State Accountability Rating Systems: A Review of School Report Cards as Indicators of School Quality

Gail L. Sunderman
Maryland Equity Project, University of Maryland (Ret.)

June 2022

National Education Policy Center

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
(802) 383-0058
nepc.colorado.edu

Acknowledgements

NEPC Staff

Faith Boninger
Publications Manager

Patricia Hinchey
Academic Editor

Elaine Duggan
Production Design

Alex Molnar
Publications Director

Kevin Welner
NEPC Director

Suggested Citation: Sunderman, G.L. (2022). *State accountability rating systems: A review of school report cards as indicators of school quality*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/report-cards

Funding: This policy brief was made possible in part by funding from the Great Lakes Center for Educational Research and Practice.

Peer Review: *State Accountability Rating Systems: A Review of School Report Cards as Indicators of School Quality* was double-blind peer-reviewed.

NEPC
NATIONAL EDUCATION POLICY CENTER

State Accountability Rating Systems: A Review of School Report Cards as Indicators of School Quality

Gail L. Sunderman
Maryland Equity Project, University of Maryland (Ret.)

June 2022

I. Executive Summary

The Every Student Succeeds Act of 2015 (ESSA) provided states with increased flexibility to design school accountability systems. A core element of the law is the requirement that states develop statewide systems allowing for meaningful differentiation among schools, and use this information to identify schools that should be the focus of improvement efforts. Individual states decide on the type of report card, or rating system, that they will use to report this information to the public.

A core premise of state school report cards is that publicizing information about how well schools and districts meet specified goals will incentivize school improvement. However, for report card rating systems to be potentially beneficial as a school improvement policy instrument, they must provide fair and valid indicators of school performance.

The increased flexibility under ESSA means that states are following different policy paths reflecting their own interests, concerns, political perspectives, and economic conditions. While states may take different approaches to measuring and reporting school performance, they have consistently adopted school report cards that collapse multiple school performance indicators into a summative rating. This is concerning because there is very limited credible research on how well a single score captures the complexity of school performance or provides information on how to improve.

Summative ratings that conflate information into a single score obscure a great deal of information about variations in school performance. They also do little to explain performance differences between or within schools or to help identify effective strategies to address low performance. Moreover, the available research evidence suggests that summative ratings fail to identify schools with high and equitable achievement, distinguishing such schools

from those with high average achievement and large achievement gaps. Indeed, available research suggests that summative ratings advantage schools serving primarily higher income students while obscuring the failure of such schools to serve *all* children.

ESSA's so-called Fifth Indicator provision allows states more flexibility to incorporate a more nuanced understanding of local contexts by adding another indicator—typically attendance, social emotional, or school climate measures. But this addition has not produced a system capturing the complexity of teaching and learning. This means that policymakers face significant challenges in designing accountability systems that take into account differences in school processes, capacity, resources, and socioeconomic variables.

It is, therefore, recommended that federal policymakers take the following four steps:

- Require that states conduct rigorous evaluations of their existing performance rating systems to determine the reliability, validity, and fairness of their summative ratings.
- Fund research on state accountability rating systems that identifies system components capable of yielding valid inferences about school performance.
- Adopt social and economic policies that address out-of-school variables that are substantially related to school performance, such as policies that increase access to health care, address the concentration of disadvantage or advantage in different neighborhoods, and expand the availability of housing and employment opportunities.
- Design federal accountability policy that promotes equitable accountability systems among states, given the reality that state political and economic variables shape divergent state systems.

Even without federal action, state policymakers can take the following two steps:

- Evaluate state accountability systems to ensure they are not misidentifying schools and potentially leading to the misappropriation of resources for school improvement.
- Adopt education, social, and economic policies that promote equitable education and address out-of-school variables related to school performance.

NEPC
NATIONAL EDUCATION POLICY CENTER

State Accountability Rating Systems: A Review of School Report Cards as Indicators of School Quality

Gail L. Sunderman
Maryland Equity Project, University of Maryland (Ret.)

June 2022

II. Introduction

School report cards, posted on state education agency (SEA) websites, are a prominent feature of federal and state accountability systems. They provide the public with information on individual schools' characteristics and performance, allowing for comparisons among them. They are also used to monitor annual school progress and to identify schools needing improvement. The premise is that providing public information about how well schools and districts are meeting identified goals will incentivize school improvement.[1] Whether report cards actually serve as a useful policy instrument, however, depends on whether they provide fair and valid indicators of school performance.

Most state school report cards arose primarily in response to the federal No Child Left Behind Act of 2001 (NCLB) requirements that states publicly report school performance. The act tied accountability to student performance on standardized achievement tests and imposed sanctions on schools falling short of requirements. Requirements stipulated that underperforming schools show steady improvement on meeting proficiency standards for specific demographic groups.[2] As more and more schools entered school improvement status and it became clear that NCLB's 2014 proficiency goal was unattainable, the act's limitations became increasingly evident and states and localities pushed back against its prescriptive nature.[3] In 2011 President Obama offered to waive key accountability mandates and the proficiency requirement for states that agreed to adopt policies the administration advocated.[4] The waiver process also introduced growth measures and the addition on non-test indicators that were incorporated into the 2015 reauthorization of the Elementary and Secondary Education Act, titled the Every Student Succeeds Act (ESSA).[5]

