

Identity Crisis: Multiple Measures and the Identification of Schools under ESSA

CORE-PACE RESEARCH PARTNERSHIP

August 1, 2016

Heather Hough (hjhough@stanford.edu)

Emily Penner (pennere@uci.edu)

Joe Witte (joewitte@stanford.edu)

Summary and Policy Implications

- The Every Student Succeeds Act (ESSA) makes sweeping changes to the way school performance is measured. We use the innovative measurement system developed by the CORE Districts in California to explore how schools can be identified for support and improvement using a multiple measures framework in direct response to the ESSA regulations. We find that:
- Different academic indicators measure very different aspects of school performance, illuminating different dimensions of schools' strengths and weaknesses. For this reason, a summative score fails to identify schools with acute levels of low performance on particular indicators. Given the value judgments inherent in the identification of schools for Comprehensive Support and Improvement, ESSA regulations should allow and encourage states to make full use of multiple measures to identify schools in the way they see fit, through the reporting and use of "dashboards" of indicators.
- If the inclusion of School Quality and Student Success measures is important, ESSA regulations should be revised to allow them to have weight in the identification of schools for Comprehensive Support and Improvement. As the regulations currently stand, these indicators must have a weight of less than one in order to not change the identification of schools that would otherwise be identified using academic indicators alone. We recommend that the SQSS indicators be allowed to contribute to identifying schools for support, even if the weight is small.
- The ESSA provision that schools will be identified for targeted support if they have subgroups performing at the 5-percent level for all students will result in a majority of schools being identified. With infinite resources, it would be possible, and potentially desirable, to support all such schools. In the absence of such resources, it would be useful to further specify how to prioritize which schools to support through TSI identification.

The Every Student Succeeds Act (ESSA) makes sweeping changes to the way school performance is measured and shifts many of the decisions about what to measure, how to identify schools for support, and what types of support to provide back to the states. ESSA's passage began a two-year process in which policymakers and educators at all levels will have to make decisions about how to improve policy and practice in this new era, including how to define and measure school quality and how to support schools in meeting this high standard. Federal policymakers are now engaged in the process of setting the proposed ESSA regulations,¹ which will be revised following a public comment period closing August 1, 2016. Simultaneously, states are engaging stakeholders in the long, hard, and important work of building consensus for the new systems that will be rolled out in the 2017-18 school year. In this shifting landscape, policymakers are tasked with making many decisions – both conceptual and technical, big and small – that will affect our country's children for decades to come. To support policymaking at both the state and federal level, we use the innovative measurement system developed by the CORE Districts in California to explore one important aspect of ESSA, the identification of schools for support and improvement using a multiple measures framework.²

ESSA requires a more comprehensive approach to measurement than was required under NCLB, with the intention of including more measures and moving away from adverse consequences of NCLB's measurement system, namely the narrowing of the curriculum towards tested subjects and content, strategic gaming of accountability structures, and cheating (Figlio & Getzler, 2002; Jacob & Levitt, 2003; Lauen & Gaddis, 2015; Neal & Schanzenbach, 2010). Specifically, ESSA requires states to include multiple measures of student academic achievement, including: academic performance as measured by proficiency on English Language Arts (ELA) and math tests; academic growth; graduation rate, development of English Learner (EL) proficiency; and at least one additional indicator of "School Quality or Student Success" (SQSS). The SQSS indicator can include measures of student engagement, educator engagement, student access to and completion of advanced coursework, post-secondary readiness, or school climate and safety. The regulations further require states to create a summative composite rating from a set of at least five indicators, and to use this indicator to identify schools for support and intervention. Many states do not currently have such measures ready for use, which raises questions about the properties of new measures and how multiple measures will be used in conjunction with one another in systems of accountability and support.

In this memo³ we address some of these measurement questions, using data from the CORE Districts in California. The CORE Districts are best known for the "waiver" they received from the U.S. Department of Education, which freed them from some of their federal obligations under NCLB. Under the terms of the waiver, six districts (Fresno, Long Beach, Los Angeles,

¹ <https://www.federalregister.gov/articles/2016/05/31/2016-12451/elementary-and-secondary-education-act-of-1965-as-amended-by-the-every-student-succeeds>

² This brief is specifically designed to provide empirical analysis of sections 200.18, 200.19, and 200.21 in the ESSA regulations.

³ This report has been modified from the original release with corrections in N size in Table 1, Figure 5, and Table 5.

Oakland, San Francisco, and Santa Ana⁴) developed and are currently implementing an innovative accountability system (the School Quality Improvement System), that focuses on the whole school and the whole child and emphasizes the importance of the “right drivers” for school improvement. CORE’s unique system focuses on academic outcomes alongside non-academic measures of student success, including chronic absenteeism, suspension/expulsion, students’ social-emotional skills, and school climate and culture. CORE’s systematic measurement of school and student performance on multiple dimensions is unparalleled, and has generated widespread national interest in the field of education and in the popular press (Blad, 2015; Bornstein, 2015; Zernike, 2016). There is much to be learned from the CORE Districts about how such measures can be integrated into state systems meeting ESSA’s requirements.

The guiding principle of the CORE Districts in their use of multiple measures is that data should be used as a “flashlight not a hammer;” what indicators reveal about school performance should be used to help them improve and not to scapegoat or punish. Indeed, there is growing agreement among policymakers, school and district leaders, and researchers that the most important use of school effectiveness measures should be in driving continuous improvement at both the local and state levels (Darling-Hammond & Plank, 2015). From this perspective, the inclusion of additional measures in an expanded accountability system is intended to provide a more comprehensive picture of a school’s successes and challenges that may be used for many purposes by various stakeholders (Brookhart, 2009; DePascale, 2012).

The identification of schools for support and improvement is only one way in which these measures should be used over time, but it remains very important, and no less important under ESSA than it was under NCLB. Although the sanctions and perceived punishments of NCLB have been replaced by intensive support provided locally and flexibly, selecting schools for improvement remains a high stakes decision, as substantial resources will be allocated to identified schools.⁵ To support policymaking at the state and federal levels, in this analysis we explore how academic measures can be used to identify schools for Comprehensive Support and Improvement (CSI) and how measures of “School Quality and Student Success” can be included in the measurement system. We also briefly illustrate the impact of regulations on the identification of schools for Targeted Support and Improvement (TSI). Similar to prior research in accountability and performance management (Chester, 2005; Jacobs & Goddard, 2007; Linn, Baker, & Betebenner, 2002; McEachin & Polikoff, 2012), our analysis illustrates the empirical effect of various policy approaches, which we hope will support thoughtful policymaking at both the state and federal levels as regulations and state-wide implementation plans are developed.

⁴ The additional CORE Districts are Sacramento City and Garden Grove Unified School Districts.

⁵ ESSA stipulates that a district that receives funds for school improvement should receive a minimum of \$500,000 for each comprehensive support school it serves and \$50,000 for each targeted support school it serves, unless the state determines that a smaller amount is sufficient.

About the CORE Districts and the School Quality Improvement System⁶

The CORE Districts together represent over one million students, nearly 20 percent of the students served in California. The CORE waiver districts have a combined 923 Title I schools, which is more than 26 of the 50 states.⁷ In this way, even though the CORE Districts' data represents a consortium of only 6 school districts, the results from this analysis are likely generalizable to the types of schools that the ESSA regulations are intended to support, both in California and in states nationwide. The CORE Districts first reported on new school performance measures in the School Quality Improvement Index (SQII) in 2014-15.⁸ In this analysis, we use a subset of the items in CORE's SQII to explore tradeoffs in some of the provisions laid out in ESSA and the accompanying regulations.

