



Creating Measurement and Accountability Systems for 21st Century Schools: A Guide for State Policymakers

David Osborne
October 2016

About the Author

David Osborne is director of the Progressive Policy Institute's project on Reinventing America's Schools. He is the author or co-author of five books on public sector reform, including the *New York Times* bestseller, *Reinventing Government*. Osborne has also authored numerous articles for the *Washington Post*, *The Atlantic*, the *New York Times Magazine*, *Harpers*, *U.S. News*, *Education Week*, *Education Next*, *The New Republic*, *Governing*, and other publications. His book on education reform will be published in 2017.

Osborne has advised governments large and small, from cities, counties, and school districts to states, federal agencies, and foreign governments. In 1993, he served as a senior advisor to Vice President Gore, to help run what the Vice President often called his "reinventing government task force," the National Performance Review. Osborne was the chief author of the NPR report, which laid out the Clinton Administration's reinvention agenda, called by *Time* "the most readable federal document in memory."

Contents

| | |
|--|----|
| EXECUTIVE SUMMARY | 4 |
| THE ESSA'S NEW RULES | 12 |
| ACCOUNTABILITY SYSTEMS THAT TREAT DIFFERENT SCHOOLS DIFFERENTLY | 14 |
| MAKE ACCOUNTABILITY REAL BY REPLACING FAILED SCHOOLS WITH PROVEN MODELS | 17 |
| ACCOUNTABILITY IS NOT THE SAME AS MEASUREMENT | 17 |
| STATES SHOULD NOT RELY SO HEAVILY ON TEST SCORES | 19 |
| GIVE MORE WEIGHT TO ACADEMIC GROWTH THAN ACHIEVEMENT LEVELS | 22 |
| STOP FOCUSING SO HEAVILY ON THE PERCENTAGE OF STUDENTS WHO ARE PROFICIENT OR ABOVE | 23 |
| WHAT ELSE SHOULD STATES USE IN ACCOUNTABILITY SYSTEMS? | 24 |
| IMPORTANT MEASURES NOT YET READY FOR PRIME TIME | 27 |
| AN IDEAL STATE RATING AND ACCOUNTABILITY SYSTEM | 32 |
| TO GRADE OR NOT TO GRADE? | 35 |
| PRESENTING THE DATA TO THE PUBLIC | 36 |
| ENDNOTES | 39 |

Executive Summary

Because Congress passed the Every Student Succeeds Act (ESSA) last December, states are revamping their federally required systems to measure school quality and hold schools accountable for performance. But most are doing so using outdated assumptions, holdovers from the Industrial Era, when cookie-cutter public schools followed orders from central headquarters and students were assigned to the closest school. Today we are migrating toward systems of diverse, fairly autonomous schools of choice, some of them operated by independent organizations. Before revising their measurement and accountability systems, states need to rethink their assumptions.

For instance, most states have assumed that they should apply one standardized, statewide accountability system to almost all public schools. Most have also assumed that measurement and accountability systems are roughly the same thing, so the only aspects of performance they need to measure are those in their federally-required accountability systems. Under the old No Child Left Behind (NCLB) Act, most of those measures were standardized test scores, and what counted was the percentage of students scoring "proficient" or better. When schools repeatedly failed to meet such standards, most states assumed the proper response was some

minor form of restructuring required by NCLB—perhaps a new principal, perhaps some new teachers, perhaps some new money.

None of these assumptions will produce the schools our children need in the 21st century. NCLB was an important step in its time, institutionalizing an expectation that states would hold schools accountable for the learning of all their children, including the poor, minorities, and those with special needs. But it relied on the fairly blunt tools used by most states back in 2001: primarily achievement scores on standardized math and reading tests. In the intervening 15 years, more tools have become available—and even more are the subject of intense research today. Fortunately, the ESSA has opened the door to these new approaches.

We need diverse schools that cultivate the joy of learning, engage their students in deep learning, and help them develop the “character skills”—such as conscientiousness and self-control—that lead to success in life.

If we want 21st century schools, we must hold them accountable to 21st century standards. For too long, we have defined and measured school quality in a way that encourages cookie-cutter schools, all focused on preparing students for tests. Instead, we need diverse schools that cultivate the joy of learning, engage their students in deep learning, and help them develop the “character skills”—such as conscientiousness and self-control—that lead to success in life.

To achieve this will require a series of fundamental changes:

We need accountability systems that focus on more than minimal standards and treat different schools differently. It still makes sense for states to adopt minimum standards that will trigger consequences for most schools if they consistently fail to meet them. After all, we want every child to learn to read, reason, do basic math, write coherently, and gain some familiarity with science, technology, history, geography, and civics. If children are not learning these things, should we really be using taxpayers' dollars to fund their schools?

We must always remember, though, that these are *minimum standards*. Beyond them, states should encourage districts and authorizers to negotiate more specific performance goals and measures that reflect the missions of individual schools. If a school is designed to provide STEM (science, technology, engineering, and math) education, for instance, it should be judged on how well it does so. If it is designed to provide dual language immersion, or career and technical education, its performance measures should reflect those goals.

Such performance agreements can motivate every school, whereas minimum state standards have little effect on schools whose students regularly score above the

minimums. Performance agreements also encourage people to open innovative schools designed to serve different kinds of students. When we apply standardized accountability to all schools, we do just the opposite—and we get far less innovation.

We need systems that make accountability real and powerful by replacing failed schools with proven models that fit students' needs. NCLB allowed states and districts to impose consequences for failure that had few teeth, and the ESSA could well make the problem worse. But if states want to improve student learning, real accountability is one of their most powerful tools. Experience has proven that the most effective way to turn around a failing school is to replace it—to bring in an entirely new team with a strong track record and a new vision for the school. This helps motivate staff at every school: when everyone in a school knows that their jobs will disappear if students are not learning enough, they usually work together to make it happen, no matter the obstacles.

In national charter school studies, the states and cities with high performing charters are those that actively monitor quality and close or replace schools whose students are falling behind. Places that failed to do this (until recently), such as Arizona, Texas, and Ohio, had charters that performed no better, and sometimes worse, than traditional public schools on standardized tests.

State measurement systems should be broader than accountability systems.

People often forget the distinction between measurement and accountability, but it is critical. Accountability systems create consequences for school performance: both rewards and penalties. Measurement systems provide information about those schools, without consequences attached. Both are necessary, but they are hardly identical.

Accountability systems create consequences for school performance: both rewards and penalties. Measurement systems provide information about those schools without consequences attached. Both are necessary, but they are hardly identical.

There are many things we measure about schools—and some we should begin measuring—because the information is useful. Some information helps parents make better choices, such as the number of Advanced Placement courses a high school offers. Other information also helps districts, authorizers, and schools learn what works and what doesn't—such as data on parental

involvement. But statewide accountability systems should focus on *outcomes* we care about as a society, while leaving schools free to choose the methods that work best for their students and teachers.

Statewide accountability systems should put only half their weight on test scores.

Since the mid-1990s, our state accountability systems have been dominated by test scores. NCLB, which required states to hold their schools accountable for delivering “proficiency” on standardized tests, intensified the problem. The U.S. Department of Education gave all but a handful of states waivers to NCLB, to measure student growth as well as proficiency levels. But by 2016, according to the Center for American Progress, the average state gave test scores (achievement and growth combined) 91 percent of the weight in elementary and middle school ratings and about 70-75 percent in high schools. Those numbers are far too high.

Don't get me wrong: Test scores are important measures of success. Without them, how will we know if students are learning to read and do math? Beginning with sixth grade, we should also test writing (tested by at least five states by 2015), science (at least 13 states), and the social sciences (a majority of states). If we don't measure these things, how will we know which schools are failing and need to be replaced?

But that doesn't mean we should rely on test scores for three quarters or more of what matters. Testing experts agree that scores bounce around from year to year, so we need to be careful how we use them. Relying so heavily on test scores creates myriad problems.

Our accountability systems should emphasize student growth more than achievement levels.

Standardized tests usually measure the level at which a student performs, not the gains he or she has made over the past year. Yet we cannot judge the performance of a school or teacher without the latter data. If a school's students arrived four years behind grade level, on average, and two years later they are only one year behind grade level, is the school failing? Of course not. This may have been NCLB's biggest flaw: it required states to measure students' achievement levels but not their rate of growth. And despite their waivers, the majority of states still give greater weight to proficiency than growth in their measurement systems—and in making decisions about intervening in low performing schools. This puts schools with high percentages of low-income students at a huge disadvantage, because their test scores are lower.

States should quit using “proficient” as the only target.

Under NCLB, states were also required to measure the percentage of students who reached some cut score, usually labeled “basic” or “proficient.” To make their schools look better, too many states lowered the proficiency bar. And as more than 40 experts in testing argued in a recent letter to Education Secretary John B. King Jr., the NCLB approach had myriad other flaws. It failed to distinguish between students who were right at the cut score and those who were far above it, for one. It gave no credit to gains made by students who remained below the cut score, no matter how large. And it created incentives for schools to concentrate on raising the scores of those just below the

cut score, the “bubble kids,” as they became known. This led to neglect of both the lowest and highest achieving students.

If we hold schools accountable for the growth of *all* their students, in contrast, they will be pressured to help those who are far behind while also providing challenging material for their advanced students. Hence states should use average scores or proficiency indexes, which reflect the percentage of students who reach each level of the scoring system.

States should construct their systems as works in progress, to be adjusted as they learn how to objectively and reliably use student surveys and measure deeper learning, character skills, and other important aspects of school performance. We should experiment with all of these aspects of measurement—to learn what works, where the pitfalls lie, and how to overcome them. If some authorizers and districts want to include them in their charters and performance agreements—and the schools in question agree—states could also learn from their experience.

AN IDEAL STATEWIDE ACCOUNTABILITY SYSTEM

I would suggest that today's statewide rating systems—those applied to all schools—have five or six basic elements, weighted roughly as follows. (The balance between achievement and growth should depend on which method states use to measure growth; with some value-added methods, achievement and growth can be combined in one value-added score.) ESSA requires that states also use English learner progress toward proficiency, but I have not specified a recommended weight because that should vary by school. In some, with many English language learners, it would be quite important; in others, with none, it would be unimportant.

For high schools:

- Student academic achievement: 20 percent
- Student academic growth: 25 percent
- English learners' progress toward proficiency: variable
- Student engagement: 10 percent
- Qualitative school assessments by experts: 15 percent
- Student outcomes: 25 percent

Elementary and middle schools would use only five elements:

- Student academic achievement: 20 percent
- Student academic growth: 30 percent
- English learners' progress toward proficiency: variable
- Student engagement: 10-20 percent
- Qualitative school assessments by experts: 25-30 percent

Indicators for each element could include the following:**Academic achievement and growth:**

- Test scores in math, ELA, writing, science, and the social sciences
- For English language learners, scores on tests designed to measure their progress in learning English
- PSAT, SAT, and/or ACT scores and/or state-approved international test scores
- Industry certifications

Qualitative assessments

- Expert site visit assessments

Student engagement

- Parent surveys


Outcomes (for high schools)

- HS Graduation rate: 4 year, 5-7 year, and with GED (perhaps 2, 2, and 1 percent for each category, respectively)
- Quality of diploma, if states offer different diplomas
- Percent of graduates enrolling in college
- Percent of enrollees required to take remedial classes in college
- Percent of college enrollees persisting to second and third years
- Percent of two-year college enrollees completing a two-year degree or credential
- Percent of non-college-bound graduates employed, in training, or in the military
- Income levels for those employed full-time

Ideally, states should give various weights to each indicator and sum them to give a grade for each area, plus an overall grade. Some states use colors or phrases, such as “meets expectations,” but parents understand grades on an A-F scale better. If we use grades for our children, we should have the courage to use them for the adults who run our schools.