Current Law

ESSA changed several aspects of school accountability in the United States by offering states more opportunities to design their own systems for measuring and reporting school performance. While the law granted states more flexibility, it maintained two key elements of its predecessor: accountability indicators and annual ratings of school performance.[6] State accountability plans, at a minimum, must include the following indicators:

- Student achievement in English language arts (ELA), mathematics, and science as measured by proficiency on annual assessments.
- A measure of student growth or another "valid and reliable statewide academic indicator" that provides meaningful differentiation in school performance.
- The four-year adjusted cohort graduation rate for high schools.
- A measure of progress in achieving English language proficiency.
- At least one measure of school quality or student success (SQ/SS) that is valid, reliable, and comparable statewide. This can include various measures such as chronic absenteeism, educator or student engagement, school safety and climate, college and career readiness, opportunity to attend and complete advanced courses, or indicators such as school discipline and ninth graders on track to graduate.[7]

A core element of the law is the requirement that each state establish a system providing meaningful differentiation among all public schools in a state, with such differentiation "be based on all indicators in the state's accountability system."[8] The law does not limit the number or kinds of indicators or specify how data are to be reported, but it does require states to assign weights to each indicator, with the first three indicators weighted greater than the school quality or student success (SQ/SS) measure. Based on this system, each state must identify 1) the lowest 5% of all public schools receiving Title I funding, and 2) any public high school with a graduation rate less than 67% for Comprehensive Support and Improvement (CSI).

State Systems

Given the flexibility to choose the type of report card or rating systems used to report school performance, states have adopted the following:

A-F Rating: An A-F letter grade is assigned to schools based on the summation of accountability indicators such as student achievement test scores, growth in academic test scores, graduation rates, ACT/SAT participation and scores, and attendance rates.

1-5 Stars: Similar to the A-F rating system, a single composite rating is assigned to schools, with 1 star being the lowest and 5 stars the highest rating.

Index Rating System: A single composite rating is assigned, but the rating scale varies. It may be a numerical rating (1-10 rankings, 1-100 rankings, and so on) or a percentile rating.

Federal Tiers of Support: Only schools that fall into the "lowest performing schools" category are identified as meeting the federal requirements for intervening in low-performing schools. Other schools are not ranked or rated.

Descriptive: Text-based labels are assigned based on performance indicators. For example, Delaware translates a summative index score into an overall text-based identification (exceeds, meets, or meets few expectations).[9] Illinois uses four descriptive summative ratings (lowest performing, underperforming, commendable, and exemplary), with exemplary schools meeting the standard of an "all students" index score at or above 80.12 (elementary schools) or 85.42 (high schools).[10]

Dashboard: Performance results on multiple indicators are assigned and reported; states may assign a summative rating to each indicator. For example, California, the only state to use a dashboard, includes summative ratings on 14 indicators, including absenteeism, suspension rate, ELA, mathematics, local climate survey, and so on.[11]

Table 1 shows the type of rating system each state adopted when reporting was first required in 2018 and which they were using in 2021. While some states switched systems, the number of states in each category remained remarkedly similar, with one exception: The number of states adopting the minimalist Federal Tiers of Support system doubled, from 5 to 10. California was the only state to adopt a dashboard.

Table 1: Comparison of Type of Rating System Adopted by States in 2018 and 2021

Type of Rating System	States 2018	States 2021	# of States 2018	# of States 2021
A-F Rating System	AZ FL IN LA MS NM NC OH OK TN TX UT	IN LA MI MS NC OH OK TN TX UT	12	10
Descriptive Rating System	DE IL KS ME MA MN NE NJ SC VT WV	DE IL KS ME MA MN NE NJ SC VT WV WY	11	12
Index Rating System	AK AR CT GA HI IA MI MO SD WA WI WY	AK AR CT FL GA HI IA MO NM ND SD WA WI	12	13
1-5 Star Rating System	DC KY MD NV RI	DC KY MD NV RI	5	5
Dashboard		CA	0	1
Federal Tiers of Support (Identifies only Schools in the Lowest Tier of Performance)	AL CO NH NY VA	AL AZ CO ID MT NH NY OR PA VA	5	10
No Summative Rating System	CA ID ND OR PA		5	0
Other Rating System	MT		1	0

Source: Education Commission of the States. *50-state comparison: States' school accountability systems*. For January 2021 data: Retrieved on January 4, 2022, from https://reports.ecs.org/comparisons/states-school-accountability-systems-2021. For January 2018 data: personal communication on January 14, 2022 with ECS staff.

These rating systems share a similar feature: they collapse multiple school performance indicators into a summative or aggregate measure.[12] Some employ five scale rating schemes, such as A-F grades or 5 stars, and others use a composite index scale (for example, 1-100 ratings) or "descriptive" rankings that, for example, range from "exemplary school" to "lowest performing." The California School Dashboard does not include an overall summative rating, but instead includes summative ratings for each of the indicators it tracks.

Issues and Questions

On the surface, summative ratings are attractive as a policy instrument because they appear to provide concise and easily understood measures of school quality as ESSA requires. However, collapsing various indicators into a composite score raises questions about how effectively the rating summarizes and communicates useful information on a school's performance. Moreover, it is difficult to determine from a single rating whether the rating systems are promoting intended policy goals.