CORE Measures that Meet ESSA Guidelines for Academic Measurement

- **Academic performance** is measured as the percentage of students testing proficient for ELA and mathematics, based on Smarter Balanced Assessment Consortium (SBAC) test scores. While CORE reports the results of the two tests separately, we averaged the percent proficient in ELA and math to form a composite.
- **Academic growth** is measured as the extent to which students in a given school have improved their performance on ELA and math tests from one year to the next relative to demographically similar students who started the school year with similar prior achievement.⁹ The result is a growth percentile (rank from 0-100) comparing schools' contribution to student growth on ELA and math test scores. While CORE reports growth percentiles on the two tests separately, we averaged the ELA and math growth percentiles into a composite measure.
- **Graduation** is reported as the percentage of students who graduate in a 4-year cohort compared with the number of students enrolled in the school (accounting for students who transfer into and out of the school).
- **EL proficiency** is represented as the percentage of students who are reclassified from English language learner status to "fluent English proficient" of the number of all the English learners who are reclassified at a school site in the current year plus all those English learners who, after five years, were not reclassified at that school.¹⁰

⁶ For further detail on the CORE Districts and the School Quality Improvement System, see <http://coredistricts.org/>.

⁷ Analysis based on data from https://nces.ed.gov/pubs2015/2015151/tables/table_03.asp

⁸ The CORE Districts first implemented the SQII in 2014-15, but not all aspects were reported in that first year. Academic growth and student reports of social-emotional skills and school culture/climate were not reported based on 2014-15 data, but they were measured and are included in the analyses presented in this paper.

⁹ The growth indicator is currently under development; the version included in these analyses is preliminary and is only available for elementary and middle schools at this time.

¹⁰ CORE's measure of EL proficiency is slightly different than what is specified in ESSA. Rather than using only test score results to determine progress on English proficiency, the CORE Districts chose to report reclassification rates, which are a combination of language proficiency scores and academic performance (Carranza, 2015).

CORE Measures that Meet ESSA Guidelines for “School Quality and Student Success”

ESSA requires that any SQSS indicator should allow for meaningful differentiation in school performance and that each indicator should be valid, reliable, comparable, and statewide (although indicator(s) may vary by grade span). For illustrative purposes, we have chosen a subset of CORE’s indicators that meet these criteria.¹¹ States could choose any of these individual indicators, a combination of them, or one or more alternate SQSS indicators not collected by CORE.

- **Chronic absence** is measured as the percent of students who have an attendance rate at or below 90 percent within a given school year. Chronic absence can severely interfere with academic achievement, while increasing attendance rates among at-risk youth can decrease achievement gaps (Chang & Romero, 2008; Ginsburg, Jordan, & Chang, 2014).
- **Suspension/expulsion** is measured as the percent of students who are suspended and/or expelled at least once in a given school year. Minority students are disproportionately suspended and expelled, resulting in decreased academic achievement (Losen, 2011; Morgan, Salomon, Plotkin, & Cohen, 2014).
- **Social-emotional skills** are measured by students’ self-report surveys in grades 5-12 that measure growth mindset, self-efficacy, self-management, and social awareness.¹² Student responses on these surveys are translated into the percent of positive responses in each school; for example, a school with a score of “80” would indicate that 80 percent of the survey questions were answered positively by students. While the CORE Districts report each of these constructs separately, for the purposes of this analysis we report an aggregated version, averaging the percent of positive responses from each construct to create a composite rating of social-emotional skills for each school. Recent analysis by West and colleagues (West, 2016; West, Scherer, & Dow, 2016) indicates that student self-reports of key social-emotional skills demonstrate acceptable levels of internal reliability and are strongly correlated with available indicators of academic performance and student behavior.¹³
- **School culture/climate** ratings are produced from surveys of students¹⁴ (grades 5-12) that include questions about the climate of support for academic learning, knowledge and perceived fairness of discipline rules and norms, school safety, and sense of belonging and school connectedness. Student responses on these surveys are translated into the percent of positive responses in each school, similar to the indicator of social-emotional skills. Over 85

¹¹ In addition to the measures detailed below, the CORE Districts also measure high school readiness in middle school.

¹² In 2014-15, all of the CORE districts surveyed grades 5-12, and some schools/districts surveyed grades 3-4 additionally. For consistency across schools, we restricted the responses to grades 5-12.

¹³ For further detail on the survey items measuring social-emotional learning, see <http://www.transformingeducation.org/measuringmesh/>.

¹⁴ CORE’s culture/climate indicator also includes survey responses from teachers and parents, which are not included in this analysis.

percent of CORE's items are from the California Healthy Kids Survey or the California School Climate Survey, both of which have been used extensively across California.¹⁵

In this brief we use CORE's indicators as implemented. We do not address the relative merits of any of the specific indicators, nor do we explore the tradeoffs within the CORE Districts' SQII. However, we acknowledge the importance of such investigations. A large body of research has shown that the statistical properties of specific indicators is important in measuring progress and comparing schools (Kane & Staiger, 2001, 2002; Koretz, 2008), and current research on specific measures should be brought to bear as policymakers are choosing how to include particular indicators in new measurement systems. For example, Polikoff (2016) has raised concerns about the use of proficiency thresholds as part of the academic performance indicator, arguing for the use of Average Scale Scores instead. Hough and Witte (2016) have examined the feasibility of using chronic absenteeism in a multi-metric accountability system, suggesting that is a viable indicator of school quality. In addition, there is an active research debate about whether to include measures of students' social-emotional skills and school culture/climate in a school accountability system and how best to do so (Duckworth, Quinn, & Tsukayama, 2012; Duckworth & Yeager, 2015; West, 2016). These discussions and critiques of individual indicators also merit careful consideration before the ESSA regulations are formalized.

The vast majority of schools in the CORE Districts have data for every indicator, with some important exceptions. (See

Table 1). First, a school does not report an EL proficiency score if it has fewer than 20 English learners, which is why only 80 percent of elementary schools report this indicator. Second, many elementary schools do not have an indicator of student social-emotional skills and schools' culture/climate, because only one grade was reported in elementary schools (5th grade), and the indicator was not included in analysis if a school had fewer than 20 responses in total. In this analysis, we retained only Title I schools with traditional grade spans to allow for straightforward comparisons, a total of 878 schools.

¹⁵ For further detail on reliability and validity of the California Healthy Kids Survey or the California School Climate Survey, visit <http://cscs.wested.org/> and <http://chks.wested.org/>, respectively.

Table 1. Percentage of schools with data for each indicator

	Elementary	Middle	High
Academic performance	100%	100%	99%
Academic growth	99%	100%	--
Graduation	--	--	98%
EL proficiency	80%	98%	93%
Chronic absence	100%	100%	99%
Suspension/expulsion	100%	100%	100%
Social-emotional skills	79%	100%	99%
School culture/climate	93%	100%	99%
Number of schools	617	136	125

Understanding the Relationship between CORE’s Academic Measures

ESSA and the proposed regulations require that states include multiple measures of academic success in their measurement system. (These measures are summarized in Table 2). The regulations specify that these academic measures must be aggregated into a single summative number to be used in identifying schools for support and improvement. The relative weighting of these measures to form the summative rating is left to the states; however, the four academic components must be afforded “substantial” weight individually and “much greater” weight than the fifth school quality component. This single summative measure, which must have at least three distinct rating categories, is to be used to identify schools for support and improvement. The nature of the support provided is largely left up to the states, but the ESSA regulations specify that schools should be identified for CSI at least once every three years if they are in the bottom 5-percent of Title 1 schools within the state based on the summative rating, within school level, or if they have a 4-year adjusted cohort graduation rate below 67 percent (high schools only). ESSA specifies that the weighting of the measure of SQSS should be such that good performance on this measure cannot remove schools from CSI identification if they would have been included based on their performance on the four academic indicators. The academic measures are thus singularly important in identifying schools for improvement, and we therefore focus initially on them, deferring discussion of the SQSS indicator for now.