The kind of system I have described—with school performance agreements in addition to minimum statewide standards that include qualitative assessments by experts—may seem like a fantasy to those steeped in the world of NCLB. But they already exist. In Massachusetts, charter schools must meet minimal state standards, but their charters also include specific goals. When their charter is up for renewal, a team of experts visits for a day and a half and writes a qualitative assessment report on the school, which the state board uses in making its decision. In Washington, D.C., the Public Charter School Board uses a performance framework much like I have advocated, but individual charters include goals specific to the schools and reviews include multiple site visits to assess the quality of schools. Denver Public Schools does much the same thing with its charters. In New Orleans, charters must meet minimum state standards, and both the Recovery School District and the Orleans Parish School Board do on-site reviews every year, plus a high stakes review when charters are up for renewal.

All three cities are among the fastest improving in the nation, and Massachusetts has one of our highest performing charter sectors. Are these not the kinds of results we want for all our children?



Creating Measurement and Accountability Systems for 21st Century Schools: A Guide for State Policymakers

OCTOBER 2016

David Osborne

In the wake of Congress's passage of the Every Student Succeeds Act (ESSA) last December, states are busy revamping their federally-required systems to measure school quality and hold schools accountable for performance. But most are doing so using outdated assumptions: holdovers from the Industrial Era, when cookie-cutter public schools followed orders from central headquarters and students were assigned to the closest school. In today's world, that is no longer the norm. We are migrating toward systems made up of diverse, fairly autonomous schools of choice, some of them operated by independent organizations, as charter, contract, or innovation schools. Before revising their measurement and accountability systems, states need to rethink their assumptions.

For instance, most states have assumed they should apply one standardized, statewide accountability system to almost all public schools. Most have also assumed that measurement and accountability systems are roughly the same thing, so the only aspects of

performance they need to measure are those in their federally-required accountability systems. Under the old No Child Left Behind (NCLB) Act, most of those measures were standardized test scores, and what counted was the percentage of students scoring "proficient" or better. Finally, when schools repeatedly failed to meet such standards, most states assumed the proper response was some minor form of restructuring required by NCLB—perhaps a new principal, perhaps some new teachers, perhaps some new money.

None of these assumptions will produce the schools our children need in the 21st century. NCLB was an important step in its time, institutionalizing an expectation that states would hold schools accountable for the learning of all their children, including the poor, minorities, and those with special needs. But it relied on the fairly blunt tools used by most states back in 2001: primarily achievement scores on standardized math and reading tests. And by setting the target at "proficiency" and ignoring

student growth, it penalized any school full of low-income children. In the intervening 15 years, more tools have become available—and even more are the subject of intense research today. Fortunately, the ESSA has opened the door to new approaches.

THE ESSA'S NEW RULES

Under the ESSA, states are still required to give greater weight to test scores than other indicators in their measurement systems, but they have significantly more leeway. The new law requires states to measure 1) student performance in math and English language arts, or ELA, 2) a second academic indicator, such as growth in math and ELA, 3) the progress English language learners make toward proficiency in English, 4) graduation rates from high school, and 5) at least one other measure of school quality or student success. (States can require schools to measure other things as well, of course.)

States have to publish the results (excluding number 3 above) for each school and for these subgroups at each school: students with disabilities, students from low-income families, students from major racial and ethnic groups, and English language learners. In rating schools, states must give “substantial weight” to categories 1-4 above and “much greater weight” to those four combined than to number 5. Their rating systems must sum the scores in these categories to create one rating or score for each school.

For Title 1 schools in the bottom five percent, plus high schools with four-year graduation rates below 67 percent, states must provide “comprehensive support and intervention”—

though the act leaves it mostly up to the states to determine what that entails.¹ (Title 1 schools are those in which at least 40 percent of students qualify for a free or reduced-price lunch.) If a subgroup at a school falls in the bottom five percent of scores for all students at Title 1 schools, the state must provide “targeted support and intervention” to the school. Unfortunately, the ESSA's requirements for determining schools needing both forms of support are unworkable and need amendment or liberal interpretation by U.S. Department of Education regulations, which are still being finalized.²

We are migrating toward systems made up of diverse, fairly autonomous schools of choice, some of them operated by independent organizations, as charter, contract, or innovation schools. Before revising their measurement and accountability systems, states need to rethink their assumptions.

States may set “alternate academic achievement standards” for students with the most significant cognitive disabilities and give them alternate tests, provided no more than one percent of all students in the state are assessed this way. This accommodation will thus apply to less than 10 percent of students who receive special education. States may also allow multiple student assessments through the year rather than one year-end test, and they may include student portfolios in the assessments. Up to seven states will be allowed to pilot

competency-based assessments and other innovations.

ESSA requires that this framework be applied to all schools and 99 percent of students, but it leaves it largely up to states to define how they will use the measures to create consequences, within broad guidelines. In other words, it dictates at least part of a state's measurement and rating systems but not the rest of its accountability system.

Some states may use their new freedom to avoid holding schools accountable for performance. This would be a huge mistake. Experience has shown that creating real consequences for school performance, including replacement by another school, is a key component of high-performing systems. In national charter school studies, the states and cities with high-performing charters are those that actively monitor quality and close or replace schools whose students are falling behind. Places where

charter authorizers did not do this until recently, such as Arizona, Texas, and Ohio, had charters that performed no better, and sometimes worse, than traditional public schools on standardized tests.³

In revising their accountability systems, states should learn from the dramatic success of emerging 21st century systems, such as those in New Orleans, Washington, D.C., Denver, Indianapolis, and Boston. New Orleans, where all but six public schools are charters, has improved faster than any other district in the state—and probably the nation as well.⁴ Washington D.C., where 45 percent of students attend charters, has improved faster on the National Assessment of Educational Progress (NAEP) than all 20 other big cities that participate, with charters leading the way.⁵ (New Orleans does not participate in NAEP). Denver Public Schools, which has embraced charters and added an equally large sector of “innovation schools,” which it treats somewhat like charters, has moved from the



slowest academic growth among Colorado's 20 largest cities in 2005 to the fastest since 2012.⁶ Indianapolis's 37 charters, authorized by the mayor, vastly outperform its traditional public schools.⁷ The traditional district has now embraced "innovation network schools," some of which are charters and all of which enjoy charter-like autonomy. And Boston's charter sector, which benefits from exemplary authorizing by the state board of education, is the highest performing in the nation, producing more than twice as much learning per year as Boston Public Schools.⁸

If we want 21st century schools, we have to hold them accountable to 21st century standards. For too long, we have defined and measured school quality in a way that encourages cookie-cutter schools, all focused on preparing students for tests. Instead, we need diverse schools that cultivate the joy of learning, engage their students in deep learning, and help them develop the "character skills"—such as conscientiousness and self-control—that lead to success in life.

As I will argue below, we need:

- accountability systems that focus on more than minimal standards and treat different schools differently;
 - systems that make accountability real and powerful by replacing failed schools with proven models that fit students' needs;
 - state measurement systems that are broader than their accountability systems;
 - accountability systems that put only half their weight on test scores and emphasize student growth more than achievement levels; and
- systems that are constructed as works in progress, to be adjusted as we learn how to objectively and reliably measure deeper learning, character skills, and other important aspects of school performance.

ACCOUNTABILITY SYSTEMS THAT TREAT DIFFERENT SCHOOLS DIFFERENTLY

The common belief that the same accountability system should be applied to every public school is outdated. In a traditional 20th century district, in which all schools operated in similar fashion and sought to educate all types of children, it may have made sense. But in 21st century systems, with parents choosing between many diverse schools designed to educate different kinds of learners, it no longer does.

It still makes sense for states to adopt minimum standards that will trigger consequences for most schools if they consistently fail to meet them. After all, we want every child to learn to read, reason, do basic math, write coherently, and gain some familiarity with science, technology, history, geography, and civics. If children are not learning these things, should we really be using taxpayers' dollars to fund their schools?

We must always remember, though, that these are only *minimum standards*. Beyond them, states should encourage districts and authorizers to negotiate more specific performance goals and measures that reflect the missions of individual schools. If a school is designed to provide STEM (science, technology, engineering, and math) education, for instance, it should be judged on how well it does so. If it is designed to provide dual language immersion, or career and technical education, its performance measures should reflect those goals.

Should alternative schools for dropouts and overage students be held to the same standards as ordinary high schools? How about schools designed specifically for students with disabilities? Or schools for students returning from the criminal justice system? Obviously, we need different standards for such “alternative” schools. While states should measure and rate these schools, just as they do all others, they should not impose consequences. Instead, districts or authorizers should hold alternative schools accountable for performance goals and measures they have negotiated with each individual school—or group of schools, if more than one use the same model.

We must be careful, however, that districts and authorizers don’t label too many schools “alternative” and dump their worst students in them to make the rest of their schools look good. State departments of education should take a close look at any district or authorizer that has more than 10 percent of its students in alternative schools.

After all, we want every child to learn to read, reason, do basic math, write coherently, and gain some familiarity with science, technology, history, geography, and civics.


Holding each school accountable to its own standards has always been the heart of the charter model—though not all states are faithful to that model. A charter should be a performance contract, which spells out what the school intends to accomplish, how it will be measured, and what will happen if the school

fails to achieve its goals. Such agreements can reflect the unique characteristics of a school far better than statewide standards can. And when schools are held accountable to their own goals, negotiated with their own authorizer or district, their leaders and staff members are far more likely to embrace responsibility for accomplishing them.

Such performance agreements can also motivate every school, whereas minimum state standards have little effect on schools whose students regularly score above the minimums. Our accountability systems should motivate every member of every school to seek improvement in student outcomes—even those who work in schools for gifted students.

Finally, by crafting different performance goals for different kinds of schools, we will encourage people to open innovative schools designed to serve different kinds of students. When we apply standardized accountability to all schools, we do the opposite—and we get far less innovation.

Consider University Preparatory Academies in inner-city Detroit. When founder Doug Ross opened his first high school, in 2003, he adopted a model heavily influenced by Big Picture Learning—particularly the MET School, based in Providence, Rhode Island, Big Picture’s first. Aware that his biggest challenge was creating motivation for college among inner-city, African-American teenagers who had rarely met anyone, other than their teachers, who had ever been to college, Ross decided that every high school student would spend two days a week in internships with local businesses, public agencies, or non-profits. It worked: when students saw African-American adults who had graduated from college and had good jobs, nice houses, and nice cars, a light bulb often went off.

A collection of school supplies including a pencil holder with colored pencils, paper clips, and sticky notes. The background is a dark, textured surface. The text is overlaid on a semi-transparent white box.

If we want 21st century schools, we have to hold them accountable to 21st century standards. For too long, we have defined and measured school quality in a way that encourages cookie-cutter schools, all focused on preparing students for tests.

On top of that, students discovered they could contribute in meaningful ways, which boosted their confidence, and some of them fell in love with particular fields of work. But when Michigan adopted statewide standards and imposed them on all schools, prompted by NCLB, University Prep had to cut internships back to half a day a week, to ensure that students covered all the material on state tests.