This brief explores two key questions about current state accountability rating systems. First, given that state policymakers have considerable flexibility under ESSA to design, amend, and tweak their accountability rating systems and to ground them in local priorities, what variables explain different state responses? Second, how well does a single score capture the complexity of school performance and provide reliable information on how to improve?

Review of the Literature

ESSA was greeted with optimism that greater state flexibility and the use of multiple measures would lead to more meaningful accountability systems. A plethora of reports and recommendations followed its enactment, suggesting how states might take advantage of the new flexibility to design systems leading to more meaningful accountability.[13] But whether states would actually experiment with new approaches, or even whether they had the capacity to engage in robust policy design and implementation, was unclear—as was the question of whether summative ratings capture the complexity of schooling. Research has since offered some insights in these areas.

The Influence of Local Context

Exploring the impact of local context, Portz & Beauchamp (2022) examined multiple state approaches to educational accountability under ESSA.[14] The authors found that states "are following *different* policy paths that will increasingly reflect their own interests and concerns, including their own political perspectives and experiences."[15] They found that state-level demographics correlated with the types of assessment policies adopted by states: Political variables (liberal versus conservative) were associated with whether states adopted alternative strategies and approaches, while economic variables (per-pupil spending and percentage of students in poverty) were associated with whether states maintained a more traditional focus on student performance measures.[16] Politically liberal states were more likely to emphasize SQ/SS measures and to incorporate such alternative indicators as growth measures.

States with a more conservative leaning and facing economic challenges maintained a focus on student test scores, while well-resourced states faced less pressure to change policies or programs.

Research on the earlier NCLB act also demonstrated that local political and policy context influenced the development, interpretation, and outcome of state responses to federal accountability mandates. A study examining the role of historical and political context in shaping assessment policy in Nebraska and Virginia, for example, found that the political culture in both strongly influenced their assessment systems.[17] In Nebraska, an historical culture rooted in local action and collaboration among the board of education, legislature, and executive branch influenced the design process. As a result, the state experienced local support for implementation, and it delayed a shift to a state standardized assessment system in favor of local assessments. In contrast, Virginia, with a tradition of centralization and top-down accountability, implemented a top-down policy model that emphasized standardized testing and constrained resources and opportunities for policy transformation at lower policy levels.

In addition, although policies with flexibility may allow for states to improve systems by updating or correcting them, changes to accountability formulas can affect the distribution of school ratings while not necessarily reflecting actual changes in school performance. A 2016 study indicated, for example, that after Oklahoma changed its formula the number of C schools dropped from 21% to 5%, and the number of F schools increased from 8% to 53%—even though school demographics remained similar and average math and reading achievement were stable.[18] Accountability rating systems may also be designed or adjusted based on preconceived perceptions and expectations.[19] For example, Florida first adopted an A-F system in 1999, when 12% of schools received an A rating. Following a series of rule changes, the percentage of schools receiving A's rose to 53% in 2005. The changes were made to better align with preconceived perceptions of what constituted a good or bad school, suggesting that ratings changed in part because the criteria changed. In fact, a simulation showed that if earlier rating rules had been applied, grades would not have differed significantly.[20] In addition, ratings can be manipulated. Under instructions from Indiana's state superintendent in 2012, staffers took advantage of a loophole in the state's system to benefit a charter school founded by a campaign donor. By eliminating the scores of some student groups, the original rating was changed from a C grade to an A.[21]

Changes to accountability formulas can affect the distribution of school ratings while not necessarility reflecting actual changes in school performance.

These studies suggest that states take different approaches to measuring accountability and prioritizing different kinds of information. Because school ratings are highly visible, they can exert a strong influence on public perceptions of the schools, whether or not they represent accurate assessments of school performance. Ratings also have implications for how information is used and interpreted. For example, weighting growth measures more strongly than achievement can alter the focus of policymakers, educators, and parents and shift the allocation of resources. Incorporating school climate surveys can direct attention to the teaching and learning environment; in contrast, an emphasis on achievement scores may simply pressure teachers to raise test scores. Different approaches can also amplify or mask

the performance of different student populations, and do not necessarily provide information about the quality of instruction a student receives.

School Report Cards as a Measure of School Quality

Little research on summative rating systems exists, despite their proliferation. What does exist casts doubt on whether a single score can capture the complexity of school performance. While there is theoretical literature on the design, content, and presentation of school rating systems,[22] there is limited empirical evidence indicating whether, in practice, ratings correlate with desired student outcomes or have unintended effects. What we do know comes primarily from a small set of papers and reports on summative grading systems (primarily A-F systems) and one report on the California School Dashboard.

