuals and groups of pupils." Applied here, states would expand that indicator and disaggregate school quality *as received* by subgroups of students. If, for example, the reviewers observed that Hispanic students were disproportionately assigned to classes where low-quality instruction occurred, that would register in the report as an issue the school should address. Similarly, the on-site reviews already include student interviews, and in this context they could over-sample particular groups of students for those conversations.

ESSA's timeline also poses challenges for states interested in using school quality reviews to satisfy the law's school quality indicator requirement. ESSA requires states to meaningfully differentiate among schools annually, beginning in the 2017–18 school year. States could not reasonably expect to conduct high-quality, formal, on-site reviews every year for every school, and states without a system already built wouldn't be ready to use it by the fall of 2017. But states could use a temporary measure, such as chronic absenteeism, while they scale up their review process.

In addition, ESSA is scaling back on NCLB's requirement for annual school improvement plans and annual interventions. School quality reviews fit better with ESSA's rule requiring three years of interventions in low-performing schools, and the school quality reviews would set up schools well to actually design and implement thoughtful improvement plans. That's because school quality reviews would serve as a tool both for accountability and improvement. Rather than low-quality, self-completed improvement plans that were common under NCLB, school quality reports would be executed by professionals trained to provide an honest review of school quality coupled with guidance on how to improve.

There's evidence that well-designed school quality reviews can identify low-performing schools and help them improve.

Perhaps most importantly, the research on Britain's inspection system suggests it is a valid and reliable way to identify and support low-performing schools. There's evidence, for example, that the school review process is effective at identifying low-performing schools. After being identified for improvement, schools responded in ways that boosted student achievement across a range of subjects, and the gains were on par with interventions like enrolling in a high-performing charter school or allowing students to transfer schools.[20]

In other words, well-designed school quality reviews can identify low-performing schools and help them improve. That evidence should satisfy the intent of ESSA, because, ultimately, that's what accountability systems should be designed to achieve.

Endnotes

1 The federal role in accountability has also shifted over time. The 1994 reauthorization of the Elementary and Secondary Education Act of 1965 relied mostly on transparency, whereas NCLB applied a more coercive accountability system. ESSA attempts to do both—it preserves the transparency requirements and requires states to have accountability systems, but leaves states broad discretion on what those systems look like.

2 Bill Jackson and Peter Cunningham, "Guest Commentary: Parents Have a Right to Know About the Public Schools," *East Bay Times*, February 28, 2015, http://www.eastbaytimes.com/opinion/ci_27599977/guest-commentary-parents-have-right-know-about-public

3 Martin Carnoy and Susanna Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24 (2002), No. 4, pp. 305–331, https://cepa.stanford.edu/sites/default/files/EEPAaccountability.pdf

4 Eric A. Hanushek and Margaret E. Raymond, "School Accountability and Student Performance," Federal Reserve Bank of St. Louis *Regional Economic Development 2*, no. 1 (2006): 51–61, http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BRaymond%202006%20RegEconDevelopment%202(1).pdf

5 Thomas Ahn and Jacob Vigdor, "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina," National Bureau of Economic Research, Working Paper 20511, September 2014, http://econ.msu.edu/seminars/docs/ahnvigdornclb_uva.pdf

6 Jay P. Greene, "An Evaluation of the Florida A-Plus Accountability and School Choice Program," Manhattan Institute for Policy Research, February 2001, http://www.manhattan-institute.org/html/cr_aplus.htm; Marcus A. Winters, "Grading Schools Promotes Accountability and Improvement," Manhattan Institute for Policy Research, May 2016, https://www.manhattan-institute.org/html/grading-schools-promotes-accountability-and-improvement-evidence-nyc-2013-15-8912.html

7 David J. Deming et al., "When Does Accountability Work?," *Education Next* 16 (2016), No. 1, http://educationnext.org/when-does-accountability-work-texas-system/

8 In other words, *districts* are responsible for within-district funding inequities, not schools. Relatedly, districts should not be held accountable for state-level funding inequities. If the state is delivering inequitable resources based on regressive funding formulas, it wouldn't make sense to hold districts accountable for cross-district spending disparities.