MAKE ACCOUNTABILITY REAL BY REPLACING FAILED SCHOOLS WITH PROVEN MODELS

Another mistake NCLB made was allowing states and districts to impose consequences for failure that had few teeth. The ESSA could well make the problem worse. But, if states want to improve student learning, real accountability is one of their most powerful tools. As noted above, experience has proven the most effective way to turn around a failing school is to replace it—to bring in an entirely new team with a strong track record and a new vision for the school. This helps motivate staff at every school: when everyone knows their jobs will disappear if students are not learning enough, they usually work together to make it happen, no matter the obstacles. This is one of the key things that separates strong charter states from weak ones.⁹

When schools fail to meet state standards or their more specific performance goals for two or three years, states should provide resources to help them hire the help they need to turn things around, as the ESSA suggests. If schools cannot do so within two years, however, authorizers and districts should replace them with more effective school operators, as they do in New Orleans, D.C., Denver, and other places. Before pulling the trigger, they should give the schools a chance to

appeal—to provide compelling evidence that they are actually succeeding, given the students they educate. Sometimes a school may have a high percentage of children with serious disabilities, or a high percentage of former dropouts or homeless children. Sometimes a middle school or high school may be helping children who arrived three or four years behind grade level achieve decent academic growth, while missing minimum state standards.

In some cases, if it's a close call, the district or authorizer might want to renew the school for only a year or two, to give it time to solve whatever problems exist. If the schools' argument is not compelling, however—or if it fails to turn things around during its extension—authorizers should replace the failing school with a better one.

ACCOUNTABILITY IS NOT THE SAME AS MEASUREMENT

People often forget the distinction between measurement and accountability, but it is critical. Accountability systems create consequences for school performance: both rewards and penalties. Measurement systems provide information about those schools, without consequences attached. Both are necessary, but they are hardly identical.

There are at least three kinds of accountability. The first is well known: measurement by the state and consequences of some kind if performance is exceptional or falls below minimum standards for several years in a row. The second is also relatively familiar: consequences (either positive or negative) imposed by parents who react to performance information by changing their children's schools,

taking public dollars with them. The third is less powerful but also important: the pride or embarrassment administrators and teachers feel when their schools are shown to be of high or low quality. All three can stimulate school leaders and their staffs to work together to improve student outcomes.

Accountability systems should focus on what we as a society most want our schools to accomplish: real-world outcomes such as graduation rates, college-going and employment rates, and student acquisition of knowledge and skills.

There are many things we measure about schools—and some we should begin measuring—because the information is useful. Some information helps parents make better choices, such as the number of Advanced Placement courses a high school offers. Other information also helps districts, authorizers, and schools learn what works and what doesn't—such as data on parental involvement, or student surveys about teacher and school quality, or student and teacher assessments of character skills such as self-control and conscientiousness. Such data may even play a role in school, district, and authorizer decisions about where to invest and what policies to adopt. But there are good reasons to keep it out of an accountability system.

Accountability systems should focus on what we as a society most want our schools to accomplish: real-world *outcomes* such as graduation rates, college-going and employment rates, and student acquisition of knowledge

and skills. But to help those who manage schools, districts, and charter networks—as well as parents facing choices—we also need information about *inputs* and *outputs*, such as attendance rates, teacher absenteeism and turnover, student-teacher ratios, numbers of Advanced Placement courses, and so on. Indeed, every school should track its own “balanced scorecard,” including data about student results, employees' views and experiences (morale, absenteeism, turnover, etc.), operational issues (spending, learning time, productivity, etc.), and customers' views (parental satisfaction, student engagement, and the like). Principals and teachers should examine such data in regular group sessions and use it to make changes that will improve performance.

If schools or networks of schools choose to use data about such things for internal accountability, that is their prerogative. Districts and authorizers may want to include any of these measures in performance agreements with particular schools, where it makes sense. But states should limit what they hold *all* schools accountable for to a handful of key outcomes that truly matter in all students' lives. And they should give schools the freedom to figure out the best methods to achieve those outcomes, given the particular students they educate. This is the formula that has worked so well in our fastest improving school systems.

Here are a few examples of data that states should require districts and authorizers to measure but *not* include in statewide accountability or rating systems:

Student demand. For most schools, demand reflects parental judgments about the school's quality. But some schools are designed for specialized populations—pregnant students,

dropouts, overage students, or special education students. It would be silly to punish such schools because demand was low or dropping, since lower pregnancy or dropout rates might be a sign of success, not failure, for the city or school system. However, districts and authorizers should feel free to include negotiated goals about demand in their performance agreements with individual schools, where it makes sense.

Retention rates. Some districts and authorizers measure the rate at which schools retain students, but this is another number that should not be attached to ratings or consequences. Some demanding schools lose students because nearby schools are easier. It would be insane to punish them for that.

Discipline rates. The same is true of rates of discipline. How could a statewide standard ever apply to every public school? How could a state agency judge whether a school was using the ideal type and amount of discipline? As charter authorizers in D.C. and elsewhere have shown, publishing data on discipline rates is useful—to keep schools honest, to encourage them to deal with the trauma that often underlies student misbehavior, and to nudge them to use methods such as restorative justice rather than suspensions and expulsions. But we need to leave judgments about discipline up to the people who run schools. Students in one school may disrupt class frequently, so high rates of discipline may be required to ensure that students can learn uninterrupted. Students in another school may rarely disrupt class and thus need little disciplinary action. Any effort to punish schools for high rates would undermine their ability to deal with the realities in their classrooms.

College-level courses. Some states, districts, and authorizers also give credit in their

performance frameworks for the number of Advanced Placement (AP) classes, International Baccalaureate (IB) programs, and/or high school students taking college classes through dual enrollment. Again, this is good information to measure and publicize, because it will help parents and students make informed choices. But how can anyone say that all schools should offer more such opportunities? Big Picture Learning Schools have concluded that internships in local businesses, nonprofits, and government offices are more valuable for their students. Given their impressive outcomes, they are surely correct. It would be silly to create incentives for them to limit internships and offer more AP courses. Different schools, with different kinds of students, need different methods. Statewide accountability systems should focus on outcomes and leave the choice of methods to schools.

STATES SHOULD NOT RELY SO HEAVILY ON TEST SCORES

Since the mid-1990s, our state accountability systems have been dominated by test scores. NCLB, which required states to hold their schools accountable for delivering “proficiency” on standardized tests, intensified the problem. President Obama’s Department of Education gave all but a handful of states waivers to NCLB, to measure student growth as well as proficiency levels. But by 2016, according to a study by the Center for American Progress, the average state gave test scores (achievement and growth combined) 91 percent of the weight in elementary and middle school ratings and about 70-75 percent in high schools.¹⁰ Those numbers are far too high.

Don’t get me wrong: Test scores are important measures of success. Without them, how will

we know if students are learning to read and do math? Beginning with 6th grade, we should also test writing (tested by at least five states by 2015), science (at least 13 states), and the social sciences (a majority of states).¹¹ If we don't measure these things, how will we know which schools are failing and need to be replaced? How will we make sure most poor children don't end up in dropout factories?

But that doesn't mean we should rely on test scores for three quarters or more of what matters. Testing experts agree that scores bounce around from year to year, so we need to be careful how we use them.

Relying so heavily on test scores creates myriad problems. One of the most important, articulated by social scientist Donald Campbell in 1976, has become known as Campbell's Law: "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." We have seen both corruption and distortion due to NCLB's accountability system: adults cheating on standardized tests and schools concentrating on rote learning to drive up test scores, undermining children's natural love of learning in the process.

That does not mean we should *stop* standardized testing. Campbell was not discouraging *measurement*; he was warning us not to rely on a single measure of quality. "Many commentators, including myself, assume that the use of multiple indicators, all recognized as imperfect, will alleviate the problem,"¹² he added in the very same essay.

In addition, we all know people who perform well in life and work but who do not test well—

because of stress, learning disabilities, trauma, or myriad other issues. As one teacher told David Kirp, author of *Improbable Scholars*, about Union City's schools in New Jersey, "The expectation is that 10-year-olds can write five paragraphs in half an hour, solve complicated math problems and have a wealth of knowledge about science, and do it all entirely on their own. But these youngsters freeze up under stress—one word on a question can throw them off...."¹³

One of my daughters taught in a K-8 school in New Orleans, where students had to achieve certain test scores to move from 4th to 5th grade and 8th to 9th. The school gave teachers rubber gloves along with the tests, because students often threw up from the stress. Fights were always more common during testing week, because students wanted to be suspended and sent home so they could avoid the test.

Yet another problem is that standardized tests often give misleading signals about students who are still learning English. Kristina Rizga, author of *Mission High*, describes a student at San Francisco's Mission High School named Maria, from El Salvador. For years, Maria failed standardized tests because she still struggled with English. But, by the time she applied for college, she had mastered the language, and she won two scholarships and was accepted at five, including the University of California at Davis.

Some studies find a correlation between good test scores and success in adult life, but others find no connection, and the question is still hotly debated.¹⁴ In 2012, Education Sector compared college enrollment rates at 21 randomly selected California high schools with their state Academic Performance Index (API) scores, which were based almost entirely on test scores. "In the sample of 'typical' schools," they found, "most

high-scoring API schools also tend to have higher postsecondary enrollment and most low-scoring API schools have lower postsecondary enrollment. But, in the sample of high-poverty schools, the relationship between high API scores and high college enrollment rates all but disappears."¹⁵

They described San Francisco's June Jordan School for Equity, "a 250-pupil school founded in 2003 to serve some of the city's poorest neighborhoods," which had a "dismal" API in 2010 of only 568 out of 1000. (The state considered 800 a good score.)

But also that year, June Jordan ranked second among San Francisco high schools in the percentage of students eligible for the UC/CSU system, behind only the prestigious, admissions-based Lowell High School. Among its 2009 graduates, 70 percent enrolled in college overall, and 49 percent enrolled in four-year colleges—higher enrollment rates than the district average. The graduates are also persisting in college. Fully 100 percent of its 2007 graduates who entered two-year colleges re-enrolled for a second year, and 85 percent of its 2008 graduates did so. For graduates entering four-year colleges, the figures are 100 percent for 2007 grads and 91 percent for 2008.

Standardized tests can also push schools to concentrate more on memorization than on deeper learning. But tests can also measure aspects of deeper learning, and those developed to measure the Common Core standards have improved the situation.¹⁶ Most experts believe the Common Core tests move in the right direction, though some argue they still need more questions that show how deeply students

can apply, analyze, and evaluate what 'they know.'¹⁷

Schools that focus deliberately on deeper learning often sacrifice their standardized test scores, because they don't prepare their students for such tests. For instance, a dozen teacher-run charter schools in Minnesota—operated or assisted by a teacher cooperative called EdVisions—use project-based learning to maximize student engagement. According to a 2010 study, they had lower scores on standardized tests than the state average but higher ACT scores (23.6 compared to a national average of 21.2) and SAT scores (1749 compared to a national average of 1518). More than 82 percent of their graduates entered two- or four-year colleges, compared to a national average of 68 percent.¹⁸

Standardized tests also fail to measure "non-cognitive" or "character" skills that are important for future success, such as self-control, conscientiousness, and the ability to work well with others. Former Minnesota Governor Rudy Perpich, who, in the 1980s, brought public school choice to Minnesota—and hence to America—used to say, "I've seen too many people who passed tests and failed life. And too many people who failed tests and passed life."¹⁹

Finally, too heavy a reliance on standardized testing in accountability systems can discourage people from creating schools for particularly challenging students, such as dropouts, children with disabilities, those convicted of crimes, or those who don't speak English. They also discourage schools from trying new methods—such as project-based learning, student internships, or career and technical education—that might deepen learning but hurt test scores. We need *more* innovation in our schools, not



less; we need to encourage people to start schools that are unique, aimed at students who do not fit well in cookie-cutter schools.

