Multiple Measure Accountability Under ESSA: Early Findings From Three States

Kerstin Le Floch, Drew Atchison, Umut Ozek, Katie Hyland, Damon Blair, Steve Hurlburt

AIR®

Advancing Evidence.
Improving Lives.

Federal law and accountability measures

Since the mid-1990s, federal law has required states to hold schools accountable by measuring and reporting school performance, with the broad and ambitious goal of improving public education for all students. By identifying schools that are persistently underperforming; sharing performance information with administrators, educators, and parents; and providing additional resources to underperforming schools; policymakers hope to stimulate and sustain school improvement.

The Elementary and Secondary Education Act (ESEA), periodically reauthorized by Congress, establishes the basic framework through which states must test students, determine performance levels, provide information to the public, and arrange for supports to schools that do not meet the needs of all students. In the broadest terms, federally mandated school accountability is intended to promote both equity and excellence in public schools: *equity*, by identifying schools in which specific groups of students are persistently underperforming, as evidenced by substantial achievement gaps; and *excellence*, by holding schools to high standards and establishing policy levers to stimulate school improvement.

Key Findings

Schools identified for comprehensive support and improvement (CSI) are often rated on fewer indicators than are other schools.

Although accountability systems incorporate multiple measures, individual measures disproportionately drive CSI identification.

Progress or growth measures are weakly associated with student poverty levels—thus putting schools on a more even playing field.

Image by Jean-Paul Jandrain from Pixabay

To accomplish these goals, however, states need to get measurement right. That is, the set of school performance measures that states use should work together to signal areas of performance that schools should prioritize, and to identify a set of schools that do, objectively, need the most improvement. And the measures should incentivize school stakeholders to behave in ways that will lead to real improvements in student outcomes—not superficial improvements or "gaming the system" to avoid identification.

Over the successive reauthorizations of ESEA—most recently as the *Every Student Succeeds Act* of 2015, or ESSA—federal and state policymakers have struggled with the appropriate mix of measures to provide valid and useful information about schools. ESSA requires states to use a broader set of measures to define school performance compared to the prior iteration of ESEA—the *No Child Left Behind Act* of 2001, or NCLB. It charges them with identifying the lowest performing 5% of Title I schools across these measures (as well as those with graduation rates below 67%) as Comprehensive Support and Improvement (CSI) schools. States and districts must then direct intensive and tailored supports intended to drive improved instructional practices and student outcomes in CSI schools.[i] In this brief, we share findings from three states—California, Florida, and Ohio—regarding the measures they selected to include in their ESSA accountability systems and how those measures influenced which schools were identified as CSI in the 2018–19 or 2019–20 school years.

Rationale for multiple measures

Schools are complex organizations. The structures, systems, and practices that adults establish in schools shape student outcomes in a multitude of ways. Fundamentally, schools should enable students to master academic content, but student success and engagement can manifest itself in other ways: high attendance rates, strong graduation rates, healthy school climate, and shrinking gaps between the achievement levels of different student groups. Early school accountability policies were criticized for over-emphasizing reading and math outcomes, while overlooking other information that signaled that a school was exerting a positive influence in students' lives.

ESSA attempted to remedy that, mandating that state accountability systems include five indicators of school performance, which states could measure as they saw fit (see box). For example, when measuring *school quality and student success*, the three states in our study elected to use suspension rates and college/career readiness (California), a "prepared for success" measure (Ohio), and achievement in science, achievement in social studies, and a "middle school acceleration" measure (Florida). (See Appendix A for more details on each state's approach.)

The premise of ESSA's "multiple measures" approach to school accountability was that the evaluation of school performance should be more comprehensive and meaningful than under previous federal policies. In addition, these measures should focus the attention of school administrators and educators on multiple facets of school performance, signaling the importance of outcomes other than student performance in reading and mathematics, and counterbalancing pressure to narrow the curriculum. In sum, schools identified as CSI should be those that are, across the board, failing to meet students' needs. The reality of CSI identification has played out a bit differently.