Table 2. Academic measures required by ESSA, by school level

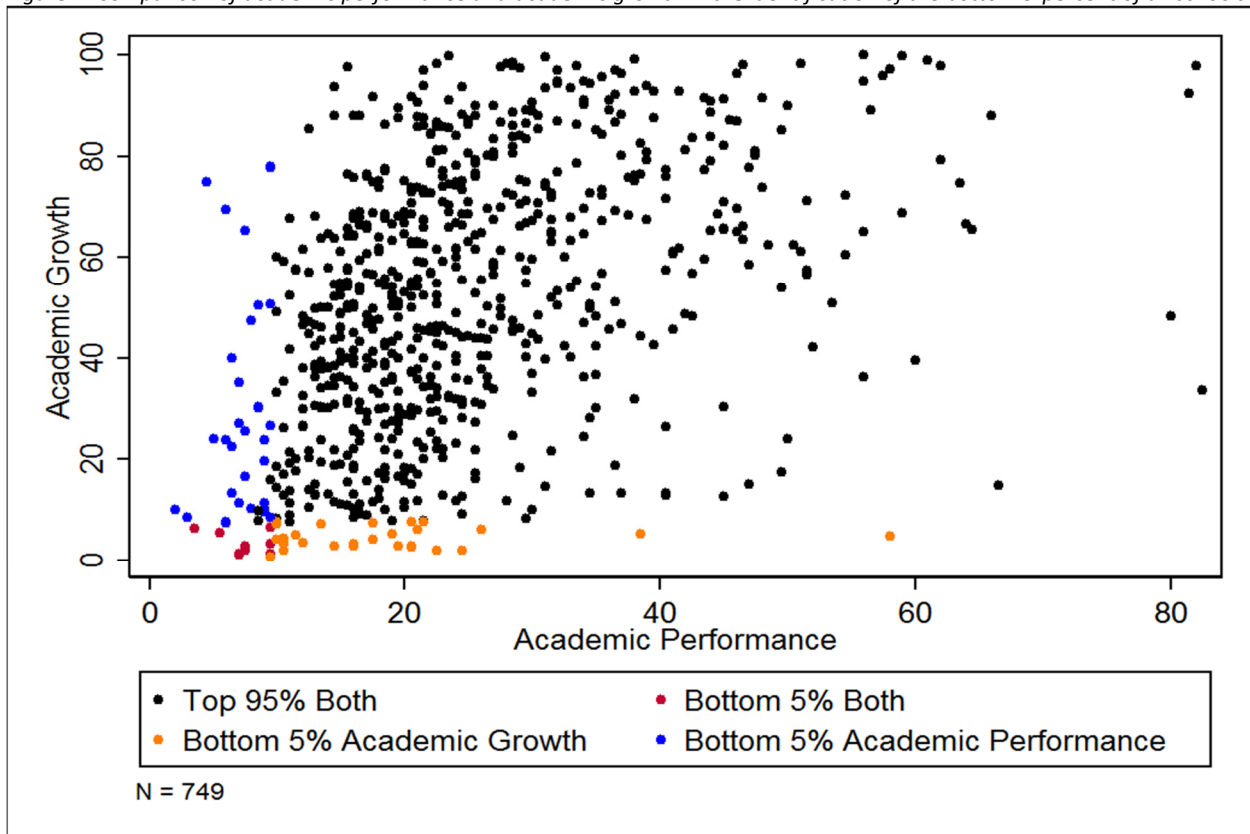
Elementary/Middle school	High school
Academic performance	Academic performance
Academic growth	Graduation
EL proficiency	EL proficiency

Relationships Among the Four Academic Indicators

In order to consider how the academic measures can best be used to identify schools for CSI we first need to understand how they are related to one another. We begin by looking at the

extent to which the different measures would “similarly identify” the same set of schools among the bottom 5-percent. As an example, Figure 1 below shows the comparison of academic performance to academic growth, for elementary and middle schools only. The red dots represent schools that would be identified in the bottom 5-percent of all schools with both measures, the blue dots represent schools that would only be identified using academic performance, and the yellow dots represent schools that would only be identified in the bottom 5-percent of all schools using academic growth. When comparing academic performance and academic growth, 70 schools are identified as being in the bottom 5-percent of all schools by either measure, but only nine schools (13 percent) are identified among the bottom 5-percent by both measures. (If the measures were identical, 100 percent of schools would be identified by both measures.)

Figure 1. Comparison of academic performance and academic growth in the identification of the bottom 5-percent of all schools



Summarizing the relationships across all of the indicators, Table 3 below shows the percentage of schools that are similarly identified in pairwise comparisons of the four academic indicators. When comparing EL proficiency and academic performance, for example, 82 schools are identified as being in the bottom 5-percent of all schools by either EL proficiency or academic performance, but only 4 percent of those 82 schools that are similarly identified by both measures. Similarly, when comparing graduation rates to academic performance, only 14

percent of schools are identified as in the bottom 5-percent of all schools under both measures (2 of 14 schools).

Table 3 . Percentage of schools similarly identified in pairwise comparisons

		Academic performance	Academic growth	Graduation	EL proficiency
Academic performance	%	100%			
	N	877			
Academic growth	%	13%	100%		
	N	70	749		
Graduation	%	14%	--	100%	
	N	14	--	123	
EL proficiency	%	4%	5%	7%	100%
	N	82	64	15	743

These analyses show that the different academic measures would identify dramatically different schools if used independently; the four academic indicators appear to be measuring very different aspects of school performance, illuminating different dimensions of schools’ strengths and weaknesses. This point is reinforced by the fact that, across measures, less than 1 percent of schools are in the bottom 5-percent of all schools on all of the measures for which they have data. Given how differently these measures identify schools, how then should they be used together to identify schools for Comprehensive Support and Improvement? We turn to this question next.

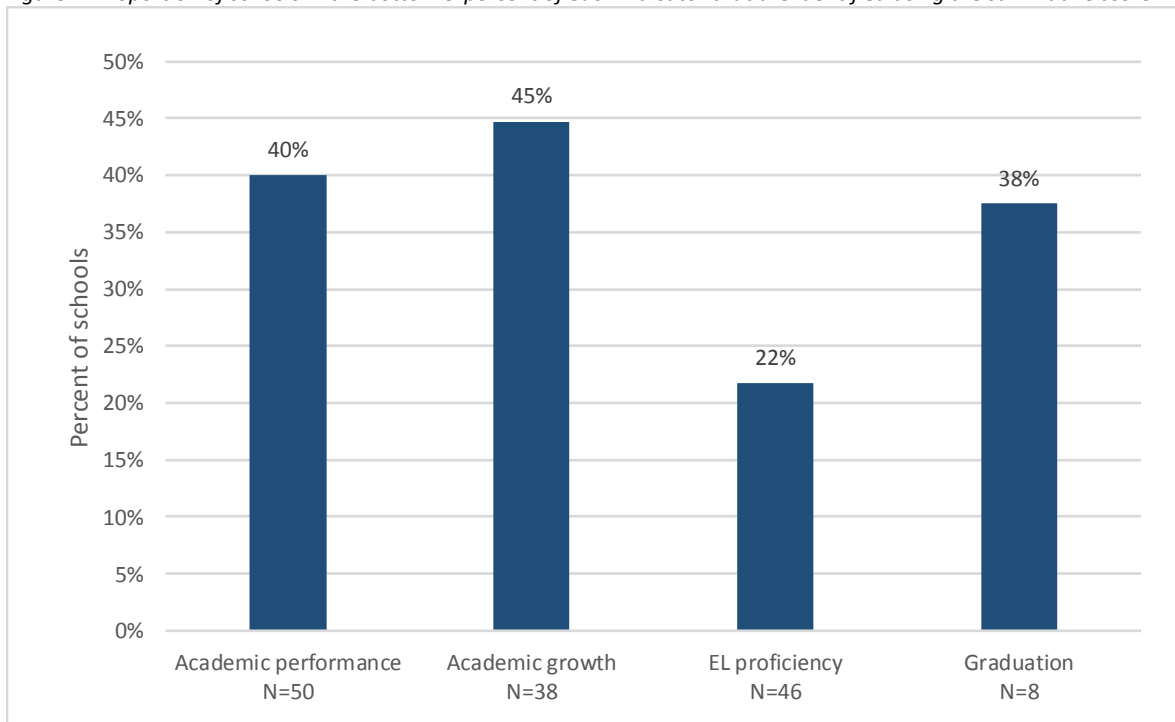
Identification of Schools for Comprehensive Support and Improvement