The American people understand this. Every year Gallup and Phi Delta Kappa collaborate on a survey to measure public opinion about education. In 2015, 64 percent of those surveyed agreed there was too much emphasis on standardized testing in their public schools. When asked about different measures of school effectiveness, almost four in five said “how engaged students are with their classwork” and “the percentage of students who feel hopeful about their future” were “very important,” and 69 percent added high school graduation rates. More than twice as many said the percentage of graduates going on to college or jobs was “very important” than said the same about standardized test scores. But 48 percent agreed tests scores were “somewhat important” in improving schools, while another 19 percent said they were “very important.” Less than a

third said they were not important or “not very important.”²⁰

In other words, Americans see the value in standardized testing, but they see more value in measuring graduation rates, employment rates, and student engagement and attitudes.

There is wisdom here. I believe that standardized test scores (including college readiness tests such as the ACT and SAT exams) should be given roughly half the weight in statewide measurement and accountability systems, depending on the school level. (I will suggest specific percentages below, after discussing some of the alternatives.) But they should be balanced by other important measures.

GIVE MORE WEIGHT TO ACADEMIC GROWTH THAN ACHIEVEMENT LEVELS

Standardized tests usually measure the level at which a student performs, not the gains he or she has made over the past year. Yet we cannot judge the performance of a school or teacher without the latter data. If a school's students arrived four years behind grade level, on average, and two years later they are only one year behind grade level, is the school failing? Of course not. This may have been NCLB's biggest flaw: it required states to measure students' achievement levels but not their rate of growth. This put schools with high percentages of low-income students at a huge disadvantage, because their test scores were so much lower.

Under waivers granted by the Bush and Obama administrations, all but five states added growth measures in English and math. But the majority of states still give greater weight to proficiency than growth in their measurement systems—and

in making decisions about intervening in low-performing schools.²¹

There are many different ways to measure student growth. Jurisdictions using student growth percentiles—a popular method first developed by Colorado—need to balance them with equal weight for proficiency. Because these systems only compare students to those who have had similar test scores in the past, a student can score above average while falling further behind grade level, because his “peer group” performs so poorly.²²

Another popular option is a “value-added model,” which attempts to isolate the contribution a school makes to student gains by controlling for student characteristics such as socioeconomic status—thus putting all schools on an equal playing field. One popular value-added model is the pioneering method developed by William L. Sanders in the early 1990s for Tennessee, the Tennessee Value-Added Assessment System (TVAAS), a version of which the firm SAS now markets as the “EVAAS Multivariate Response Model.”²³ Another is a two-step value-added model advocated by a group of economists at the University of Missouri. This method first performs a value-added analysis, then compares each school to schools and teachers serving students with similar characteristics—income levels, race, English language ability, and so on. Whereas student growth percentiles and some single-step value-added models produce results in which schools with higher-income students have an advantage, this two-step model appears to eliminate any such advantage.²⁴

There are a few other alternatives used by states,²⁵ but my purpose here is not to analyze which methods are superior. It is simply to argue

that any academic performance scores must focus more on growth than on proficiency levels.

STOP FOCUSING SO HEAVILY ON THE PERCENTAGE OF STUDENTS WHO ARE PROFICIENT OR ABOVE

Under NCLB, states were required to measure the percentage of students who reached some cut score, usually labeled “basic” or “proficient.” To make their schools look better, too many states lowered the proficiency bar. And, as more than 40 experts in testing argued in a recent letter to Education Secretary John B. King Jr., the NCLB approach had myriad other flaws. It failed to distinguish between students who were right at the cut score and those who were far above it, for one. It gave no credit to gains made by students who remained below the cut score, no matter how large. And it created incentives for schools to concentrate on raising the scores of those just below the cut score—the “bubble kids,” as they became known.

Focusing only on proficiency also led to neglect of both the lowest and highest achieving students.

Focusing only on proficiency also led to neglect of both the lowest and highest achieving students. In contrast, if we hold schools accountable for the growth of all their students, they will be pressured to help those who are far behind while also providing challenging material for their advanced students. There are a surprising number of the latter: A recent study by the Institute for Education Policy found that 20-40 percent of elementary and middle school students perform at least one grade level above

their grade in reading, while 11-30 percent do so in math. That equates to 1.4 to two million students in California alone.²⁶

NCLB gave schools incentives to ignore such students, and some of them did. According to the Fordham Institute:

Research has demonstrated that students just below the bar were most likely to make large gains in the NCLB era, while high achievers made lesser gains. Those most victimized by this regime were high-achieving poor and minority students—kids who were dependent on the school system to cultivate their potential and accelerate their achievement. (Equally able youngsters from middle-class circumstances have other people and educational resources to keep them moving forward.)²⁷

In their letter, the experts urged the Education Department to write regulations allowing states to use either of two approaches under ESSA: 1) average scores for each grade and subject, or 2) proficiency indexes, which would reflect the percentage of students who reached each level in the scoring system—not just one level, “proficiency.” Either would give a truer picture of school performance than proficiency rates.²⁸

The Fordham Institute recommended the latter, an “achievement index that gives schools partial credit for getting students to ‘basic,’ full credit for getting students to ‘proficient,’ and additional credit for getting students to ‘advanced.’”²⁹ It noted that eight states already used this kind of index, under NCLB waivers.³⁰ It also recommended that states add “high-achieving students” as a subgroup on which they report results, just like the other subgroups required by ESSA.

Some states also use growth models that only include “growth to proficiency”—hence excluding all students who are already above proficient levels.³¹ Schools that know they are being judged this way tend to ignore the needs of those advanced students. Leslie Jacobs, a veteran of almost two decades on New Orleans’s and Louisiana’s school boards, puts it this way: “If you don’t measure it, it doesn’t count. And if kids don’t count, they will be ignored.”

WHAT ELSE SHOULD STATES USE IN ACCOUNTABILITY SYSTEMS?

Graduation rates. NCLB required states to measure graduation rates, so most use a four-year adjusted cohort graduation rate, which includes all those who start 9th grade at a high school but subtracts those who transfer to another high school. Some states have included five-year, six-year, and even seven-year rates.³² This is wise, because we want high schools to work hard to help students graduate, even if it takes longer than normal. At least a third of those who enroll in college are not ready—they have to take remedial courses, and many of them later drop out.³³ Some of the nation’s best charter schools—which want their graduates to be ready for college—require students to do an extra year if they have not fulfilled all graduation requirements in four years. We should reward such behavior, not punish it. Extended year graduations should receive equal weight with four-year graduations; there should be no assumption that one is better than another.

Some states award special diplomas to recognize high achievement—“distinguished achievement” programs in Texas,³⁴ for example, or “regents diplomas with advanced designation with honors” in New York.³⁵ Some of them award

points in their performance indexes for the number of high-achievement diplomas.³⁶ Florida, Indiana, Louisiana, Maryland, New Mexico, Oklahoma, and New York City also give credit for students who have earned industry-based certifications, to incent schools to make such options available.³⁷ Both are excellent ideas.

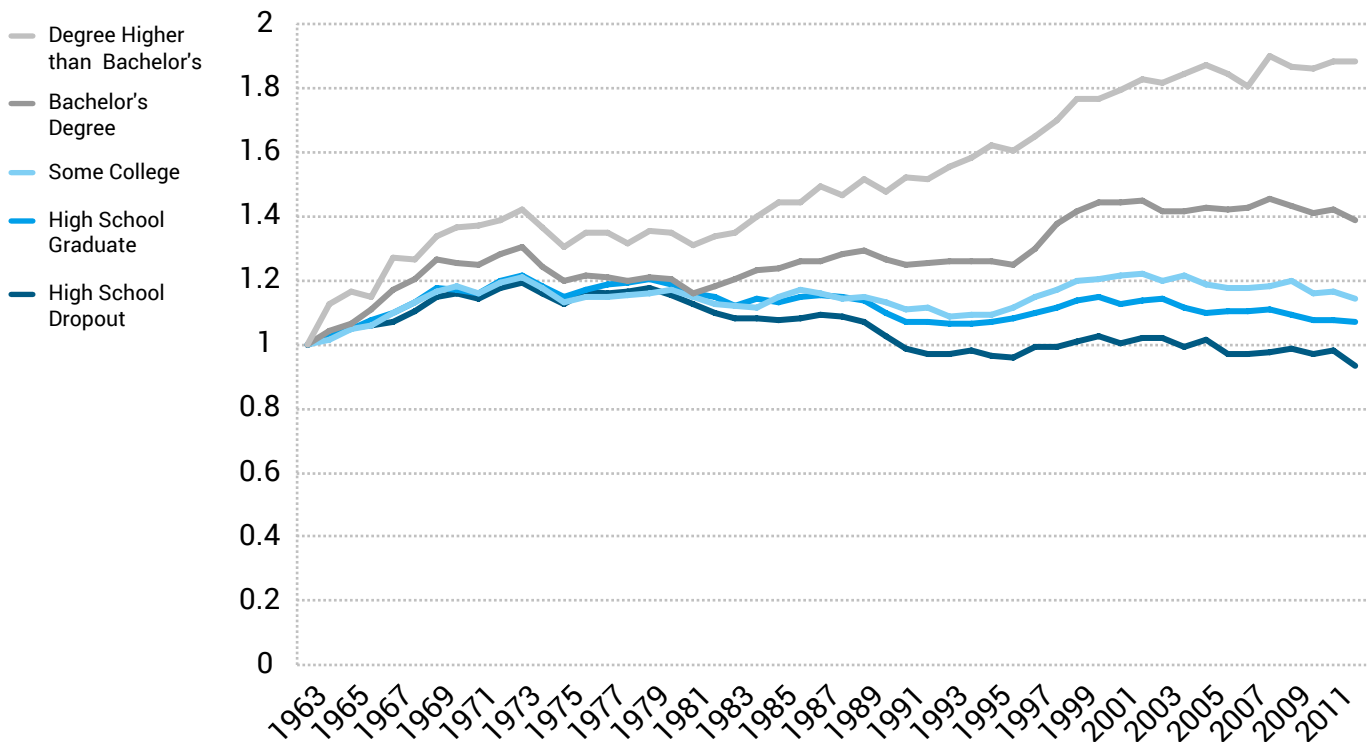
College entrance and persistence rates.

Ultimately, the most important factor in judging schools should be how they prepare students for success in life. Graduating from college is often an important milestone on that path. Indeed, the value of a college degree has increased steadily in the Information Age, while the value of a high school degree has fallen, as the graph below shows. (The graph focuses on male earnings,

but the data for female workers looks much the same.) For most, college has become the gateway to a middle-class career.

Hence many charter authorizers and a few states include the percentage of graduates who enter college in their measurement systems.³⁸ In addition, Denver measures the percentage of college entrants who must take remedial classes, an important indicator everyone should use.³⁹ The percentage who move on to a second year, the percentage entering a two-year college who earn a two-year degree or credential, and the percentage moving on to a third year are also worth measuring. We should give high schools an incentive to actively help their graduates make it through college, as many of the best charters

FIGURE 1: Change in Real Wage Levels of Full-Time Male Workers by Education, 1963-2012



Source: David H. Autor, "Skills, Education, and the Rise of Earnings Inequality Among the 'Other 99 Percent,'" *Science*, 23 May 2014: 344 (6186), 843-851, Figure 6a.

do. Unfortunately, using college graduation rates is probably unfair, because college graduation occurs so long after high school graduation. A school might make huge strides in helping its graduates prepare for and get through college, but the improvement would take four to six years to show up in graduation rates.