9 Many measures of grit suffer from reference bias, where students compare themselves to peers at the same school, see: Angela L. Duckworth and David Scott Yeager, "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes," Educational Researcher 44 (May 2015), No. 4, pp. 237–251.

10 For more on high school accountability systems, see Chad Aldeman, "Mind the Gap: The Case for Re-Imagining the Way States Judge High School Quality," Bellwether Education Partners, July 2015.

11 To accommodate schools attempting to show improvements, states could weight current-year results more heavily than older data, but that would come at the cost of simplicity and potentially undercut some of the benefits of additional years of data.

12 For a detailed look at how one year of results unfairly harms small schools, see Thomas J. Kane and Douglas O. Staiger, "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," in Diane Ravitch, ed., *Brookings Papers on Education Policy* (Washington, D.C.: Brookings Institution, 2002), pp. 235–283.

13 Abigail Potts, Rolf K. Blank, and Andra Williams, "Key State Education Policies on PK-12 Education: 2002," Council of Chief State School Officers, 2002, http://programs.ccsso.org/content/pdfs/KeyState2002.pdf

14 See Chapter IV, "Final Report on the Evaluation of the Growth Model Pilot Project," U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, , Washington, D.C., 2011.

15 This could also be accomplished through a back-end statistical manipulation, but states should attempt to limit backdoor elements as much as possible.

16 See Exhibit 56 from "Final Report on the Evaluation of the Growth Model Pilot Project," available at: http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf.

17 Eric A. Hanushek and Margaret E. Raymond, "Lessons about the Design of State Accountability Systems" (prepared for the "Taking Account of Accountability: Assessing Policy and Politics" conference, Cambridge, MA, Jun 9–11, 2002), accessed May 19, 2015, http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BRaymond%202003%20NCLB.pdf

18 For more information, see Craig D. Jerald, "On Her Majesty's School Inspection Service," Education Sector, 2012, http://educationpolicy.air.org/sites/default/files/publications/UKInspections-RELEASED.pdf

19 Sir Michael Wilshaw, "The Annual Report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2014/15," Ofsted (London: The Stationery Office Limited, 2015), Figure 25a.

20 See Rebecca Allen and Simon Burgess, "How Should We Treat Under-Performing Schools? A Regression Discontinuity Analysis of School Inspections in England," Working Paper No. 12/287, Centre for Market and Public Organisation, March 2012, http://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/wp287.pdf; and Iftikhar Hussain, "Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections," *The Journal of Human Resources* 50 (2015), No. 1, pp. 189–221, http://jhr.uwpress.org/content/50/1/189.abstract

Acknowledgments

I presented earlier versions of this paper at the Thomas B. Fordham Institute's ESSA Accountability Design Competition and at a working meeting hosted by Education Counsel. Thanks to the hosts of those opportunities as well as to all who provided feedback. Those presentations challenged my thinking in important ways and helped me shape the final product. In addition, I would like to thank Christy Hovanetz, Craig Jerald, Sara Mead, Kaitlin Pennington, Elliot Regenstein, Ryan Reyna, Andy Rotherham, and Scott Sargrad for various pieces of oral and written feedback. Thanks as well to Super Copy Editors and Five Line Creative for copy editing and design support. Finally, I'd like to thank the Bill & Melinda Gates Foundation for providing funding for this paper. As always, the findings and conclusions are mine alone.

About the Authors

Chad Aldeman

Chad Aldeman is a Principal at Bellwether Education Partners and the Editor of TeacherPensions.org. He can be reached at chad.aldeman@bellwethereducation.org.

BELLWETHER
EDUCATION PARTNERS
IDEAS | PEOPLE | RESULTS

About Bellwether Education Partners

Bellwether Education Partners is a nonprofit dedicated to helping education organizations—in the public, private, and nonprofit sectors—become more effective in their work and achieve dramatic results, especially for high-need students. To do so, we provide a unique combination of exceptional thinking, talent, and hands-on strategic support.