States are being asked to include multiple measures of school performance in their accountability systems because policymakers, and the public they serve, now recognize that schools should be held accountable for more than just increased test scores. Holding schools accountable for their performance on multiple measures is naturally far more complex than tracking a single measure. Multiple measures offer multiple ways for states to identify schools for improvement, and there are tradeoffs between various approaches. Here we explore the tradeoffs between several possible decision-rules for CSI identification, including a) identifying schools using an equally-weighted, continuous summative score; b) adjusting for low performance in the use of indicators; or c) using the full set of indicators to create decision rules about school identification using a tiered approach. Our analysis does not provide an exhaustive evaluation of each of the possible permutations that such a system could take, but instead assesses several of the important alternatives that states might consider in devising their systems for identifying schools in need of CSI.

Equally weighted, continuous summative score. To test the way a summative score would identify schools, we standardized and averaged the four academic indicators. We then identified schools for CSI when this continuous summative score placed them in the bottom 5-percent of all schools by school level. (Note that due to clustering in the variables, this method actually identifies 5.1 percent of schools.) We use equal weighting for illustrative purposes only; the ESSA regulations do not require equal weighting, but rather “substantial weight” accorded to each measure.

As shown above, these four academic measures are not highly correlated and appear to measure performance along different dimensions. As a result, the summative measure, as an average, fails to identify many schools for improvement that have very low performance on individual indicators. Figure 2 shows that many schools that are low on specific indicators are not identified for CSI using the summative score. Of the schools in the bottom 5-percent of all schools on the measure of academic performance, only 40 percent are identified for CSI using the summative measure. Similarly, only 45 percent of the schools in the bottom 5-percent of all schools on academic growth are identified by the summative measure. The corresponding numbers are 22 percent for EL proficiency and 38 percent for graduation. By aggregating across measures that represent very different dimensions of performance, the summative score may not identify schools that are low on one measure if they are even average on another. Indeed, of the 14 percent of schools that are in the bottom 5-percent on any single indicator, 71 percent are *not* identified as being in the bottom 5-percent of all schools on the summative index.

Figure 2. Proportion of schools in the bottom 5-percent of each indicator that are identified using the summative score



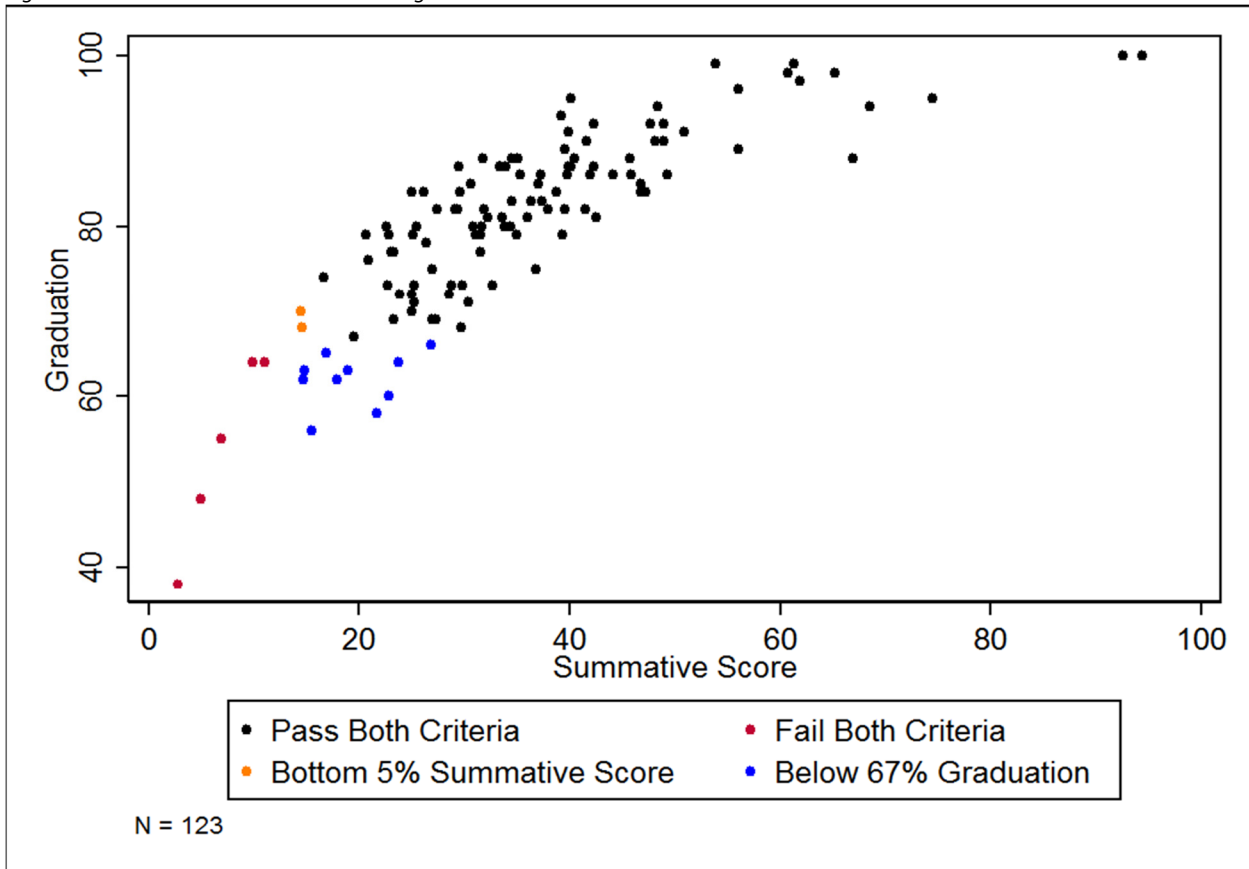
We find that this equally-weighted summative score works best to identify schools that are low on all of the indicators. For example, less than 1 percent of schools are in the bottom 5-percent of all schools on every indicator, and all of these schools are identified using the summative measure. However, the summative score also identifies a large proportion of schools that have only *relatively* low levels of performance. For example, 5 percent of schools are in the bottom 20-percent of all schools on every indicator, and 76 percent of these schools are identified for CSI using the summative measure.

The summative score appears to do well identifying schools that are relatively low on all measures but not well with those schools that are very low on one measure and average or below average on others. The judgement that states will have to make is on which of these problems to focus their limited resources. This is akin to the dilemma facing a school counselor, who can only provide intensive support to a limited number of students. Should she target her support to a student with all Ds, or to a student with mostly Cs and one F? The decision reflects a value judgement, and may depend on other characteristics of the student.