Training and employment rates. There are many skilled positions in our economy that do not require college degrees; instead, they require some technical training, whether through an apprenticeship, a training program, or a community college. And many students have no desire to go to college. For some, finding a full-time job (or joining the military) after high school is an indicator of success. States should measure employment and training rates (including military service) for graduates who do not go on to college and include them in their accountability systems. They should also include the income levels of recent graduates (from the past two years) who are employed full time. They should be careful not to put so much weight on college enrollment that they give high schools an incentive to ignore career and technical education that leads to further training, apprenticeships, and jobs. These outcomes should be given substantial weight as well.

Qualitative assessments by experts. In England, small teams of experts, many of them former school leaders and teachers, visit each school roughly every three years, with two days notice, and spend two to three days gauging its quality. They sit in on classes; examine student work; talk with groups of students, staff, and members of the governing board; look over documents, records; and test scores; review parent surveys; solicit written input from parents; and often meet with parents. Then they publish reports—distributed to all parents—full of qualitative

judgments, rating areas such as “quality of teaching, learning and assessment,” “outcomes for pupils,” “early years provision,” “effectiveness of leadership and management,” and “personal development, behavior and welfare.” There are four possible ratings: outstanding, good, requires improvement, and inadequate. The work is overseen by the Office for Standards in Education, Children’s Services and Skills, an independent government agency created in 1992.

New York City adopted a similar model to evaluate its own schools; Vermont is doing the same; Massachusetts, Indianapolis, and others use visits based on the English model in reviewing their charter schools; and large charter networks such as KIPP have done the same.⁴⁰ This approach can yield valuable information about school quality that test scores don’t reveal—particularly about school culture and aspects of deeper learning, such as critical thinking, problem solving, researching, and speaking skills.

Obviously, inspections such as this are more expensive than standardized tests, though they don’t have to be done every year. If we want a balanced set of quality measures that reflect the whole child’s experience, however, they are indispensable. We already spend a lot accrediting most public schools every six to ten years (depending on the region), paid for by the schools. It is a voluntary process that relies heavily on a “self-study” by the school, and, in some regions, it focuses less on academic outcomes than on facilities and process (guidance services, curriculum, instructional model, etc.). Whether school leaders implement the accrediting agency’s recommendations depends entirely on them, and accrediting agencies rarely refuse a public school accreditation. As one former superintendent

who has been through multiple accreditations and also done charter reviews put it, "If we're not keeping score, what are we doing?"⁴¹

Ted Sizer, the late headmaster of the elite Phillips Academy, author of *Horace's Compromise*, founder of the Coalition of Essential Schools, and chair of the Brown University Education Department, told me that the assessment done when he ran a charter school, at the end of his career, was far more valuable than any accreditation process he had ever been through. In some regions, what we spend on accreditation would be more productively spent on an English-style assessment of each school every three years. If it costs more than accreditation every six years, which is unlikely, the investment would still be worthwhile. Accountability is an arena in which we should never be penny wise and pound foolish.

In large states, the scale of these qualitative assessments might require be phased in over multiple years. But organizations already exist that know how to do them, and there are plenty of retired teachers and administrators who would be happy for the part-time work. With 53 million people, England is far larger than any state, and the English government has managed just fine for decades. New York City is larger than most states, and it has done likewise. Since states are highly unlikely to embrace this method all at once, there would rarely be more than one large state launching such a system at any given time.

Student engagement, measured by parent surveys.

In most industries, customer satisfaction is a key indicator of quality. Education is no exception: both parents and students have important perspectives on the quality of their schools. There is some risk in

using student surveys, as I will discuss below. But there is less risk with parent surveys, because parents are more likely to express their true feelings about their children's schools and less likely to comply with principals' and teachers' wishes.

Attendance rates are another measure of student engagement, but they are easily manipulated by schools and difficult to audit effectively.

IMPORTANT MEASURES NOT YET READY FOR PRIME TIME

Many critics of standardized testing rightly point out that tests cannot capture many valuable aspects of student learning. For instance, a good test can capture some aspects of what people call deeper learning, but not all. Some schools and districts use other methods, but they are difficult to standardize when applied to many schools. For these and the two other forms of assessment below, we should experiment, moving as fast as possible to develop reliable, objective measures. States should treat their measurement and accountability systems as works in progress, to be improved as we learn more.

Qualitative assessments of student performance.

New Hampshire is gradually moving to system in which students advance by proving their competence in a subject matter, rather than putting in seat time and passing courses. In 2015, the state received a federal waiver to work with eight school districts to develop performance tasks to assess student learning, in place of standardized tests. The pilot, the Performance Assessment of Competency Education (PACE), is designed to develop

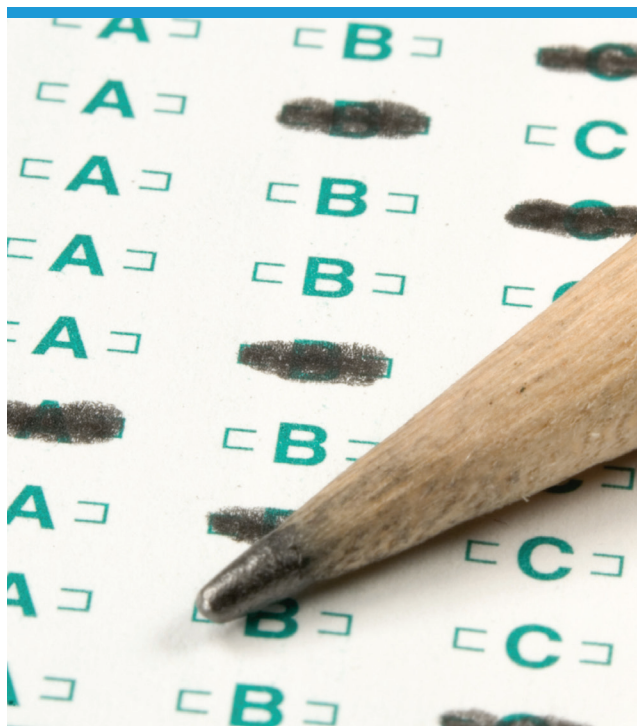
assessments more in line with competency-based education. Some grades take the state's standardized (Smarter Balanced) tests, but others are assessed with multi-step tasks that seek to measure deeper learning. For instance, geometry students at Spaulding High School were asked to design two water towers that would each hold about 45,000 cubic feet of water—one a simple solid and one a compound solid—with the least amount of construction materials possible, then to write a proposal recommending the best approach.⁴²

States should treat their measurement and accountability systems as works in progress, to be improved as we learn more.

Deeper learning assessments are common in other developed countries, but they're just beginning to spread in the U.S.⁴³ The Council of Chief State School Officers' Innovation Lab Network is working with schools and districts in a dozen states to assess Common Core skills through "performance-based measures of deeper learning."⁴⁴ And the New York Performance Standards consortium, a group of 38 small high schools that focus on deeper learning, has a state waiver allowing them to use continuous, project-based assessments. Students write essays and research papers, solve math problems, do science experiments, and orally present their work to external assessors. All but two of the schools are in New York City. With demographics similar to the rest of the city's high schools, their dropout rates are half what the city's are and their college acceptance rates are almost 30 percentage points higher.⁴⁵

Unfortunately, all these approaches involve subjective assessment of student work, and it is difficult to ensure that such assessments are standardized across thousands of schools. Other nations, including the Netherlands, Singapore, and parts of Australia, train teachers and education professors who serve on assessment panels, using common rubrics. The results are audited to ensure that roughly the same standards are being applied in all schools.⁴⁶ If research on these efforts can find a method that proves objective and reliable when used statewide, it would enrich our accountability systems. For now, though, all we can do is encourage such research.

Student surveys. In the U.S., most colleges use student surveys as part of professors' evaluations, but K-12 schools rarely do. Yet, when they respond to surveys honestly, students have proven to be very accurate barometers of teacher and school quality. Starting in 2010, the Bill and Melinda Gates Foundation funded research involving 3,000 teachers in six urban school districts to identify the most accurate measures of teaching quality. They analyzed test scores, videotaped 20,000 teacher lessons, studied thousands of classroom observations of teaching, and implemented dozens of student surveys. They asked tens of thousands of students to respond to statements on a survey developed through a decade of research by Dr. Ronald Ferguson of Harvard University, in collaboration with teachers, students, and colleagues.⁴⁷ The survey asked students to agree or disagree with statements like, "My teacher knows when the class understands, and when we do not"; "When I turn in my work, my teacher gives me useful feedback that helps me improve"; and "In this class, the teacher accepts nothing less than our full effort." Students'



answers correlated with how much they had learned (measured by growth on standardized tests). Indeed, they were more reliable than ratings by trained observers who watched videos of classrooms. And they bounced around less than test scores, over time.⁴⁸

At least 100 districts and a thousand schools are now using the Tripod survey, as Ferguson and his company call it, as part of their teacher evaluation systems. Some are attaching consequences. In Pittsburgh, PA., for instance, survey results collected multiple times over two years account for 15 percent of a teacher's overall evaluation score.⁴⁹ Tripod Education Partners has also developed and used student surveys to deliver feedback on "school climate," "student engagement," "peer support," and character skills such as "conscientiousness."

If surveys are used to create real consequences, however, schools and teachers may push students to emphasize the positive. Some

students may also use their answers to punish teachers who are more demanding or tougher graders—a phenomenon well known to college professors. According to Ferguson's partner, Rob Ramsdell, Tripod finds very consistent patterns across classrooms, even in districts that give 15 percent of the weight in teacher evaluations to surveys. If students were not answering honestly, he says, this would not be the case. But he agrees we can't rule out the possibility of such problems. If you've ever bought a new car and had the salesperson tell you to expect a phone call from the company asking you to rate his performance—and that his *entire* bonus rests on your answer—you're familiar with one potential problem.

With standardized tests, states analyze data and investigate instances where the unexpected happens, to detect cheating. They could do the same with student surveys. Normally, survey results are fairly consistent over time, and they correlate with students' academic gains on tests. When surveys suddenly show a big change or more positive results than tests, states could investigate.

For now, every school should use feedback from its customers—its students and their families—to improve its performance. If the data is trustworthy, there are few better measures of school quality. States should require that the data be collected and distributed to schools and parents, to help them choose appropriate schools for their children. But, given the risk of schools influencing student responses—or students punishing demanding teachers—we should not yet include student surveys in statewide accountability systems. Instead states should experiment, to learn what works, where the pitfalls lie, and how to overcome them. If some authorizers or districts want to include

survey results in their charters and performance agreements—and the schools agree—states could also learn from their experience.

Finally, if states use qualitative assessments by experts, as the English do, assessment teams could include parental and student survey data in their analyses. If the data looked fishy, they could discount it and ask the state, district, or authorizer to investigate.