ESSA requires the following indicators of school performance:

- Student achievement in reading and math
- Growth or another state-selected academic indicator for elementary and middle schools
- Graduation rate (for high schools)
- English language proficiency progress (for English learners)
- School quality and student success

Image by cherylt23 from Pixabay

CSI schools are often rated on fewer indicators than other schools

CSI schools are often evaluated on fewer indicators[ii] than are other schools, rendering school performance ratings leading to CSI designation less comprehensive than policymakers had intended. However, when a school has too few students to reliably calculate the results for a given indicator (that is, the number of students does not meet the "minimum n") states cannot include the indicator in accountability determinations. This issue is most pronounced among high schools. Among the three states in our sample, the median number of indicators on which high schools are evaluated is consistently higher among all schools compared to CSI schools: In California, the median number for all high schools is five, compared to two indicators for CSI schools. In Florida, the median number of indicators for all high schools is 10, compared to six for CSI high schools; in Ohio, the relevant numbers are five for all high schools and three for CSI high schools. Across the three states, those

schools rated on fewer indicators tended to be substantially smaller than average and were more likely to be alternative or have nontraditional grade configurations.

Schools that are evaluated based on more data points have more ways in which they can demonstrate success, and hence more ways to demonstrate that they do not deserve to be classified among the very lowest performing schools in the state. In contrast, schools with few indicators do not have as many "escape routes" from CSI status.

- **Over half of California CSI high schools were identified based on two or fewer indicators, and 43% were identified based on only one indicator.** In almost all cases where schools were identified on a single indicator, this single indicator was the suspension rate. At the elementary/middle school level, 11% of CSI schools were evaluated on one or two indicators, in contrast with 2% of all schools. At most, California evaluates schools based on six indicators, but some indicators only apply to students of certain grade levels, and schools may not have sufficient data to be rated for certain indicators. For example, California chose to apply chronic absenteeism only to elementary and middle school grades, whereas college and career readiness and graduation rates only apply to high schools. Furthermore, schools with data on fewer than 30 students for a given indicator do not receive a rating for that indicator. Indeed, two thirds of CSI high schools in California were ungraded on the ELA and math indicators because of insufficient data. But all CSI high schools in California have sufficient data to be held accountable for their suspension rate.

- **CSI schools in Florida have fewer graded indicators than what is typical of all schools**. Florida evaluates schools' performance based on as many as 12 measures; the median number of measures on which Florida schools are actually measured is eight. The difference between the number of graded indicators for CSI and other schools is evident among elementary and middle schools, but is particularly striking at the high school level. While 82% of all high schools were evaluated based on seven or more indicators, the same was true for only 39% of CSI high schools. One in 10 CSI high schools was identified based on a single measure.

- **Likewise, in Ohio, CSI elementary and high schools often had fewer graded indicators than what was typical of all schools**. This was particularly acute at the high school level: Of all high schools in Ohio, 92% were rated on five indicators, and only about 3% of high schools were held accountable for three or fewer indicators. Among CSI high schools, however, only 42% were rated on all five indicators that typically apply to high schools, and 58% were rated on three indicators or fewer. Among all elementary schools in Ohio, less than 10% were evaluated based on one or two indicators, collectively, compared with 18% of CSI elementary schools.

When schools are rated on fewer indicators, this compromises the intent of both federal and state policymakers. Accountability evaluations are less meaningful when based on few indicators; local stakeholders may discount the CSI designation if it is well known that it is driven by limited data. Moreover, if specific indicators are frequently missing, this diminishes their influence overall.

That said, states designed accountability systems to operate within the parameters of ESSA guidelines, which emphasized the inclusion of small schools and alternative schools in accountability systems. As they did so, they balanced a series of trade-offs between local priorities and federal requirements. Thus, the implementation of state systems reflected these tensions and compromises.