Within and Between School Differences

What do summative, single score systems tell us about within- and between-school differences? A study of the Oklahoma A-F rating system examined the "informational significance" of its grades—that is, "the ability of the grade to yield meaningful and useful information about achievement differences within and between schools."[23] The study found that while average math and reading scores were higher in A and B schools compared to C, D, and F schools, the averages masked unequal test score distribution and important information about achievement gaps. The largest achievement gaps were in schools ranked as the most effective. For example, test score gaps for low-income students were found in all schools, but the gaps were smaller in C, D, and F schools. The largest minority gaps in reading and math (over one standard deviation) were found in B rated schools. The authors concluded that, "grades do not sort out schools with high and equitable achievement from schools with high average achievement and large achievement gaps," which can create inaccurate judgements about school quality and diminish the usefulness of letter grades for low-income and minority students.[24]

Because letter grades obscure achievement differences within schools, the authors show that the distribution of letter grades would change drastically if the state changed its accountability formula. For example, if the state assigned the same weight to achievement gaps as it assigned to overall achievement, several D and F schools would become B or C schools, and A and B schools would become C and D schools. Since ESSA requires intervention only in the lowest performing schools, higher ranked schools in Oklahoma were exempt from any pressure to improve the education of its low-income and minority students. To meaningfully represent school performance, the A to F grade would need to capture within school fluctuations between subgroups of students. Otherwise, the authors argue, the summative ratings provided little information of instructional value to teachers and administrators, and they are no help in measuring the effectiveness of improvement strategies or interventions.[25]

The authors identified three components of A-F grading systems that contribute to inequitable results. First, these systems use proficiency scores based on achievement tests to calculate student achievement and student growth. However, aggregating proficiency scores into a single metric and using them to rank schools embeds inaccuracies in the results, because

that metric does not reflect the many variables that contribute to school performance.[26] Second, the methods used to calculate growth can hide the achievement of low-income and minority students. Since there is no one method of measuring growth, the type and use of growth measures and what each says about a school varies.[27] And third, A-F grades do not reveal anything about the dynamic nature of schools and student learning. Instead, grades treat teaching and learning as a fixed process.[28]

Other studies report similar findings. A study examining the impact of being rated a "failing" school, defined as receiving a D or F on the New York City Progress Report (NYCPR), concluded that identifying failing schools by a single summary statistic was inappropriate, suggesting that "the ability to educate students performing at grade-level and the ability to educate students performing below grade-level may be separate dimensions."[29] The author describes the NYCPR system as aspirational in that the indicator weights reflected the best judgement of stakeholders in what they would like a quality school to be. While the NYCPR system included a range of indicators, by collapsing all the indicators into a unidimensional metric, any nuance gained from using multiple indicators was "lost from both the measurement perspective and from the public accountability perspective."[30]

Unmediated Outcomes

Summative ratings likely obscure many variables that contribute to school performance that are outside the control of educators or are unknown. Accountability rating systems implicitly assume that the school itself is primarily responsible for student performance even though empirical evidence shows that school effects typically account for less than 30% of student academic performance.[31] The Oklahoma study, for example, found that student differences accounted for 72% of the variance in reading achievement scores and 70% in math. Schools, on the other hand, accounted for 28% of the reading variance and 30% of the math variance.[32]

Summative ratings obscure the well documented relationship between student achievement scores and demographic variables, most notably race and socioeconomic status. Research shows that school composition matters: Attending a high-poverty or a highly segregated African American school has a negative effect on a student's achievement outcomes, above and beyond the effect of individual poverty or minority status.[33] When student data is aggregated to the school level, the relationship between demographic variables and outcomes becomes even stronger.[34] If school composition is not taken into account, summative ratings over-identify schools serving low-income and minority students as low performing and miss the contribution schools make to student achievement. Other variables that may contribute to achievement such as teacher characteristics and knowledge, instructional support, or school characteristics such as class size, are largely missing from these summative rating systems.

The NYCPR analysis found statistically significant differences between schools receiving passing grades (A, B, or C) and those receiving failing grades (D or F): Schools serving special education students, Black or Hispanic students, and Title I students were more likely to receive a failing grade.[35] To reflect each school's contribution to student academic progress, NYCPR heavily weighted the academic growth indicator. Nonetheless, schools receiv-

ing lower grades (D or F) had notably higher concentrations of Black or Hispanic students, low-income students, and students receiving special education services compared to schools receiving A, B, or C. When the authors controlled for demographic context of schools, a different pattern emerged. These findings suggest that "the mis-specification of weights under the NYCPR may be obscuring actual differences in school quality for students of different demographics."[36]

Likewise, an analysis of the Maryland five-star rating system documented the relationship between report card rating and the percentage of a school's economically disadvantaged students.[37] No high-poverty school earned a five-star rating under the Maryland system. When the authors adjusted ratings to account for economic disadvantage, the number of five-star schools increased.

In general, findings on A-F and five-star rating systems suggest that any system that collapses school performance into a single composite rating does not provide policymakers or parents the nuanced and accurate information they need to understand achievement patterns in schools.[38]

Multiple Indicators of School Performance

The theory behind dashboards is that a system of multiple measures of school performance provides more—and better—information than a single rating. California is the only state using a dashboard.[39] While it does not assign a single, summative rating to schools, it does provide an aggregated rating for each performance category. These include assigning a performance color (red, orange, yellow, green, blue) for some categories, and noting whether the standard was met on others.