Assessments of non-cognitive skills. In recent years scholars have focused increasing attention on character skills not measured by standardized tests, such as self-control, persistence, and conscientiousness. These are also known as non-cognitive skills, social-emotional skills, or habits of success. Both common sense and academic research suggest they are extremely important in determining whether a student will succeed in later life. In a recent paper, *Transforming Education's* Chris Gabrieli, Dana Ansel, and Sara Bartolino Krachman recounted the research findings:

In the Dunedin Multidisciplinary Health & Development Study [in a city in New Zealand], 95 percent of the young people in the top quintile of self-control were likely to graduate from high school, compared with 58 percent for those in the lowest quintile and about 80 percent for those in the next two quintiles. In James Heckman's 2006 analysis of the [U.S.] National Longitudinal Survey of Youth from 1979, non-cognitive factors were as equally predictive as cognitive factors in accounting for which young men earned a college degree by age 30. In the Fast Track longitudinal study, kindergartners with high social competency were 1.5 times more likely to graduate from high school and twice as likely to graduate

from college. Among the Dunedin Study cohort, those from the lowest quintile of self-control in their elementary school years were more than three times as likely as those in the highest quintile of self-control to ever have been convicted of a crime (43 percent versus 13 percent).⁵⁰

The authors summed up the research results this way:

Academics

- 1. Non-cognitive skills predict high school and college completion.*
- 2. Students with strong non-cognitive skills have greater academic achievement within K–12 schooling and college.*
- 3. Fostering non-cognitive skills as early as preschool has both immediate and long-term impact.*

Career

- 1. Employers value non-cognitive skills and seek employees who have them.*
- 2. Higher non-cognitive skills predict a greater likelihood of being employed.*
- 3. Stronger non-cognitive skills in childhood predict higher adult earnings and greater financial stability.*

Well-Being

- 1. Adults with stronger non-cognitive skills are less likely to commit a crime and be incarcerated.*
- 2. Strong non-cognitive skills decrease the likelihood of being a single or unplanned teenage parent.*
- 3. The positive health effects associated with stronger non-cognitive skills include*

*reduced mortality and lower rates of obesity, smoking, substance abuse, and mental health disorders.*⁵¹

Educators have long understood the importance of character skills, of course. A 2013 national teacher survey found that 93 percent agreed it was important for schools to promote these skills, while 88 percent said their schools were already trying to do so.⁵² Indeed, schools have long graded students' conduct, and increasingly they create incentives for good behavior and penalties for poor behavior. Many charter schools focus heavily on helping students develop a series of values. Summit Public Schools, a charter network in California and Washington state, asks students and their teachers to fill out surveys on students' habits of success twice a year, and mentors then initiate discussions about them with their students.⁵³

According to Ted Dintersmith and Tony Wagner, authors of *Most Likely to Succeed*, employers also recognize how important social-emotional skills are:

Google, for instance, changed its hiring strategies after Laszlo Bock, senior vice president of people operations, analyzed their data and found no correlation between job performance and an employee's GPA, SATs, or college pedigree. Google now considers an applicant's ability to collaborate and to perform authentic job-related challenges. Now, they hire many new employees who never went to college.

Our education goals have lost touch with what matters most—helping students develop essential skills, competencies, and character traits. It's time to reimagine the goals for U.S. education, and hold all

*schools—from kindergarten through college—accountable for teaching the skills and nurturing the dispositions most needed for learning, work, and citizenship.*⁵⁴

Who could disagree? The problem is that almost all measurement of non-cognitive skills is done through surveys. (The other option is observations by trained personnel, which would be quite expensive.) Surveys raise several issues, one of which is known as "reference bias." Imagine a demanding school, with a lot of homework, and another school that is more laissez-faire, with little homework. If we ask students and teachers at both schools to rate kids on conscientiousness, those at the first school will likely have much higher expectations, so they will not rate themselves or their students as highly as those in the laissez-faire school would. Harvard Professor Martin West says it well: "To the extent that students attending schools with more demanding expectations for student behavior hold themselves to a higher standard when completing questionnaires, reference bias could make comparisons of their responses across schools misleading. If schools with high expectations are actually more effective in improving students' non-cognitive skills (something not yet known but often assumed), conclusions about school performance based on self-reports could even be precisely backward."

And what would happen if teachers started talking with students a great deal about things like self-control, conscientiousness, and persistence? Would some kids decide, "I'm just low on persistence, and I always will be"? Experts who promote the development of non-cognitive skills worry about that. According to Gabrieli, lowering kids' self-perception on these skills leads to lower grades, lower test scores, and worse behavior.

Finally, if a state made school scores on these skills part of the measurement of school quality, would teachers start coaching students to influence the way they answered surveys? What other perverse behavior would emerge?

We don't know the answers to these questions. But we may soon begin to find out, because six large districts in California, including San Francisco, Los Angeles, and Oakland, volunteered to measure non-cognitive skills as part of their measurement and rating system, and the Obama administration gave them waivers to do so. By late 2016, the experiment was entering its third year. These "CORE" districts use a School Quality Improvement Index that includes test scores, graduation rates, suspension and expulsion rates, chronic absenteeism, and school culture and climate surveys by students, teachers, and parents. But it also includes student surveys on four habits of success: growth mindset (the belief that one's abilities can grow with effort); self-efficacy (a belief in one's ability to meet goals); self-management (the ability to control one's emotions); and social awareness (interpersonal skills such as empathy, collaboration, and the ability to listen). The results account for eight percent of the School Quality Improvement Index, though by 2016 the districts had not attached any consequences. According to Martin West, who led research efforts examining the results, the first year of non-cognitive skills data showed the expected correlation between social-emotional skills, grade point averages, standardized test scores, absenteeism, and suspensions, suggesting that the measures were fairly accurate.⁵⁵

Clearly, we don't know enough yet to use such data as part of accountability systems that impose consequences on schools. But these

skills are critical to success, and the best way to learn about measuring something is to start doing so. We need to measure student progress on social-emotional skills, learn more about their role in future success and the relative impact of home life vs. school life, develop better ways to measure these skills, and figure out how schools can improve them.

In August 2016 the Collaborative for Academic, Social, and Emotional Learning (CASEL) announced it will help eight states "create and implement plans to encourage social-emotional learning in their schools."⁵⁶ But CASEL also warns it is too early to begin attaching consequences—or even to include the data in rating systems. For now, states should include non-cognitive skills only in their measurement systems. This information would be valuable to both schools and parents, when choosing schools. If a few states or districts want to experiment with including the data in their rating and accountability systems, as the CORE districts are, they could speed up our learning curve. Charter authorizers should also be encouraged to negotiate performance goals that include such measures with schools that are interested in being accountable for improving students' non-cognitive skills. But it is far too early to force accountability for such improvement on schools that don't want it.

AN IDEAL STATE RATING AND ACCOUNTABILITY SYSTEM

At the risk of repeating myself, state systems enforce the minimum levels of performance we will accept from our public schools. We need other means of accountability—performance agreements and charters—to create incentives for alternative schools, other schools that differ

from the norm, and schools that routinely score above the minimum.

That said, I would suggest that today's statewide rating systems—applied to all schools—should have five or six basic elements, weighted roughly as follows. (The balance between achievement and growth should depend on which method states use to measure growth; with some value-added methods, such as EVAAS, achievement and growth can be combined in one value-added score.) ESSA requires that states also use English learner progress toward proficiency, but I have not specified a recommended weight because it should vary by school. In some, with many English language learners, it would be quite important; in others, with none, it would be unimportant.

For high schools:

- Student academic achievement: 20 percent*
- Student academic growth: 25 percent
- English learners' progress toward proficiency: variable
- Student engagement: 10 percent
- Qualitative school assessments by experts: 15 percent
- Student outcomes: 25 percent

Elementary and middle schools would use only four elements:

- Student academic achievement: 20 percent*
- Student academic growth: 30 percent
- English learners' progress toward proficiency: variable
- Student engagement: 10-20 percent
- Qualitative school assessments by experts: 30 percent

Indicators for each element could include the following:

Academic achievement and growth:

- Test scores in math, ELA, writing, science, and the social sciences
- For English language learners, scores on tests designed to measure their progress in learning English
- PSAT, SAT ACT, and/or state-approved international test scores
- Industry certifications

Qualitative assessments

- Expert site visit assessments (see pp. 26-27)

Student engagement

- Parent surveys

Outcomes (for high schools)

- HS graduation rate: 4 year, 5-7 year, and with GED (perhaps 2, 2, and 1 percent for each category, respectively)
- Percent of graduates enrolling in college
- Percent of enrollees required to take remedial classes in college

** Can be eliminated using certain value-added models for measuring growth, in which case all 45 or 50 percent of the weight would go to the value-added score.*

- Percent of college enrollees persisting to second and third years
- Percent of two-year college enrollees completing a two-year degree or credential
- Percent of non-college-bound graduates employed, in training, or in the military
- Income levels for those employed full-time

Measures Recommended for State Accountability Systems, Measurement, and Research

INCLUDE IN A STATEWIDE ACCOUNTABILITY SYSTEM:

Student academic growth and achievement, measured by:

- Test scores in math, ELA, writing, science, and the social sciences
- For English language learners, scores on tests designed to measure their progress in learning English
- PSAT, SAT, and/or ACT scores and/or state approved international test scores
- Industry certifications
- Progress toward proficiency of English language learners

Qualitative assessment, measured by:

- Expert site visit assessments

Student engagement, measured by:

- Parent survey

For high schools only: Student outcomes, measured by:

- HS graduation rate: 4 year, 5-7 year, and with GED
- Quality of diploma, if states offer different diplomas
- Percent of graduates enrolling in college
- Percent of enrollees required to take remedial classes in college
- Percent of college enrollees persisting to second and third years
- Percent of two-year college enrollees completing a two-year degree or credential
- Percent of non-college bound graduates employed, in training, or in the military
- Income levels for non-college bound graduates employed full-time

REQUIRE DISTRICTS AND AUTHORIZERS TO MEASURE BUT DON'T INCLUDE IN STATEWIDE RATING OR ACCOUNTABILITY SYSTEM:

- Student attendance rates
- Rates of chronic student absenteeism
- Rates of teacher absenteeism
- Student surveys
- Parent surveys
- Student demand
- Student retention
- Teacher retention
- Safety
- Discipline rates
- Numbers of advanced courses (AP, IB, dual credit, etc.)
- Numbers of student internships
- Student-teacher ratios

FUND RESEARCH TO FIND OBJECTIVE, RELIABLE WAYS TO MEASURE:

- Qualitative assessments of student performance (performance tasks, portfolios, etc.)
- Student surveys
- Assessments of non-cognitive "character" skills

TO GRADE OR NOT TO GRADE?

Ideally, states should give various weights to each indicator and sum them to give a grade for each area, plus an overall grade. Some states use colors or phrases, such as “meets expectations.” But parents understand grades on an A-F scale better. (I would urge states to use pluses and minuses for more precision, just as schools do). If we use grades for our children, we should have the courage to use them for the adults who run our schools.

Some experts argue against a summative grade. California’s Superintendent of Public Instruction, Tom Torlakson, and the president of its State Board of Education, Michael Kirst, expressed this view in a recent letter to the U.S. Department of Education: “A summative rating, in contrast, necessarily glosses over differences in performance across indicators and inappropriately draws school leaders, stakeholders, and the public to focus on the single rating rather than a more robust reflection of performance demonstrated by the individual indicators. We reach this conclusion having over 15 years’ experience with a single rating where the public paid little attention to the individual components that comprised that single rating.”⁵⁷

Without a single, summative grade for each school, however, accountability becomes squishy: schools face much less pressure to improve. Chad Aldeman of Bellwether Partners cites several examples to make the point: “When Wales dropped its school rating system, student achievement dropped significantly, particularly for lower-performing students. Similarly, after New York City dropped its A-F rating system and stopped applying pressure on low-performing schools, achievement in F-rated schools immediately fell.”⁵⁸ And to say that parents don’t

have the ability to look at multiple indicators and the four or five grades that sum to the final grade sells them short. For years, we have asked them to look at five or six grades on a typical report card, not to speak of marks for conduct. Most parents can understand where a school is strong and where it is weak and make their decisions accordingly.

“Summative ratings are all around us,” Aldeman points out.