Exhibit 1. Percentage of All Schools and CSI Schools in California, Florida, and Ohio, by Number of Indicators
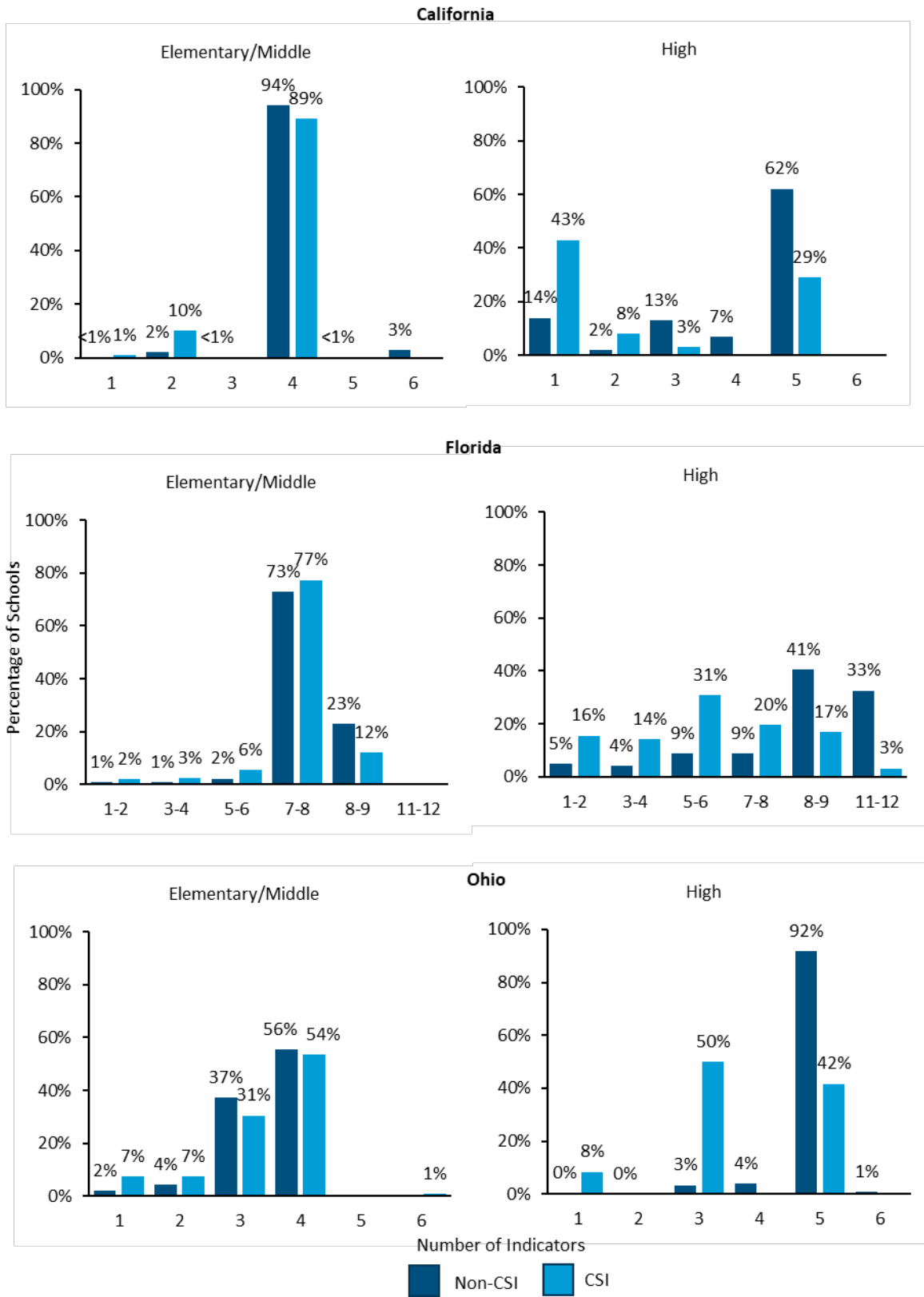
Multiple Measure Accountability Under ESSA: Early Findings From Three States

Although accountability systems incorporate multiple measures, individual measures disproportionately drive CSI identification

Within a multiple measure accountability system, each performance measure should contribute unique information and have some influence in determining which schools are identified as CSI. The weights that states attribute to various measures and indicators have some bearing on a given measure's influence on CSI ratings. For example, the weighting of achievement growth measures can vary from 15% to 40% of a school's accountability rating, depending on the state (AIR, 2017). But weights are not the only factor that determines the influence of a specific measure on a school's CSI designation: other considerations include the extent to which measures are highly correlated or provide new information, or whether a measure is frequently missing.