Correcting for unacceptable levels of performance. One of the downsides of using a summative score to identify schools for CSI is that it excludes many very low-performing schools from identification. To adjust, states could introduce minimum thresholds into their measurement system that adjust for unacceptable levels of performance on particular indicators. For example, states could define a level of performance that is deemed unacceptable and under which schools are identified for intervention, similar to what is already required by ESSA for graduation. ESSA regulations specify that graduation rates should be set at 67 percent and that schools should be identified for CSI even if their summative index score does not include them naturally. In the CORE Districts, 15 schools (12 percent) would be identified for CSI because they have graduation rates below 67 percent, and these schools are different than the ones that would have been identified by the summative measure. Figure 3 below compares the identification of the bottom 5-percent of all schools using the summative score versus schools with graduation rates less than 67 percent. The red dots represent schools that would be identified in the bottom 5-percent of all schools with both measures, the yellow dots represent schools that would only be identified using the summative score, and the blue dots represent schools that would only be identified because of the graduation rate threshold. Adding the graduation rate threshold to the schools identified by the summative score brings in an additional 10 schools, which increases the number of high schools identified for CSI from 5 percent (5.6 percent because of clustering) to 14 percent.¹⁶

¹⁶ It is important to note that ESSA requires states to include alternative schools in the regular accountability system. Under the CORE Districts' waiver, districts were allowed to exclude select alternative schools such as credit recovery programs and schools serving students with significant special needs. The inclusion of these schools will likely change dramatically which schools are identified in the bottom 5-percent of all schools.

Figure 3. Summative academic score with graduation threshold added



Along similar lines, states could define levels of “unacceptable performance” on each academic indicator and identify the schools that fall below this level, rather than just the 5-percent of all schools that would be identified by a summative measure. The benefit of this approach is that it enables a clear, criterion referenced definition of low-performance. The downside, of course, is that many more than 5 percent of schools could be identified, which would have strong implications for resource allocation.

Similarly, states could use a school’s relative low performance on *any* indicator to identify schools rather than to use the summative index. For example, to identify 5-percent of schools we could select schools that are in the bottom 2-percent on *any* indicator for elementary schools, and in the bottom 1-percent on any indicator in middle and high schools. If we believe that the measures of academic performance used in these accountability systems are valid and reliable and represent the highest priorities for our schools, it is difficult to argue that a school in the bottom 1 percent of all schools in the state on any indicator should *not* receive additional support for improvement. Another approach that could be employed in the context of the summative score is that states could overweight the lowest scores on specific measures in

calculating the summative score, thus ensuring that schools that perform very badly on these measures have a better chance of being identified than with a simple equal weighting of measures.

The above examples illustrate the centrality of weighting decisions in identifying the “lowest performing” schools. Very specific, technical decisions about the identification scheme will have large impacts on which schools are identified for CSI and on whether low-performing schools are excluded from CSI support. Despite their apparent precisions these decisions are fundamentally arbitrary; why should a school in the bottom 5 percent be identified but not one in the bottom 5.1 percent? Regardless of the method adopted, states should undertake the kind of empirical tests that we have presented here to better understand the unintended consequences of specific policy choices.

Using a tiered approach. To this point we have assumed the use of a summative score in the identification of schools for CSI. While a summative score of some sort is required in the proposed regulations, there is substantial disagreement about whether states should be required to create a summative score at all and what form it should take.¹⁷ Proponents of a summative score argue that it makes the identification of schools transparent and straightforward, and that flexibility in weighting can ensure that states’ systems reflect what is most valued locally. Opponents of a single score argue that reducing school performance to a single rating undermines the intent of a multiple measure system, which is designed to provide a richer view of the successes and challenges in specific schools. A single rating reduces this richness to one dimension, and is likely to focus attention on how to improve the summative score – which identifies a school as “good” or “bad” – rather than on the various dimensions that make up that score.

Instead of using a summative score, states could identify schools for CSI using a method that leverages the full information of each indicator. Using a “dashboard” of measures, states could base identification on a series of decisions about school performance on particular indicators. This process could start with two academic indicators of highest importance within the state. By way of illustration, we use academic performance and academic growth as the two first-tier indicators, since (for elementary and middle schools) these are the two academic indicators that represent the performance of all students. To determine how schools are performing on these two dimensions we use CORE’s rating system, which places schools in one of 10 levels for each indicator. These levels are set in the first year of measurement and maintained over time, rather than being recalculated every year, which gives schools the opportunity to move out of the lowest categories from year-to-year. In contrast, assigning 10 percent of schools to each performance level each year makes it difficult if not impossible for schools at the bottom of the performance distribution to show improvement or escape

¹⁷ http://blogs.edweek.org/edweek/campaign-k-12/2016/06/senate_essa_oversight_hearing_school_ratings_timeline.html

identification.¹⁸ We then collapsed the levels into four categories of performance: acute (level 1), below average (levels 2-3), average (levels 4-7), and above average (levels 8-10).

Table 4 below shows the comparison between academic performance and academic growth. In the CORE Districts, 2 percent of the schools are in the lowest performance category (red) on both academic performance and growth; these schools are the most likely candidates for comprehensive support, as their students are both struggling academically and not making progress. The orange shaded cells identify schools that have acutely-low performance on *either* academic performance or growth, or that are below average on both. These are schools that might be good candidates for CSI, but more information is needed about their performance; this is where additional measures, including SQSS indicators, can provide useful information. The yellow shaded cells indicate schools that have average or below average performance but are making reasonable growth; these schools should have clear and measureable improvement goals, but may not need comprehensive support as specified under ESSA. Finally, the green shaded cells are schools that are doing well on both their performance and growth measures, and thus do not need substantial support to improve.¹⁹

Table 4. Percentage of elementary and middle schools across academic performance and growth categories

		Academic growth			
		Acute	Below average	Average	Above average
Academic performance	Acute	2%	4%	3%	0%
	Below average	1%	5%	10%	1%
	Average	2%	8%	24%	14%
	Above average	0%	2%	10%	14%