If you want to go to a movie, you might consult a site like IMDb or Rotten Tomatoes. Cars, colleges, neighborhoods, restaurants, etc. You name it, if there’s some sort of choice that people can make, there’s probably at least one, if not more than one, rating system to help them decide. Even the National Education Association, which opposes the idea of rating schools, has its own A-F grading systems for individual legislators.

... It’s not just that summative ratings exist; they’re also extremely popular. Consumer Reports is an entire magazine devoted to rating everyday household products, and it’s been around since 1936 for a reason.

*...**Summative ratings are simple and easy to understand, but they’re not one-dimensional.** All of the rating systems mentioned above have various factors that go into them (in education-speak, we might say they’re based on “multiple measures”). And, while the overall rating provides a useful method for people to make decisions, none of these systems stop at a numeric rating. They all include much more information for people who want to dig in further.*

The ESSA requires states to report performance data also for subgroups: by gender, race, ethnicity, English language learners, low-income children, and those receiving special education. As the Fordham Institute recommends, states should add high-achieving students to this list. In all cases, they should average two years of data whenever possible, to smooth out annual variation and more accurately reflect school performance.

Students who arrive at a school more than six weeks into an academic year should not be included in these measures. Schools should not be held accountable—or even measured—based on students they did not have an opportunity to educate for at least six months before a test.

States must ensure the data is audited, analyzed, and spot-checked to detect cheating. In the past, districts and schools have been caught cheating on standardized tests and manipulating attendance, graduation, and dropout rates.

Some districts have labeled their worst schools “alternative schools” to avoid negative consequences. The lesson: we should be on alert for efforts to manipulate all indicators.

PRESENTING THE DATA TO THE PUBLIC

States, districts, and authorizers should publish brief performance reports on each school, showing their scores and grades. These reports should include other information that is of value to parents and the public but is not included in the ratings, such as the number of students, the student-teacher ratio, demand for the school, the school’s mission and focus, the percentage of students receiving special education, and the like.

To put school grades in perspective, states and districts could also give schools a percentile rating based on their overall score—that is, an elementary school that outperformed 62 percent of all elementary schools would have a rating of



62. (Utah, New York City, and Philadelphia have all done this in the past.⁵⁹) States would also be wise to divide schools into groups with similar demographics, then provide bar graphs to show how all those schools compare. Bar graphs give readers visual evidence of how a school stacks up against schools with similar students. One might be rated at the 62nd percentile but be very close in performance to schools at the 90th percentile, for instance. Another might be at the 62nd percentile but be quite close to those at the 40th percentile, because schools are bunched around the middle. Bar graphs reveal this, whereas percentiles do not.

We must remember, however, that individual schools should have their own performance agreements, in addition to the state standards. These would include other goals aligned with the school's particular mission and focus. Performance reports should give equal space to performance against these goals, so parents and others can see what the school feels is most important and how well it achieves those goals. Some schools might be just average when it comes to state standards but be outstanding in their own focus areas, whether that be music, art, drama, debate, STEM, student projects, community service, languages, leadership development, character development, or real-world internships. Performance reports should reflect these realities.

On the other hand, parents have indicated their strong preference for brief reports—from two to four pages.⁶⁰ More in-depth information should be provided on websites identified on and linked to the reports.

Finally, districts and authorizers need to give families help in understanding the data. They should publish explanatory material, as the

D.C. Public Charter School Board does with its *Parent Guide to Public Charter School Performance*. They should also create or contract for information centers where parents can get help choosing a school for their children, as the Recovery School District does in New Orleans. When we buy houses, most of us use real estate brokers to help us sort through the plethora of options and make the best choice. Our decisions about our children's education deserve equal care, if not more.

Some schools might be just average when it comes to state standards but be outstanding in their own focus areas, whether that be music, art, drama, debate, STEM, student projects, community service, languages, leadership development, character development, or real-world internships. Performance reports should reflect these realities.

SPECIAL EDUCATION STUDENTS

Students with severe learning disabilities should not be included in these measurement and accountability systems, for obvious reasons. States should create separate systems for them that use different indicators and give less weight to academic achievement and growth. But most students receiving special education services should be included. Students without severe disabilities may learn differently from others, but they can still learn. Some may need

accommodations during tests, such as more time. But we don't want to exempt all students receiving special education services, because that would give schools incentives to label students as needing special education. It would also be illegal under ESSA.

IN CONCLUSION

The kind of system I have described may seem like a fantasy to those steeped in the world of NCLB, but they already exist. In Massachusetts, charter schools must meet minimal state standards, but their charters also include specific goals.⁶¹ When their charter is up for renewal, a team of experts visits for a day and a half and writes a qualitative assessment

report on the school, which the state board uses in making its decision. In Washington, D.C., the Public Charter School Board uses a performance framework much like I have advocated, but individual charters include goals specific to the schools and reviews include multiple site visits to assess the quality of schools. Denver Public Schools does much the same thing with its charters. In New Orleans, charters must meet minimum state standards, and both the RSD and OPSB do on-site reviews every year, plus a high-stakes review when they are up for renewal.

All three cities are among the fastest improving in the nation, and Massachusetts has one of our highest performing charter sectors.⁶² Are these not the kinds of results we want for all our children?

Endnotes

Endnotes: All quotations not cited in the endnotes are from interviews with the author.

1. The description of ESSA's requirements: Carmel Martin, Scott Sargrad, and Samantha Batel, *Making the Grade: A 50-State Analysis of School Accountability Systems* (Washington, D.C.: Center for American Progress, May 2016) at <https://www.americanprogress.org/issues/education/report/2016/05/19/137444/making-the-grade/>; and Heather Hough, Emily Penner, and Joe Witte, "Identity Crisis: Multiple Measures and the Identification of Schools under ESSA," Policy Memo 16-3, CORE-PACE Research Partnership, Aug. 1, 2016, at http://edpolicyinca.org/sites/default/files/PACE_PolicyMemo_1603.pdf.
2. For an analysis of this problem, see Hough et. al., *ibid.* Using school quality and student success (SQSS) measures from the six CORE districts that already have a new measurement and accountability system operating under federal waivers, in California, they concluded, p. 18: "Across each of the four indicators, we found that SQSS measures would have to account for *less than one percent* of the summative measure to not change which schools are identified for CSI. This effectively removes the non-academic measures from the accountability system, since it has been shown that measures reported without consequences will not receive the same attention as measures for which schools are held accountable (Jacob, 2005)." The paper did show how a "tiered" approach to choosing schools for comprehensive support and intervention could work, but the complexity of such an approach could be problematic.
As "for targeted structured intervention," see p. 19: "We find that this method of identifying additional schools for TSI has the potential to identify an enormous number of schools. As shown in Figure 6, using academic performance, an additional 69 percent of schools (beyond those identified for CSI) would be identified for at least one subgroup. Using graduation rate, an additional 31 percent of high schools would be identified for at least one subgroup. This effect seems to be driven mainly by students with disabilities. In 61 percent of schools, students with disabilities are performing at or below the 5-percent level for all schools on academics, and 15 percent of schools have students with disabilities graduating at the 5-percent level."
3. See *National Charter School Study 2013* (Stanford, CA: Center for Research on Education Outcomes, 2013), available at <http://credo.stanford.edu/documents/NCSS%202013%20Final%20Draft.pdf>, and *Urban Charter School Study: Report on 41 Regions, 2015* (Stanford, CA: Center for Research on Education Outcomes, 2015), at <http://urbancharters.stanford.edu/download/Urban%20Charter%20School%20Study%20Report%20on%2041%20Regions.pdf>.
4. For details, see David Osborne, "How New Orleans Made Charter Schools Work," *The Washington Monthly*, June-July-August 2015, at http://www.washingtonmonthly.com/magazine/junejulyaugust_2015/features/how_new_orleans_made_charter_s055905.php?page=al; and Douglas N. Harris, "Good News for New Orleans," *Education Next*, Fall 2015, Vol. 15 No. 4, at <http://educationnext.org/good-news-new-orleans-evidencereform-student-achievement/>.
5. For details, see David Osborne, *A Tale of Two Systems: Education Reform in Washington D.C.* (Washington, D.C.: Progressive Policy Institute, 2015), at <http://www.progressivepolicy.org/slider/tale-of-two-systems-education-reform-in-washington-d-c/>.
6. For details, see David Osborne, *A 21st Century School System in the Mile-High City* (Washington, D.C.: Progressive Policy Institute, 2016), at <http://www.progressivepolicy.org/issues/education/21st-century-school-system-mile-high-city/>.
7. See *Urban Charter School Study: Report on 41 Regions, 2015*, CREDO.; and "Results Deck" powerpoint presentation from the Mayor's Office of Education Innovation, 2015.
8. *Urban Charter School Study: Report on 41 Regions*, CREDO, 2015.
9. For the weakness of turnaround strategies, see Thomas Dee, "School Turnarounds: Evidence from the 2009 Stimulus," NBER Working Paper No. 17990, National Bureau of Economic Research, April 2012; David A. Stuit, *Are Bad Schools Immortal?* (Washington D.C.: Basis Policy Research and the Thomas B. Fordham Institute, Dec. 2010); and Andy Smarick, "The Turnaround Fallacy," *Education Next*, Winter 2010, Vol. 10, No. 1. For the superiority of charters in states and cities where failing charters are closed or replaced, such as Massachusetts, D.C., New Orleans, and New York, see the 2013 and 2015 CREDO studies cited above.