An analysis of the California dashboard found little evidence that the system untangled the relationship between student demographics and school ratings.[40] Nor did the system take into account differences in the resources, policies, and practices of schools that play a significant role in producing outcomes. While the study notes that the system did a much better job of including a range of performance measures when compared to the system it replaced, challenges remained. These included: (1) few parents were aware of, or reported having used the dashboard, with gaps between more and less educated voters; (2) the dashboard did not offer comparisons between schools, and school ratings were highly related to student demographics in a way that unfairly punished schools serving more disadvantaged students; (3) the method used to calculate growth was poorly aligned with what research says about how to use growth models to isolate school effectiveness; and (4), the dashboard gave insufficient weight and priority to the performance of student groups, thereby allowing schools to have high overall ratings even through subgroup performance was weak.[41]

The study suggests that the inability of the system to make appropriate inferences about school performance stemmed in part from its heavy reliance on achievement measures, with math and reading test scores accounting for at least 50% of the ratings and raised questions about how accurately it captured school performance. A follow-up a year later found that while the state made a number of substantive, technical, and stylistic changes, many challenges persisted.[42] Among these were the inability to compare schools on multiple indicators

or to compare schools with similar demographics, resulting in a tendency for more affluent schools to look better and less affluent schools to look worse. The state also had not changed its approach to measuring student growth or remedied inequities in parental awareness and use of the dashboard.

In Summary

These studies suggest that a single, composite measure of school performance is a poor indicator of school quality and does not provide the information needed to understand performance differences between and within schools.[43] Aggregating data to the school level leaves summative measures open to misinterpretation when used to explain student-level differences in performance.[44] Such systems tend to disproportionally harm some schools (poor and minority) and inflate the quality of other schools (wealthy and white) based on demographics and income. In addition, they cast doubt on the ability of indicator dashboards to remedy the flaws inherent in summative ranking systems.

Recent Developments

The COVID-19 pandemic brought unprecedented challenges to education systems nationwide and will present further challenges to accountability rating systems in the years to come.[45] Attempting to limit the spread of the virus, states closed schools and shifted to online teaching and learning, provisions that disproportionally affected low-income and students of color.[46] As a result, all 50 states received waivers from the U.S. Department of Education from state accountability requirements, including ESSA assessments, school identification, and related reporting requirements for the 2020-21 school year.[47]

While waivers were welcome relief in the short-term, the longer term impact of COVID-19 on the measurement of indicators and interpretation of summative ratings remains.[48] The transition from face-to-face learning to digital or hybrid learning means that students had different, often inequitable, learning experiences affected by their access to Internet service and computers, ability to adapt to and engage with online learning, and/or level of participation. Teachers likely approached online instruction differently depending on their skill and knowledge of technology and typically without standardized school or district guidelines for their efforts. The variability in instruction is likely to create larger variability in test scores within and across schools that, if not accounted for, will affect the reliability and validity of achievement scores—a major indicator used in summative ratings. Achievement scores are also likely to be impacted depending on how interrupted learning is addressed and how much time teachers spend on review rather than on new material. Growth scores, which compare test scores to similar scores in prior years, may not be calculated for several years. Both test and growth scores will particularly impact ratings for schools serving low-income students, students of color, and students with disabilities.

Discussion and Analysis

Despite the proliferation of summative rating systems across the states, there is very little empirical research on the effects of summative ratings on student performance or school and teacher practices. The studies reviewed here suggest that a rating system using a single score does not capture the complexity of school performance or provide information on how to improve. Several limitations have been identified. Summative ratings do not sort out schools with high and equitable achievement from schools with high average achievement and large achievement gaps. Indeed, available research suggests that they advantage schools serving primarily higher income students while obscuring the failure of such schools to serve *all* children. By conflating information into a single score, summative ratings obscure a great deal of information about variations in school performance and have not been very useful in explaining performance differences between or within schools or in suggesting how to address them. Allowing states more flexibility to incorporate a more nuanced understanding of local contexts by adding other indicators— attendance, social emotional, and school climate measures, for example—has not yet produced a system capturing the complexity of teaching and learning.

ESSA was meant to fix problems in the earlier NCLB accountability system, but it produced only technical fixes influenced by political and economic dynamics in each state. This review suggests that these technical fixes are inadequate, allowing for blunt summative ratings that fail to capture the contributions schools actually make to student performance, particularly those with high concentrations of poverty and of minority students.

Despite known weaknesses, accountability ranking systems persist for a number of reasons. They are useful to bureaucracies as a management tool because a single indicator of school quality is widely perceived as providing a clear, concise, and easily understood measure of school performance. There is also a sense that accountability is here to stay, and that the way forward is to tweak current systems by making better use of measurement methodology.[49] However, since accountability is in part a political process, it is not clear that technical fixes can lead to systems that are more reliable, fair, or valid.

By conflating information into a single score, summative ratings obscure a great deal of information about variations in school performance.

The challenge for policymakers is twofold. The short-term challenge is how to interpret summative ratings so they account for the interrupted learning that some students experienced during the COVID-19 pandemic. Summative ratings, which rely on achievement and growth scores, are unlikely to take into account the variability in student learning experiences during the pandemic. The longer term challenge is how to design an accountability system that takes into account differences in school capacity, resources, and processes. Inputs, an important aspect of school performance, are not included in accountability systems or reported on school report cards, although financial and budget information may be reported elsewhere. By focusing primarily on outputs (high test scores in math and reading, attendance, graduate rates, for example), summative ratings likely reflect uneven inputs since access to curriculum, diversity of textbooks, adequate staffing, materials and equipment,

technology, and facilities differ among schools. Focusing on these conditions could be useful in directing resources and attention. The expansion of rating systems to include inputs, often referred to as "opportunity to learn" standards, could provide a more nuanced appraisal of school performance.[50]

Finally, research shows that neighborhood social context and school composition substantially affect school performance.[51] Both racial and socioeconomic segregation is associated with poor academic performance. To achieve educational equity and improve student outcomes means acknowledging and addressing the role that segregation by both race and income plays in school performance. Summative ratings are limited indicators of student learning and can misidentify schools, potentially leading to misappropriation of resources for school improvement.