Importantly, some school performance measures used in state accountability systems are closely related, while others provide distinct and complementary information. For example, among elementary and middle schools in California, the math and ELA performance levels are closely correlated (0.92). Likewise, in Florida, achievement measures are highly correlated with each other—particularly math with ELA, but also science with social studies. In Ohio, the achievement indicator is closely correlated with the "gap closing" indicator—a measure of subgroup performance relative to subgroup-specific goals. This strong correlation suggests the measures are capturing very similar types of information about schools' performance.

In contrast, other measures are weakly correlated, suggesting that they are capturing different aspects of school performance, but these correlations vary across all three states. In California, the suspension rate and English learner progress (ELP) measure are the two with the weakest correlation with other measures. In Florida, the ELP measure is also weakly correlated with others, as well as other measures that are specific to certain grade levels—for example, a "middle school acceleration" measure. In Ohio, the "progress" component is the least correlated with other components across all school levels. When measures are weakly correlated, each reflects a different facet of a school's performance, meaning that each measure contributes to school accountability ratings independent of the other measures, potentially making it more influential with respect to school performance designations.

To further explore the influence of specific measures of school performance on CSI designations, we conducted a series of simulations in which we removed one of the accountability measures and then recalculated which schools were identified as CSI. The change in CSI schools (that is, the number and percentage of schools that would either be newly designated as CSI or lose their CSI status) provided information on the relative influence of each measure.

- **In California, the suspension rate measure played a larger role than other measures in determining which schools are identified as CSI.** If the suspension rate measure was dropped in California, more than one quarter of the 465 CSI schools would no longer be identified as such, and 378 schools (roughly 80% of the original CSI count) would be newly identified. Thus, the suspension rate measure introduced volatility into school accountability designations in California. Moreover, as we saw earlier, a relatively high percentage of CSI schools in California were identified only based on one indicator, which was most often suspension rate: nearly 9% of all CSI schools and 15% of CSI high schools in California were identified *only* because of the rate at which they suspend students (no other measures included).

- **In Florida, the middle school acceleration and social studies achievement indicators were very influential in shifting CSI identification at the middle school level.** Dropping these indicators from the calculation would result in 82% and 54% new CSI schools, respectively. Consequently, inclusion of these two indicators meant that a substantial proportion of middle schools avoided CSI identification. In 2018–19, only 28 middle

schools were identified as CSI in Florida, which is only 6% of all CSI schools—a proportion that would be much higher if one or both of these indicators were dropped from Florida's accountability model. However, when looking across all schools, no single indicator stands out as being more influential in determining CSI identification.

- **In Ohio, the progress and gap closing indicators had the greatest influence on the CSI designation, taking all schools into account.** At the elementary school level, the K–3 Literacy indicator was also influential: among elementary and middle schools, approximately 37% of the original count of CSI schools would become newly identified if this indicator were dropped. However, the influence of each indicator appears to be somewhat balanced in Ohio, relative to California and Florida.