10. Martin, Sargrad, and Batel, *Making the Grade*, pp. 15-16 and 24-26: "Together, academic achievement and student growth make up a combined average of 91 percent of elementary and middle school ratings—with a minimum of 71 percent and maximum of 100 percent—and an average of 63 percent of high school ratings—with a minimum of 40 percent and a maximum of 100 percent." The high school figures do not include scores on college readiness tests, such as the PSAT, SAT, ACT, and AP exams. These make up a significant portion of high school college and career readiness indicators, which together make up an average of 15 percent of high school accountability scores.
11. *Ibid.*, p. 9. These numbers are from their analysis of 35 states plus the District of Columbia. So the total numbers are probably higher.
12. Donald T. Campbell, "Assessing the Impact of Planned Social Change," Occasional Paper Series, #8, Dartmouth College Public Affairs Center, December 1976, at <http://citeseerx.ist.psu.edu/viewdoc/nload?doi=10.1.1.170.6988&rep=rep1&type=pdf>.
13. David L. Kirp, *Improbable Scholars* (New York: Oxford University Press, 2013), pp. 173-4.
14. See for instance the exchange between Michael Petrilli and Jay Greene in the Fordham Institute's blog, May, 2016. You can access all six articles at <https://edexcellence.net/articles/test-score-gains-predict-long-term-outcomes-so-we-shouldnt-be-too-shy-about-using-them>.
15. Anne Hyslop and Bill Tucker, *Ready by Design: A College and Career Agenda for California* (Washington, D.C.: Education Sector, 2012), at <http://educationpolicy.air.org/sites/default/files/publications/ReadybyDesign-RELEASED.pdf>.
16. Nancy Doorey and Morgan Polikoff, *Evaluating the Content and Quality of Next Generation Assessments* (Washington, D.C.: Fordham Institute, Feb. 2016), at <https://edexcellence.net/publications/evaluating-the-content-and-quality-of-next-generation-assessments>.
17. See, for instance, Mikhail Zinshteyn, "Should the U.S. Make Standardized Tests Harder?", *The Atlantic.com*, Nov. 24, 2014, at <http://www.theatlantic.com/education/archive/2014/11/should-the-us-make-standardized-tests-harder/383084/>.
18. Charles Taylor Kerchner, *Can Teachers Run Their Own Schools? Tales from the Islands of Teacher Cooperatives* (Claremont, Ca.: Claremont Graduate University, Oct. 2010), at http://charlestkerchner.com/cr/uploadImages/Teacher_run_case.pdf.
19. Quoted in Ted Kolderie, *Creating the Capacity for Change* (St. Paul, MN.: Education Week Press, 2007), p. 129.
20. *Testing Doesn't Measure Up for Americans: The 47th Annual PDK/Gallup Poll of the Public's Attitudes Toward the Public Schools, A Special Supplement to Kappan Magazine* (Bloomington, IN.: Phi Delta Kappan, Nov. 2015), at http://pdkpoll2015.pdkintl.org/wp-content/uploads/2015/10/pdkpoll47_2015.pdf.
21. Martin, Sargrad, and Batel, *Making the Grade*, pp. 13-16; and Morgan S. Polikoff, Andrew J. McEachin, Stehapani L. Wrabel, and Matthew Duque, Alyssa Schwenk, "The Waive of the Future: School Accountability in the Waiver Era," *Education Researcher*, Vol. 43, No. 1, Jan.-Feb. 2014, at <http://www.aera.net/Newsroom/Recent-AERA-Research/The-Waive-of-the-Future-School-Accountability-in-the-Waiver-Era>.
22. For more about this, see Osborne, *A 21st Century School System in the Mile-High City*, pp. 6-7.
23. William L. Sanders, "Comparisons Among Various Education Assessment Value-added Models," SAS Institute white paper presented at the Power of Two National Value-Added Conference, October 16, 2006, available at <http://www.sas.com/resources/asset/vaconferencepaper.pdf>
24. Mark Ehlert, Cory Koedel, Eric Parsons and Michael Podgursky, "Choosing the Right Growth Measure," *Education Next*, Spring 2014, Vol. 14, No. 1., at <http://educationnext.org/choosing-the-right-growth-measure/>.
25. For a description of those alternatives, see Lyria Boast and Tim Field, *Quality School Ratings: Trends In Evaluating School Academic Quality* (Washington, D.C.: National Alliance For Public Charter Schools, 2013), at http://www.publiccharters.org/wp-content/uploads/2014/01/Quality-Ratings-Report_20131010T114517.pdf.
26. Matthew C. Makel, Michael S. Matthews, Scott J. Peters, Karen Rambo-Hernandez, and Jonathan A. Plucker, *How Can So Many Students Be Invisible: Large Percentages of American Students Perform Above Grade Level* (Baltimore: Institute for Education Policy at Johns Hopkins University, August 16, 2016), at <http://education.jhu.edu/edpolicy/commentary/PerformAboveGradeLevel>.
27. Michael J. Petrilli, David Griffith, Audrey Kim, and Brandon Wright, *High Stakes for High Achievers: An Analysis of State Accountability Systems in the Age of ESSA* (Washington, D.C.: Fordham Institute, Aug. 31, 2016), at <https://edexcellence.net/publications/high-stakes-for-high-achievers>.
28. Sarah D. Sparks, "Should ESSA Jettison Proficiency Rates in School Accountability?" *Education Week*, July 18, 2016, at http://blogs.edweek.org/edweek/inside-school-research/2016/07/testing_researchers_want_to_je.html.
29. Petrilli et. al., *High Stakes for High Achievers*.

30. Michael Petrilli, "Two Changes to the Department of Education's ESSA Implementation Rule," *Flypaper*, Fordham Institute, August 1, 2016, at <https://edexcellence.net/articles/two-changes-to-the-department-of-educations-essa-implementation-rule>.
31. Petrilli et. al., *High Stakes for High Achievers*, p. 16.
32. Martin, Sargrad, and Batel, *Making the Grade*, p. 22.
33. Richard Whitmire. *On the Rocketship: How Top Charter Schools are Pushing the Envelope* (New York: Wiley, Kindle Edition, 2014) p. 2 "... about a third of the 2013 high school graduates who took the ACT tests, nearly all of them students planning to go to college, were not prepared for college-level writing, biology, algebra, or social studies, reported the testing company. The College Board reported even more dismal findings from the SAT: 57 percent were not prepared for college." Gene I. Maeroff, *Reforming a School System, Reviving a City: The Promise of Say Yes to Education in Syracuse* (New York: Palgrave MacMillan, 2013), p. 142. Three of four college students who take remedial courses will not graduate within 8 years, according to the Gates Foundation.
34. For information on the Texas program, see Texas Education Agency, "Questions and Answers on the Distinguished Achievement Program," at tea.texas.gov/WorkArea/DownloadAsset.aspx?id=2147487536.
35. For information on New York diplomas, see New York State Department of Education, "Diploma/Credential Requirements," at <http://www.p12.nysed.gov/specialed/diploma-credentials.html>.
36. Martin, Sargrad, and Batel, *Making the Grade*, p. 25, and Boast and Field, *Quality School Ratings: Trends In Evaluating School Academic Quality* p. 15.
37. Boast and Field, *Quality School Ratings*, p. 15.
38. *Ibid.*, p. 15, and Martin, Sargrad, and Batel, *Making the Grade*, p. 24.
39. Boast and Field, *Quality School Ratings*, p. 16.
40. For more on New York City, Vermont, and Massachusetts, see Linda Darling-Hammond, Soung Bae, Channa M. Cook- Harvey, Livia Lam, Charmaine Mercer, Anne Podolsky, and Elizabeth Leisy Stosich, *Pathways to New Accountability Through the Every Student Succeeds Act* (Palo Alto: Learning Policy Institute, 2016), pp. 31-33, at <http://learningpolicyinstitute.org/our-work/publications-resources/pathways-new-accountability-every-student-succeeds-act>.
41. Interview with Tom Pandiscio, September 2016.
42. Shaina Cavazos, "Building Better Student Assessments," *The Educated Reporter*, Education Writers Association blog, June 23, 2016, http://www.ewa.org/blog-educated-reporter/building-better-student-assessments?utm_source=salsa&utm_medium=email&utm_campaign=newsletter.
43. David T. Conley and Linda Darling-Hammond, *Creating Systems of Assessment for Deeper Learning* (Stanford, CA: Stanford Center for Opportunity Policy in Education, 2013) at https://edpolicy.stanford.edu/sites/default/files/publications/creating-systems-assessment-deeper-learning_0.pdf; and Zinshteyn, "Should the U.S. Make Standardized Tests Harder?," *The Atlantic*.
44. See the Council of Chief State School Officers' website: http://www.ccsso.org/Resources/Programs/Balanced_Systems_of_Assessment_and_Aligned_Accountability.html.
45. *Educating for the 21st Century: Data Report on the New York Performance Standards Consortium* (New York: New York Performance Standards Consortium, undated), at http://www.nyclu.org/files/releases/testing_consortium_report.pdf; and Gail Robinson, "NYC Schools That Skip Standardized Tests Have Higher Graduation Rates," *The Hechinger Report*, Oct. 30, 2015, at <http://hechingerreport.org/nyc-schools-that-skip-standardized-tests-have-higher-graduation-rates/>; and the consortium web site, <http://performanceassessment.org/consortium/>.
46. Conley and Darling-Hammond, *Creating Systems of Assessment for Deeper Learning*, pp. 10-13.
47. Personal communication from Rob Ramsdell of Tripod Education Partners, via email, July 25, 2016.
48. MET Project, *Ensuring Fair and Reliable Measures of Effective Teaching: Illuminating Findings from the MET Project's Three-Year Study* (Seattle: Bill and Melinda Gates Foundation, January 2013), pp. 11-13; and interview with Rob Ramsdell of Tripod Education Partners, July 2016.
49. Interview with Rob Ramsdell, July 2016.
50. Chris Gabrieli, Dana Ansel, and Sara Bartolino Krachman, "Ready To Be Counted: The Research Case for Education Policy Action on Non-Cognitive Skills, version 1.0," *Transforming Education*, December 2015, p. v, at http://static1.squarespace.com/static/55bb6b62e4b00dce923f1666/t/5665e1c30e4c114d99b28889/1449517507245/ReadytoBeCounted_Release.pdf.
51. *Ibid.*, p. vi.

52. *Ibid.*
53. For more on Summit Public Schools, see David Osborne, *Schools of the Future: California's Summit Public Schools* (Washington, D.C.: Progressive Policy Institute, January 2016), at <http://www.progressivepolicy.org/slider/schools-of-the-future-californias-summit-public-schools/>.
54. Ted Dintersmith and Tony Wagner, "America Desperately Needs To Redefine 'College And Career Ready'", *Market Watch*, August 8, 2016, at <http://www.marketwatch.com/story/america-desperately-needs-to-redefine-college-and-career-ready-2016-08-05>.
55. Martin West, "Should Non-Cognitive Skills Be Included In School Accountability Systems? Preliminary Evidence From California's CORE Districts," *Evidence Speaks Reports*, Economic Studies at Brookings Institution, Vol. 1, #13, March 17, 2016, at <https://www.brookings.edu/wp-content/uploads/2016/07/EvidenceSpeaksWest031716.pdf>.
56. Evie Blad, "Social-Emotional Learning: States Collaborate to Craft Standards, Policies," *Education Week Blogs*, August 1, 2016, at http://blogs.edweek.org/edweek/rulesforengagement/2016/08/social-emotional_learning_states_collaborate_to_craft_standards_policies.html?cmp=eml-eb-popweek+08052016.
57. Letter from Torlakson and Kirst to Meredith Miller, U.S. Department of Education, Aug. 1, 2016, Docket ID: ED-2016-OESE-0032, at <http://www.cde.ca.gov/re/es/regletter1.asp>.
58. Chad Aldeman, "Summative Ratings Are All Around Us. Why Are We Afraid of Them in K-12 Education?" *Education Next*, Aug. 9, 2016, at <http://educationnext.org/summative-ratings-are-all-around-us-why-are-we-afraid-of-them-in-k-12-education/>.
59. Boast and Field, "Quality School Ratings," p. 21.
60. *Reporting Results: What the Public Wants to Know* (Arlington, Va.: A-Plus Communications, 1999), p. 9. (A-Plus Communications is now KSA-Plus Communications, at <http://www.ksaplus.com/about-us.html>.)
61. See Massachusetts Office of Charter Schools and School Redesign, *Charter School Performance Criteria*, Version 3.3 (Boston, MA.: Massachusetts Department of Elementary and Secondary Education, June 2016), at <http://www.doe.mass.edu/charter/acct.html?section=criteria>.
62. Center for Research on Education Outcomes, *National Charter School Study 2013*, CREDO.



The Progressive Policy Institute is a catalyst for policy innovation and political reform based in Washington, D.C. Its mission is to create radically pragmatic ideas for moving America beyond ideological and partisan deadlock.

Founded in 1989, PPI started as the intellectual home of the New Democrats and earned a reputation as President Bill Clinton's "idea mill." Many of its mold-breaking ideas have been translated into public policy and law and have influenced international efforts to modernize progressive politics.

Today, PPI is developing fresh proposals for stimulating U.S. economic innovation and growth; equipping all Americans with the skills and assets that social mobility in the knowledge economy requires; modernizing an overly bureaucratic and centralized public sector; and, defending liberal democracy in a dangerous world.

© 2016

Progressive Policy Institute
All rights reserved.

Progressive Policy Institute
1200 New Hampshire Ave NW,
Suite 575
Washington, DC 20036

Tel 202.525.3926
Fax 202.525.3941

info@ppionline.org
progressivepolicy.org