Recommendations

It is recommended that federal policymakers take the following four steps:

- Require that states conduct rigorous evaluations of their existing performance rating systems to determine the reliability, validity, and fairness of their summative ratings.
- Fund research on state accountability rating systems that identifies system components capable of yielding valid inferences about school performance.[52]
- Adopt social and economic policies that address out-of-school variables that are substantially related to school performance, such as policies that increase access to health care, address the concentration of disadvantage or advantage in different neighborhoods, and expand the availability of housing and employment opportunities.[53]
- Design federal accountability policy that promotes equitable accountability systems among states, given the reality that state political and economic variables shape divergent state systems.

Even without federal action, state policymakers can take the following two steps:

- Evaluate state accountability systems to ensure they are not misidentifying schools and potentially leading to the misappropriation of resources for school improvement.
- Adopt education, social, and economic policies that promote equitable education and address out-of-school variables related to school performance.

Notes and References

1 Figlio, D.N. & Ladd, H.F. (2007). School accountability and student achievement. In H.F. Ladd & E. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 166-182). New York: Routledge.

Smith, M.S. & O'Day, J.A. (1993). Systemic school reform and educational opportunity. In S.H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 250-312). San Francisco: Jossey-Bass.

Schneider, A. & Ingram, H. (1990). Behavioral assumptions of policy tools. *Journal of Politics, 52*(2), 510-529. Retrieved April 18, 2022, from http://dx.doi.org/10.2307/2131904

2 Mathis, W.J. & Trujillo, T.M. (2016). *Lessons from NCLB for the Every Student Succeeds Act*. Boulder, CO: National Education Policy Center. Retrieved February 3, 2022, from http://nepc.colorado.edu/publication/lessons-from-NCLB

Manna, P. (2011). *Collision course: Federal education policy meets state and local realities*. CQ Press.

Henrich, R. & Sunderman, G.L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement and why we may retain it anyway. *Educational Researcher, 38*(5), 353-364. Retrieved March 23, 2022, from https://doi.org/10.3102/0013189X09339055

3 Porter, A.C., Linn, R.L., & Trimble, C.S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice, 24*(4), 32-39. Retrieved March 3, 2022, from https://doi.org/10.1111/j.1745-3992.2005.00021.x

Balfanz, R., Legters, N., West, T.C., & Weber, L.M. (2007). Are NCLB's measures, incentives and improvement strategies the right ones for the nation's low-performing schools? *American Educational Research Journal, 44*(3), 593. Retrieved March 3, 2022, from https://doi.org/10.3102/0002831207306768

Sunderman, G.L. & Kim, J. (2007, May). The expansion of federal power and the politics of implementing the No Child Left Behind Act. *Teachers College Record, 109*(5), 1057-1085.

Sunderman, G.L., Ed. (2008). *Holding NCLB accountable: Achieving accountability, equity, & school reform*. Thousand Oaks, CA: Corwin Press.

Valli, L., Croninger, R.G., Chambliss, M.H., Graeber, A.O., & Buese, D. (2008). *Test driven, high-stakes accountability in elementary schools*. New York: Teachers College Press.

4 U.S. Department of Education. (2011). *Letter from the Education Secretary, September 23, 2011*. Retrieved January 25, 2022, from https://www2.ed.gov/policy/gen/guid/secletter/110923.html

5 U.S. Department of Education (2015). *Every Student Succeeds Act (ESSA)*. Retrieved January 17, 2022, from https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

6 U.S. Department of Education (2015). *Every Student Succeeds Act (ESSA)*. Retrieved January 17, 2022, from https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

7 U.S. Department of Education (2015). *Every Student Succeeds Act (ESSA)*. 129 STAT. 1837 (C). Retrieved January 17, 2022, from https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

8 U.S. Department of Education (2015). *Every Student Succeeds Act (ESSA)*. 129 STAT. 1837 (C). Retrieved January 17, 2022, from https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

9 Delaware Department of Education (2017, April 3). *State template for the consolidated state plan under the Every Student Succeeds Act: CCSSO template for revised template questions embedded* (p. 50). Retrieved March 3, 2022, from https://education.delaware.gov/wp-content/uploads/2020/07/essa_200620_consolidated_state_plan.pdf

10 Illinois State Board of Education. *Support & accountability summative designations*. Retrieved January 27, 2022, from https://www.isbe.net/Pages/Summative-Designations.aspx

11 California School Dashboard. Retrieved January 27, 2022, from https://www.caschooldashboard.org

12 For information on indicators and weights used in each rating system by state, see Education Commission of the States. *50-state comparison: States' school accountability systems*. For January 2021 data: Retrieved January 4, 2022, from https://reports.ecs.org/comparisons/states-school-accountability-systems-2021