Exhibit 2. Percentage of California Schools That Would Be Reclassified if Certain Indicators Were Dropped
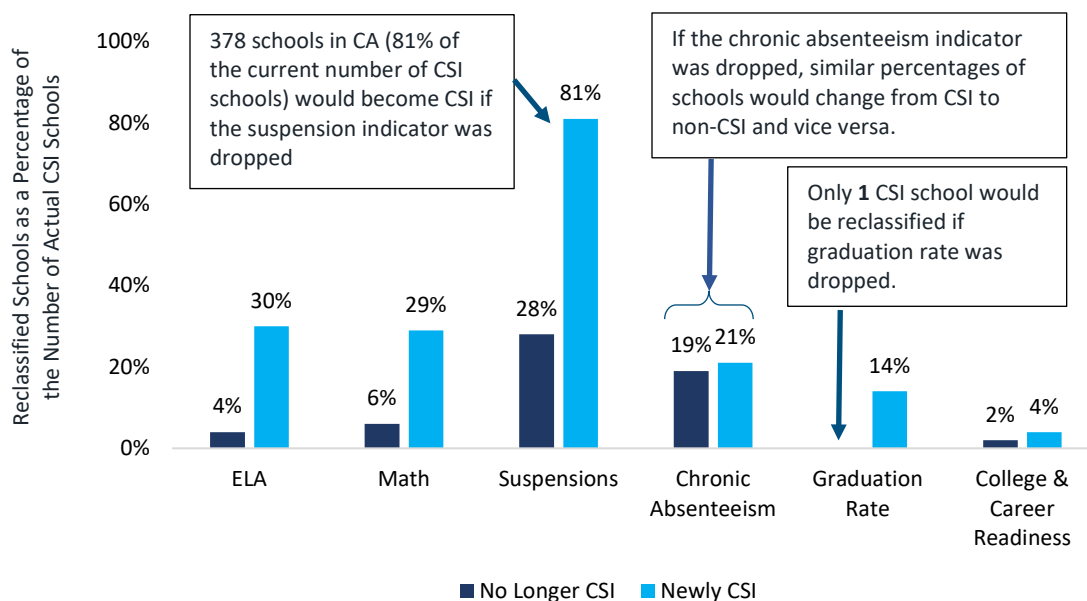
Note. Percentages are calculated using 465 CSI schools identified based on performance in 2019–20 as the denominator for all schools, 324 as the denominator for elementary and middle schools, and 106 as the denominator for high schools. The elementary/middle school and high school figures do not account for 35 CSI schools with nontraditional grade configurations.

Progress or growth measures are weakly associated with student poverty levels— thus putting schools on a more even playing field

School accountability systems are intended to measure school performance, not simply to reflect features of the student population. Measures that are strongly correlated with demographics may not reveal the true contribution of schools to student outcomes (Di Carlo, 2019) and may be seen as unfair—particularly to high-poverty schools (Wright & Petrilli, 2017). In contrast, measures that are *not* associated with the characteristics of enrolled students are instead reflective of school practices, capacity, and supports for students.

Because groups of students perform differently on various accountability measures, the selection of these measures has implications for students of different ethnic, socioeconomic, and linguistic groups, as well as the schools in which they are enrolled. Reliance on certain accountability measures will increase the likelihood of

identifying schools with low-income students, students of color, or English learners—potentially stigmatizing those groups, while only measuring one facet of school performance.

Across all three states, our analyses show that measures of achievement are closely associated with the poverty level of students, whereas measures of progress or growth are less reflective of family income. The pattern is remarkably similar across all three states. In California, accountability measures that focus on the overall performance *level* tend to distinguish between schools in the highest and lowest poverty quartiles, whereas both sets of schools perform similarly on measures of *change,* which reflect improvement from one year to the next. For example, Exhibit 3 shows that schools in the lowest poverty quartile outperform schools in the highest poverty quartile when we look only at the ELA performance *level*. But when we look at ELA *change*, both sets of schools overlap neatly, showing that neither set of school is favored by this measure. Similarly, in Florida, student achievement in ELA clearly distinguishes between the student poverty levels, whereas *learning gains* in ELA brings high- and low-poverty schools closer together. Finally, in Ohio, the *achievement* component shows very little overlap in performance among high- and low-poverty schools, but the *progress* component enables both high- and low-poverty schools to demonstrate their contributions to student learning.