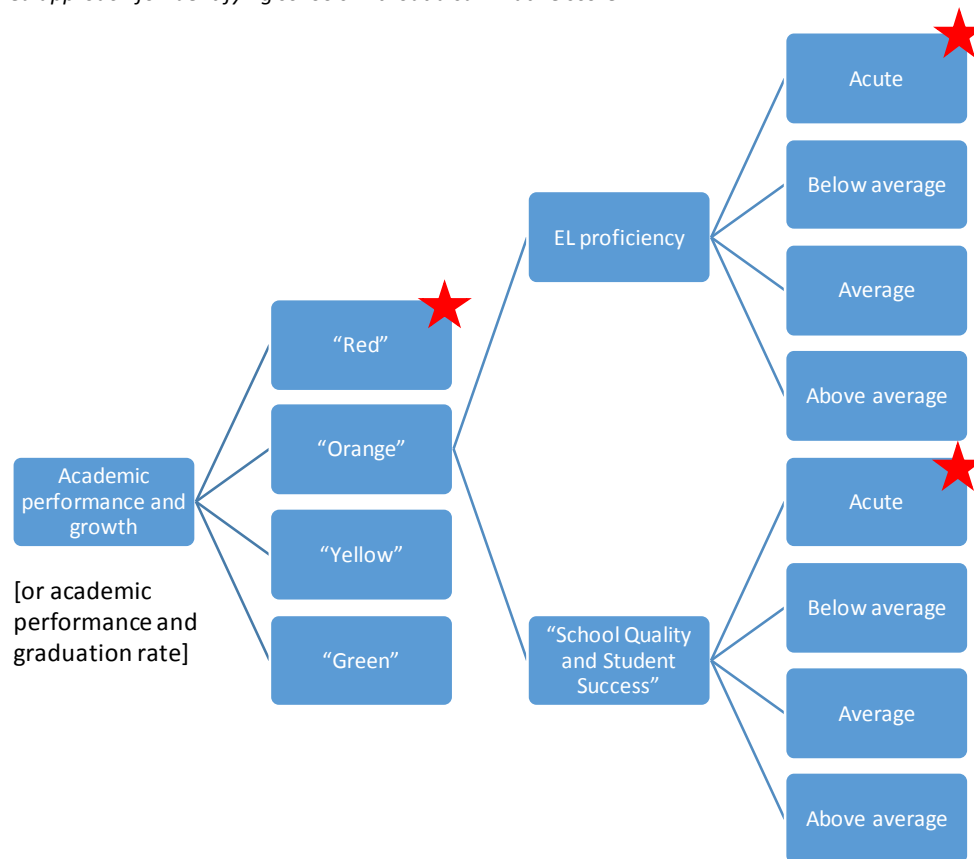
N=749

This kind of information could provide the foundation for an identification system that brings in information from the various academic and non-academic measures. An example of this approach is shown in Figure 4. The 2 percent of schools in the “red” cell would be automatically included in CSI, and the 83 percent of schools in the “yellow” and “green” categories would be excluded from identification. For the 15 percent of schools in the “orange” cells, however, the schools’ performance on additional measures can determine whether they are identified for CSI or not. For example, if “acute” performance on EL proficiency qualified a school for CSI, then an additional 2.4 percent of schools would be included in CSI. SQSS measures could be utilized here as well, as we discuss in detail below.

¹⁸ For most indicators, the levels were set based on 2013-14 data and applied to 2014-15 data; for this reason, not all indicators have 10 percent of schools in each level. Academic growth, social-emotional skills, and school culture-climate levels were set based on 2014-15 data, as that was the first year they were measured.

¹⁹ Indeed, many states have utilized similar contingency tables in their NCLB waivers, and the state of California is using a similar approach in the development of the Local Control Accountability Plans.

Figure 4. A tiered approach for identifying schools without a summative score



The framework laid out here is just an illustration. A method like this would enable states to decide what combination of performance indicators should identify a school for comprehensive support under ESSA versus which schools would likely fare well and have a good chance of meeting their improvement goals without CSI identification.

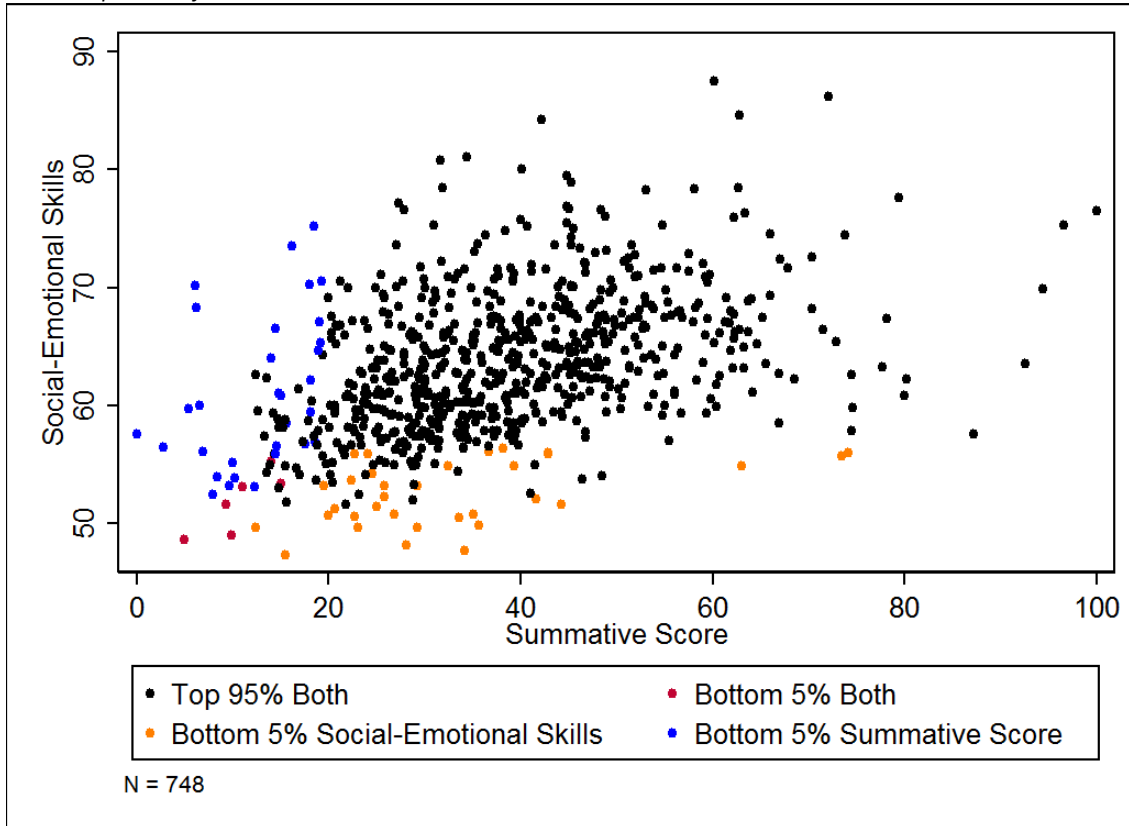
The Inclusion of Measures of “School Quality and Student Success”

In addition to the academic indicators, ESSA specifies that states must include at least one indicator of “School Quality or Student Success” that measures such factors as student access to and completion of advanced coursework, postsecondary readiness, school climate and safety, student engagement, educator engagement, or any other measure the state chooses. The ESSA regulations specify that the four academic measures should have “much greater” weight in the aggregate measure than the fifth school quality component. Furthermore, the weighting of the measure of “School Quality or Student Success” should be such that it cannot remove schools from CSI identification that would have been included when only the four academic indicators are considered.

The CORE Districts have already collected and integrated a number of these SQSS indicators, including chronic absenteeism, suspension/expulsion, social-emotional skills, and school culture/climate. States could choose any of these individual indicators, a combination of them, or one or more alternate SQSS indicators not collected by CORE. We next illustrate how these measures are related to one another and to the academic measures, and explore how they could be used in the context of the larger measurement system to identify schools for CSI.

ESSA regulations specify that the academic measures must have “substantial weight” compared to the non-academic measures, but to explore these relationships we compare each SQSS measure to a summative academic measure that weights the four academic measures equally, to see how each would identify the bottom 5-percent of all schools. As an example, Figure 5 below shows the comparison of the summative academic score to social-emotional skills. The red dots represent schools that would be identified in the bottom 5-percent of all schools with both measures, the blue dots represent schools that would only be identified using the summative academic score, and the yellow dots represent schools that would only be identified in the bottom 5-percent of all schools using the indicator of social-emotional learning. When comparing the summative academic score and social-emotional skills, 70 schools are identified as being in the bottom 5-percent of all schools by either measure, but only six schools (9 percent) are similarly identified among the bottom 5-percent of all schools by both measures. (If the measures were identical, 100 percent of schools would be identified by both measures.)

Figure 5. Comparison of the summative academic score to the indicator of social-emotional learning in the identification of the bottom 5-percent of all schools



Summarizing the relationships across all of the indicators, Table 5 shows the percentage of schools that are similarly identified in pairwise comparisons between the four SQSS indicators and the summative academic score. As with the four academic measures, schools are not often low on multiple measures simultaneously, and there is a wide range in how the measures similarly identify schools in the bottom 5-percent compared to one another and compared to the summative academic score. When comparing suspension/expulsion and chronic absence, for example, 90 schools are identified as being in the bottom 5-percent of all schools by either measure, and 16 percent of those schools are similarly identified by both measures. Suspension/expulsion has the weakest relationship with the summative academic score, with only 6 percent of schools similarly identified in the bottom 5-percent by both measures. These data suggest that the SQSS indicators are truly measuring different dimensions of school performance for the schools in the CORE Districts, despite the fact that research shows that each of these measures is related to academic outcomes at the student level. For example, a school could have high test scores but a very unhealthy climate for students, or could have high suspension/expulsion rates even though the majority of students are performing well academically.