13 See, for example: Darling-Hammond, L., Bae, S., Cook-Harvey, C.M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E.L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. Palo Alto: Learning Policy Institute. Retrieved January 4, 2022, from https://learningpolicyinstitute.org/product/pathways-new-accountability-through-every-student-succeeds-act

The Education Trust (2016). *The Every Student Succeeds Act: What's in it? What does it mean for equity? Overview*. Retrieved January 11, 2022, from https://edtrust.org/wp-content/uploads/2014/09/What-is-in-ESSA-Overview.pdf

Rouland, K. (2018). *Exploring equity issues: How ESSA provides the opportunity to target inequitable education practices and outcomes*. Bethesda, MD: MAEC. Retrieved March 1, 2022, from https://maec.org/wp-content/uploads/2018/06/Exploring-Equity-ESSA-and-Equity.pdf

Schanzenbach, D.W., Bauer, L., & Mumford, M. (2016). *Lessons for broadening school accountability under the Every Student succeeds Act*. Washington, DC: The Hamilton Project. Retrieved March 1, 2022, from https://www.hamiltonproject.org/assets/files/lessons_school_accountability_essa.pdf

Martin, C., Sargrad. S., & Batel, S. (2016). *Making the grade: A 50-state analysis of school accountability systems*. Washington, DC: Center for American Progress. Retrieved January 18, 2022, from https://cdn.americanprogress.org/wp-content/uploads/2016/05/17094420/AccountabilityLandscape-report2.pdf

14 Portz, J. & Beauchamp, N. (2020). Educational accountability and state ESSA plans. *Educational Policy, 1*(31). Retrieved January 4, 2022, from https://doi.org/10.1177/0895904820917364

15 Portz, J. & Beauchamp, N. (2020). Educational accountability and state ESSA plans. *Educational Policy, 1*(31). Retrieved January 4, 2022, from https://doi.org/10.1177/0895904820917364

16 Pollical variables focused on the ideological orientation of state policymakers and the development and innovation of public policy. It was measured using a policy liberalism index and political party of the governor in 2017. The hypothesis was that more liberal states are more likely to explore alternative indicators, while conservative states are more likely to focus on existing and traditional test score indicators. Economic variables focused on the resources available to policymakers and was measured by percent of students in poverty and per pupil expenditures. The hypothesis was that states with more economic resources (fewer students in poverty and higher per pupil expenditures) are more likely to consider alternative accountability strategies while states with fewer economic resources are more likely to focus on existing and traditional test score indicators.

17 Ruff, R.R. (2019). State-level autonomy in the era of accountability: A comparative analysis of Virginia and Nebraska education policy through No Child Left Behind. *Education Policy Analysis Archives, 27*(6). Retrieved January 11, 2022, from http://dx.doi.org/10.14507/epaa.27.4013

18 Adams, C.M., Forsyth, P.B., Mwavita, M., Barnes, L.L., & Khojasteh, J. (2016). An empirical test of Oklahoma's A-F school grades. *Education Policy Analysis Archives, 24*(4). Retrieved January 3, 2022, from http://dx.doi.org/10.14507/epaa.v24.2127

19 Tanner, J. (2016). *The A-F accountability mistake: The Texas accountability series*. Austin, TX: The Texas Association of School Administrators. Retrieved January 18, 2022, from https://www.roscoe.esc14.net/upload/page/0086/docs/TASA%20A-F-Essay.pdf

20 Rouse, C.E., Hannaway, J., Goldhaber, D., Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251-81. Retrieved January 18, 2022, from https://doi.org/10.1257/pol.5.2.251

DiCarlo, M., (2013, March 5). *Why did Florida schools' grades improve dramatically between 1999 and 2005?* Boulder, CO: National Education Policy Center. Retrieved January 18, 2022, from https://nepc.colorado.edu/blog/why-did-florida-schools'-grades-improve-dramatically-between-1999-and-2005

21 Layton, L. (2013, August 3). A-F systems for grading public schools get new scrutiny. *The Washington Post*. Retrieved February 4, 2022, from https://www.washingtonpost.com/local/education/a-to-f-grading-systems-for-public-schools-get-new-scrutiny/2013/08/03/03533aa2-fbab-11e2-a369-d1954abcb7e3_story.html

22 Dalton, B. (2017). *The landscape of school rating systems.* Research Triangle Park, NC: RTI Press. Retrieved April 19, 2022, from https://files.eric.ed.gov/fulltext/ED582356.pdf

Koretz, D. (2015). Adapting educational measurement to the demands of test-based accountability. *Measurement: Interdisciplinary Research and Perspectives, 13*(1), 1-25. Retrieved April 28, 2022, from https://doi.org/10.1080/15366367.2015.1000712

Polikoff, M.S., McEachin, A.J., Wrabel, S.L., & Duque, M. (2013). The waive of the future? School accountability in the waiver era. *Educational Researcher, (43)*1, 45-54. Retrieved April 28, 2022, from https://doi.org/10.3102/0013189X13517137

23 Adams, C.M., Forsyth, P.B., Ware, J., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*(7), 1-31. Retrieved January 18, 2022, from https://www.tcrecord.org/content.asp?contentid=20925

24 Adams, C.M., Forsyth, P.B., Ware, J., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*(7), 1-31. Retrieved January 18, 2022, from https://www.tcrecord.org/content.asp?contentid=20925

25 Adams, C.M., Forsyth, P.B., Mwavita, M., Barnes, L.L., & Khojasteh, J. (2016). An empirical test of Oklahoma's A-F school grades. *Education Policy Analysis Archives, 24*(4). Retrieved January 3, 2022, from http://dx.doi.org/10.14507/epaa.v24.2127

26 For a detailed discussion on variables that contribute to school performance, see White, K.R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461-481. Retrieved March 29, 2022, from https://doi.org/10.1037/0033-2909.91.3.461

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417-453.