Measures that put schools on an even playing field—rather than

Exhibit 3. Performance on Achievement and Growth Measures in Each State by School Poverty Quartile
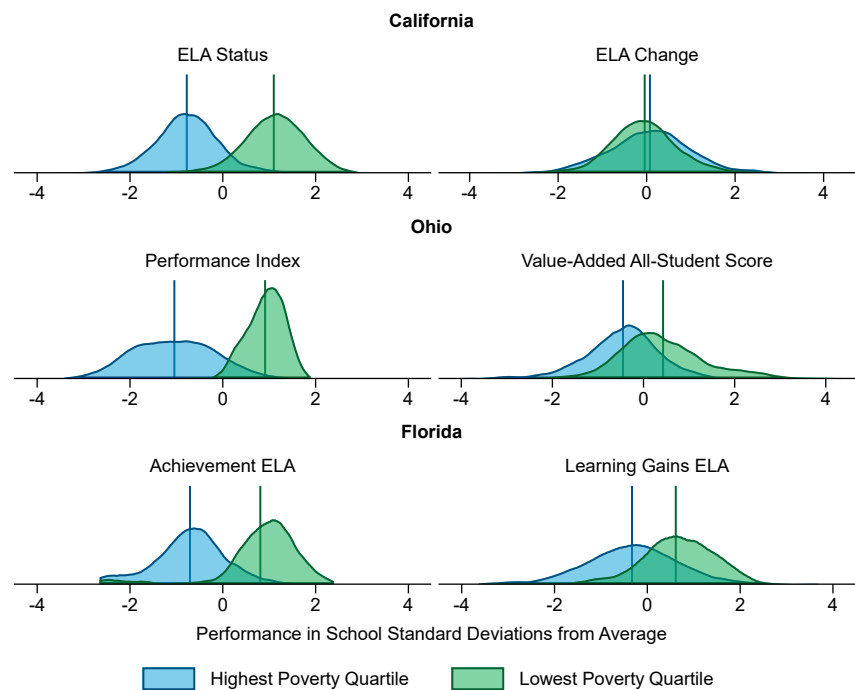
Note. Vertical lines represent the average performance for the highest and lowest poverty quartiles, respectively. The highest and lowest poverty quartiles are comprised of approximately 2,300 schools each in California, 700 schools in Florida, and 800 schools in Ohio.

simply reflecting student demographics—enable a state to most effectively target schools that need support. Arguably, an equitable and fair accountability system is one in which progress or growth measures exert a strong influence on the identification of CSI schools. Because California accounts for both *level* and *change* in all accountability measures, this state's system is the one that most comprehensively incorporates change over time. In Ohio, progress is one of the measures that exerts the greatest influence on accountability designations, helping high-poverty schools demonstrate their successes. In Florida, the influence of learning gains is generally on par with other measures. Thus, all three states have prioritized these asset-based measures.

Multiple Measure Accountability Under ESSA: Early Findings From Three States

CONCLUSION

These analyses of CSI policies in three states provide some important insights with regard to the identification of CSI schools under ESSA. In all three states, CSI schools more often lack sufficient data to be held accountable for certain indicators. As a result, CSI schools are evaluated on fewer indicators than other schools, suggesting that identification is often based on an incomplete picture of school performance. Although most schools were evaluated on multiple indicators, this was less frequently the case for those identified as requiring the most intensive supports. This raises concerns that the accountability system may not be identifying schools that are low performing when taking other important indicators into account.

In California, the suspension rate measure influenced CSI ratings in a way that was perhaps not intended by policymakers; in both Florida and Ohio, measures associated with specific school levels exerted a disproportionate influence. These results may be consistent with policymakers' intent—for example, Ohio policymakers may have intended to shine a spotlight on K–3 literacy—or they may instead be unanticipated quirks of new systems.

Finally, all three states structured systems in which the growth (or progress) measures substantially influenced CSI ratings. Because growth measures are less reflective of student characteristics—and more reflective of school quality—than are traditional academic measures, all three states have advanced equity while seeking to uphold standards of excellence.