Table 5. Percentage of schools similarly identified in pairwise comparisons

Similarity		Chronic absence	Suspension/expulsion	Social-emotional skills	School culture/climate	Summative academic score
Chronic absence	%	100%				
	N	877				
Suspension/expulsion	%	16%	100%			
	N	90	878			
Social-emotional skills	%	10%	8%	100%		
	N	81	80	748		
School culture/climate	%	10%	2%	16%	100%	
	N	87	88	69	836	
Summative academic score	%	11%	6%	9%	12%	100%
	N	91	88	70	76	878

ESSA regulations specify that non-academic measures cannot remove a school from CSI that would otherwise have been identified using the academic measures. Given how different the non-academic measures are from the summative academic score, this means that states have one of two options. First, they can include schools that are identified under either the academic and SQSS metrics, which by definition means that more than 5 percent of schools will be identified. In the case of our example above, equally weighting suspensions and the summative academic score would identify an additional 43 schools for CSI, constituting 10 percent of all schools in the CORE Districts. However, given that additional resources may not be available for CSI if states identify a larger number than 5 percent of schools, identifying additional schools with the non-academic indicator may simply result in spreading limited resources too sparsely over a larger number of schools.

The other option is for states to weight the SQSS indicators so that they do not change the identification of schools using the academic measures. 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).

Policymakers, practitioners, and communities have expressed their desire for a measurement system that better reflects the multiple goals that we as a society want our students and schools to pursue. The CORE Districts chose to include these specific indicators because they, as leaders of their own local systems, believe that these elements reflect aspects of school and student performance that are important above and beyond test scores. If these measures have to be weighted at zero to be compliant with federal law, however, the reason for including them in state accountability systems is fatally compromised.

While the SQSS indicators are not easily included in the identification of schools for CSI using a summative score under the current regulations, they could work well in a tiered identification system as laid out in Figure 4. In that example, the 15 percent of schools that have

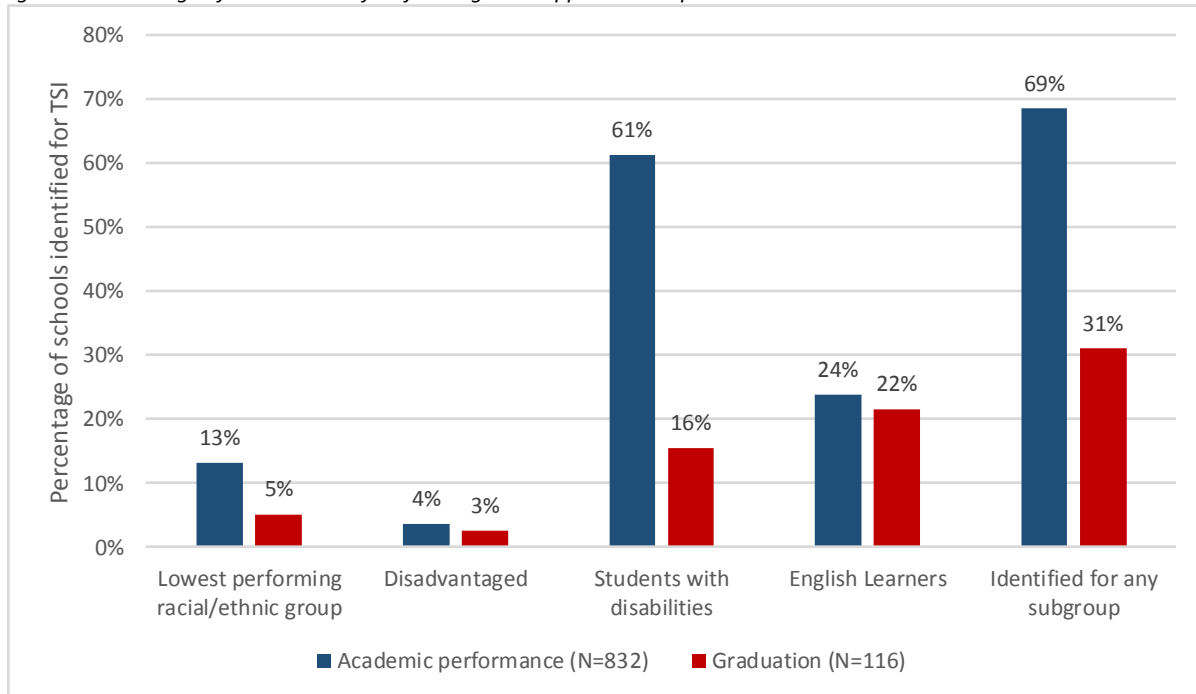
low, but not acutely low, performance enter into a second tier of review to determine whether they should be identified for CSI. In this secondary review, if one school has low scores on the non-academic measures and the other has high scores, a state could reasonably decide that the school with positive non-academic outcomes is already on the road to improvement and could focus limited resources on the school that is lower on all measures. In our CORE example, if “acute” performance on any of the SQSS indicators qualified a school for CSI in this secondary review, then an additional 5.1 percent of schools would be included (0.4 percent identified for suspension/expulsion, 3.9 percent identified for chronic absence, 1.3 percent identified for student reports of a school’s culture/climate, and none, in this example, identified for social-emotional skills). The benefit of this tiered approach is that it makes full use of all indicators by painting a comprehensive picture of school performance. Identifying schools in this way could ultimately help a state or district provide more effective supports for improvement. After all, a school identified for acutely low performance may benefit from a different set of supports than a school identified for low EL proficiency rates, or poor non-academic student outcomes.

A Note on Identifying Schools for Targeted Support and Improvement

The focus of this memo is the identification of schools for comprehensive support, but under ESSA states must also identify schools for TSI if the performance of any subgroup falls below the bottom 5-percent level on the summative rating for the “all students” group of Title 1 schools.²⁰ ESSA requires that performance be reported for each of the following subgroups: economically disadvantaged students, students from major racial/ethnic groups, children with disabilities, and English learners. To estimate the number of schools that would be identified under ESSA’s specified method for TSI, we show which subgroups would be identified using academic performance and graduation rate, as CORE’s academic growth and EL proficiency indicators are not currently broken down by subgroup. 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.

²⁰ Title 1 schools can also be identified for TSI if they have a “consistently underperforming subgroup”, one with underperformance defined by the state for two or more years. We will not explore this provision here.

Figure 6. Percentage of schools identified for Targeted Support and Improvement under individual metrics



In a world with infinite resources, supplemental funding for 5 percent of the schools for CSI and an additional 69 percent for TSI could be hugely beneficial for students. However, we do not live in a world with infinite resources, and it is hard to imagine that this many schools would actually receive targeted support. This suggests that the ESSA regulations will need to be revised to provide a different mechanism for identifying schools for TSI, perhaps using subgroup performance in the decision-tree approach, where performance of subgroups is considered in relation to other school performance indicators.

Discussion

As policymakers consider the future of accountability in the United States, one central concern is how to identify under-performing schools for additional support and improvement. ESSA attempts to strike out a new path that moves away from the achievement-only, penalty-based approach of NCLB and moves towards a multiple-measure, intensive-support approach. As with prior legislation, the details of how this new approach is designed and implemented will have a profound impact on which schools are identified and how schools demonstrate improvement. In a world of finite resources and capacity, tradeoffs must be made in order to best pursue these new objectives. The good news about this decision process is that it does not exist in a vacuum. There is much to be learned from prior accountability systems, NCLB waivers, and innovative practices being adopted across the country.

In this memo we have examined data from the CORE Districts’ multiple measure accountability system to explore some of the different methods that could be employed to identify schools for support and improvement under ESSA. Our analyses show that decisions about how to identify schools are highly sensitive to the specific definitions employed, and that these technical decisions reflect value judgements. Given the huge variation in the ways in which schools can be identified, and the costly consequences that could result if low-performing schools are not provided with adequate support, we make the following recommendations for refining the ESSA regulations when it comes to school identification:

- 1. Do not require states to report school performance using a summative rating.** Different academic indicators measure very different aspects of school performance, illuminating different dimensions of schools’ strengths and weaknesses. A single score rating, even with only three categories, reduces this richness to one dimension, and is likely to focus attention on how to improve the summative score – which identifies a school as “good” or “bad” – rather than on the various dimensions that make up that score. A central goal of ESSA is to shift control to the local level and measure school performance in a more meaningful way, specifically in a way that is meaningful to local decision-makers. A summative score might align best with the priorities of one state, but another might prefer to support schools that are particularly low on a few key indicators. States will have to decide how they want to identify schools for CSI, and in so doing, determine how their limited resources are best allocated. Given the value judgments inherent in the identification of schools for Comprehensive Support and Improvement, ESSA regulations should allow and encourage states to make full use of multiple measures to identify schools in the way they see fit, whether through a summative index, a minimum threshold of performance, a tiered approach, or another alternative.
- 2. Allow the states to include School Quality and Student Success in the identification of schools for CSI.** The “5th indicator”, which describes a school’s School Quality and Student Success, provides a promising path for helping schools to improve on dimensions beyond academic performance. However, its current role in the accountability and school support regulations minimizes its potential to support school improvement. Giving SQSS measures no meaningful role in determining which schools to target for CSI, as the current ESSA regulations do, makes it very unlikely that schools and communities will take non-academic measures seriously. While it does not need to be afforded the “substantial weight” given to the academic indicators, if the SQSS indicators are valued as an important component in understanding school performance, they indicators should contribute to identifying schools for support, perhaps as part of a tiered approach, or contributing a small weight to an overall score, even if its inclusion changes the identification of schools for CSI.
- 3. Revise the TSI identification or be prepared to provide supplemental support to a majority of schools.** ESSA’s guidelines for identifying schools for TSI are aggressively focused on illuminating the performance of subgroups and supporting schools to improve

their outcomes. We show, however, that the majority of schools could easily be identified for TSI, depending on the measure chosen, which is hardly targeted support. With infinite resources, it would be possible, and potentially desirable, to support all such schools. In the absence of such resources, however, it would be useful to further specify how to prioritize which schools to support through TSI identification. In particular, states should also be allowed to consider subgroup performance in relation to other indicators of school performance when selecting schools for TSI.

ESSA presents an opportunity for states, districts, and schools to benefit from a new orientation towards flexibility and support. With this flexibility comes great responsibility. Too many students at too many schools continue to underperform, in spite of great potential. States and districts will be given substantial authority to adopt accountability policies that suit their contexts and meet the needs of the students they serve. In keeping with this spirit of local expertise, ESSA regulations could, with minimal revisions for clarity and specificity, provide a strong mandate for identifying the right schools for support through instituting a multiple-measures system that affords careful attention to the varying needs of an increasingly diverse student population. Defining the rules for identification too tightly will needlessly constrict what states and districts can do to advance school improvement. If ESSA's goal is to promote innovation in school improvement, identifying the right schools to support is as important as offering the right kinds of support.

References

- Blad, E. (2015, December 10, 2015). CORE Districts to Share Social-Emotional Measures to Inform Accountability Plans. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/rulesforengagement/2015/12/core_districts__social-emotional_measures_to_inform_accountability_essa.html
- Bornstein, D. (2015). Teaching social skills to improve grades and lives. Retrieved from http://opinionator.blogs.nytimes.com/2015/07/24/building-social-skills-to-do-well-in-math/?_r=0
- Brookhart, S. (2009). The many meanings of “multiple measures”. *Educational Leadership*, 67(3), 6-12.
- Carranza, R. (2015) How best to measure success with English learners. *Redesigning Accountability*. Retrieved from <https://edsources.org/2015/holding-school-districts-accountable-for-success-with-english-learners/87121>.
- Chang, H. N., & Romero, M. (2008). Present, Engaged, and Accounted for: The Critical Importance of Addressing Chronic Absence in the Early Grades. Report. *National Center for Children in Poverty*.
- Chester, M. D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice*, 24(4), 40-52.
- Darling-Hammond, L., & Plank, D. N. (2015). Supporting Continuous Improvement in California's Education System. *Policy Analysis for California Education, PACE*.
- DePascale, C. A. (2012). Managing Multiple Measures. *Principal*, 91(5), 6-10.
- Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of educational psychology*, 104(2), 439.
- Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, 44(4), 237-251.
- Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: Gaming the system*. Retrieved from <http://www.nber.org/papers/w9307>
- Ginsburg, A., Jordan, P., & Chang, H. (2014). Absences add up: How school attendance influences student success. *Portland, OR: Attendance Works*.
- Hough, H. J., & Witte, J. (2016). *Using Chronic Absence in a Multi-Metric Accountability System*. Retrieved from <http://www.edpolicyinca.org/publications/using-chronic-absence-multi-metric-accountability-system>
- Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5), 761-796.
- Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating*. Retrieved from <http://www.nber.org/papers/w9413>
- Jacobs, R., & Goddard, M. (2007). How do performance indicators add up? An examination of composite indicators in public services. *Public Money and Management*, 27(2), 103-110.
- Kane, T. J., & Staiger, D. O. (2001). *Improving school accountability measures*. Retrieved from <http://www.nber.org/papers/w8156>

- Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 91-114.
- Koretz, D. M. (2008). *Measuring up*: Harvard University Press.
- Lauen, D. L., & Gaddis, S. M. (2015). Accountability pressure, academic standards, and educational triage. *Educational Evaluation and Policy Analysis*, 0162373715598577.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the no child left behind act of 2001. *Educational Researcher*, 31(6), 3-16.
- Losen, D. (2011). Discipline policies, successful schools, and racial justice.
- McEachin, A., & Polikoff, M. S. (2012). We Are the 5% Which Schools Would Be Held Accountable Under a Proposed Revision of the Elementary and Secondary Education Act? *Educational Researcher*, 41(7), 243-251.
- Morgan, E., Salomon, N., Plotkin, M., & Cohen, R. (2014). The school discipline consensus report: Strategies from the field to keep students engaged in school and out of the juvenile justice system. *The Council of State Governments Justice Center: June, 4*, 112.
- Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2), 263-283.
- West, M. R. (2016) Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE Districts. *Vol. 29. Evidence Speaks*.
- West, M. R., Scherer, E., & Dow, A. W. (2016). *Measuring social-emotional skills at scale: Evidence from California's CORE districts*. Paper presented at the Association of Education Finance and Policy, Denver, CO.
- Zernike, K. (2016, February 29, 2016). Testing for joy and grit: Schools nationwide push to measure students' emotional skills. *New York Times*. Retrieved from <http://www.nytimes.com/2016/03/01/us/testing-for-joy-and-grit-schools-nationwide-push-to-measure-students-emotional-skills.html?action=click&contentCollection=Opinion&module=RelatedCoverage®ion=EndOfArticle&pgtype=article>

We are grateful to the S.D. Bechtel, Jr. Foundation for their support for the preparation and publication of this report. The views expressed are those of the authors, and do not necessarily reflect the views of PACE, its funders, or the CORE Districts.