References

American Institutes for Research (2017). *State Accountability Under ESSA: Fall 2017 Submissions*. https://www.air.org/project/state-accountability-under-essa-fall-2017-submissions

Di Carlo, M. (2019). *The false choice of growth versus proficiency.* Albert Shanker Institute.

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Wright, B. L., & Petrilli, M. J. (2017). *Rating the ratings: An analysis of the 51 ESSA accountability plans*. Fordham Institute.

[i] ESSA also requires states to identify schools for targeted support and improvement (TSI) and additional targeted support and improvement (ATSI), based on the performance of specific groups of students. However, TSI and ATSI schools are not the focus of these analyses.

[ii] In this brief, we use the term "indicator" to describe the core elements of each state's accountability system, such as achievement. Note that Ohio uses the term "component," but we have opted to use consistent terminology across all states.

Appendix: Overview of Accountability in California, Florida, & Ohio

California

To evaluate school performance, California assigns schools ratings on seven possible indicators: (1) math performance, (2) ELA performance, (3) suspension rate, (4) chronic absenteeism (elementary and middle schools only), (5) graduation rate (high schools only), (6) college and career readiness (high schools only), and (7) English learner progress (ELP). Because certain indicators apply to specific school levels, few schools in California would be held accountable for all indicators: typical elementary and middle schools could receive ratings on a maximum of five indicators, while high schools could be rated on six indicators. Moreover, a school would not be held accountable for a given indicator if the data for the indicator were available for fewer than 30 students in either the current or prior year.

For each indicator, there are five possible ratings, distinguished by different colors, with red indicating lowest performing and blue indicating highest performing. The color rating for each indicator is determined based on two measures: the level of performance and the year-to-year change in performance on the indicator. Schools performing lower for both level and change would receive a lower color rating for the indicator.

A school's CSI status is determined based on the combination of color ratings across all performance indicators. Rating combinations that result in CSI designation are as follows: red ratings for all indicators, red ratings for all but one indicator, red or orange ratings for all indicators, and five or more indicators where the majority have red ratings. In addition to being identified based on performance, high schools can be identified as CSI if the 2-year average of their combined 4- and 5-year graduation rate is below 68%.

Florida

Florida is one of many states that uses an index-based system to evaluate school performance by aggregating many different measures into a single overall score. Florida's index scores are calculated by combining four indicators that vary by school level. All school levels have three indicators in common: (a) academic achievement, (b) school quality and student success, and (c) English learner progress. Elementary and middle schools are also graded on academic progress, while high school calculations additionally include 4-year graduation rates. These indicators are each calculated from an aggregation of a larger set of more fine-grained measures, or components, that vary by school level and are graded on a scale of 0 to 80, 100, or 120 points, depending on component and school level. Because certain indicator and component combinations apply to certain school levels, very few schools would be graded on all components. Moreover, a school would not be held accountable for a given indicator if the data for the indicator were available for fewer than 10 students.

The rules for identification as a CSI school in Florida are as follows: (a) attain an index score of less than 40, (b) attain an F or D on Florida's state accountability system (also scoring less than 40), or (c) attain a graduation rate of less than 67%.

Ohio

Like many states, Ohio uses an index-based system, which combines multiple measures in a single index or score to describe the overall performance of a school. Prior to 2022 (at which point the state introduced several changes to how measures and ratings are calculated), the Ohio accountability system included 11 accountability measures, which were aggregated into six components further combined into an overall index of school

performance. These six aggregate components include: (1) achievement in ELA, math, science, and social studies; (2) progress for all students in the school, gifted students, students with disabilities, and low-performing students; (3) gap closing in ELA, math, graduation rates, and, for English learners only, progress in achieving English language progress; (4) K–3 literacy; (5) graduation; and (6) "prepared for success," which includes multiple measures of college and career readiness. Because several components were specific to a given level of schooling, almost no schools would be held accountable for all components. Moreover, a school would not be held accountable for a given indicator if the data for the indicator were available for fewer than 15 students.

Schools in the bottom 5% based on the overall index were designated for Comprehensive Support and Improvement (CSI).