27 For more information on measuring growth, see Data Quality Campaign (2019). *Growth data: It matters, and it's complicated.* Washington, DC: Author. Retrieved January 3, 2022, from https://dataqualitycampaign.org/wp-content/uploads/2019/04/DQC-Growth-Data-Resources.pdf

28 For an in-depth analysis of A-F grading systems and teaching and learning, see Murray, K. & Howe, K.R. (2017). Neglecting democracy in education policy: A-F school report card accountability systems. *Education Policy Analysis Archives, 25*(109). Retrieved January 18, 2022, from http://dx.doi.org/10.14507/epaa.25.3017

29 Tsui, J. (2018). *The effect of grading in school accountability systems: An investigation using propensity scores in second-order growth modes*. A dissertation submitted to UCLA (p. 72). Retrieved January 18, 2022, from https://escholarship.org/uc/item/3rg4h5hx

30 Tsui, J. (2018). *The effect of grading in school accountability systems: An investigation using propensity scores in second-order growth modes*. A dissertation submitted to UCLA. Retrieved January 18, 2022, from https://escholarship.org/uc/item/3rg4h5hx

31 Rowan, B., Correnti, R. & Miller, R.J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Nye, R., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.

32 Adams, C.M., Forsyth, P.B., Mwavita, M., Barnes, L.L., & Khojasteh, J. (2016). An empirical test of Oklahoma's A-F school grades. *Education Policy Analysis Archives, 24*(4). Retrieved January 3, 2022, from http://dx.doi.org/10.14507/epaa.v24.2127

Adams, C.M., Forsyth, P.B., Ware, J., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*,(7), 1-31. Retrieved January 18, 2022, from https://www.tcrecord.org/content.asp?contentid=20925

33 Borman, G. & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's Equality of Educational Opportunity data. *Teachers College Record, 112*(5), 1201-1246.

See also, Angrist, J., Hull, P., Pathak, P.A. & Walters, C.R. (2021). *Race and the mismeasure of school quality*. Cambridge, MA: National Bureau of Economic Research. Retrieved March 28, 2002, from http://www.nber.org/papers/w29608

34 White, K.R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461=481. Retrieved March 29, 2022, from https://doi.org/10.1037/0033-2909.91.3.461

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417-453.

35 Tsui, J. (2018). *The effect of grading in school accountability systems: An investigation using propensity scores in second-order growth modes*. A dissertation submitted to UCLA (p. 72). Retrieved January 18, 2022, from https://escholarship.org/uc/item/3rg4h5hx

36 Tsui, J. (2018). *The effect of grading in school accountability systems: An investigation using propensity scores in second-order growth modes*. A dissertation submitted to UCLA (p. 72). Retrieved January 18, 2022, from https://escholarship.org/uc/item/3rg4h5hx

37 Wohn, C. & Boberiene, L. (2019). *Looking at school performance from a different perspective: A Maryland report card reflecting students served.* Baltimore, MD: Baltimore Education Research Consortium. Retrieved April 4, 2022, from https://files.eric.ed.gov/fulltext/ED609144.pdf

38 Adams, C.M., Forsyth, P.B., Ware, J., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*,(7), 1-31. Retrieved January 18, 2022, from https://www.tcrecord.org/content.asp?contentid=20925

39 California School Dashboard. https://www.caschooldashboard.org

40 Polikof, M.S., Korn, S., & McFall, R. (2018). *In need of improvement? Assessing the California Dashboard after one year*. Palo Alto, CA: Stanford University, Policy Analysis for California Education. Retrieved January 3, 2022, from https://gettingdowntofacts.com/sites/default/files/2018-09/GDTFII_Report_Polikoff.pdf

Polikof, M.S. (2019). *Gauging the revised California School Dashboard: Evidence from the 2019 PACE/USC Rossier Voter Poll*. Palo Alto, CA: Policy Analysis for California Education. Retrieved, January 4, 2022, from https://files.eric.ed.gov/fulltext/ED594791.pdf

41 Polikof, M.S., Korn, S., & McFall, R. (2018). *In need of improvement? Assessing the California Dashboard after one year*. Palo Alto, CA: Stanford University, Policy Analysis for California Education. Retrieved January

3, 2022, from https://gettingdowntofacts.com/sites/default/files/2018-09/GDTFII_Report_Polikoff.pdf

Polikof, M.S. (2019). *Gauging the revised California School Dashboard: Evidence from the 2019 PACE/USC Rossier Voter Poll.* Palo Alto, CA: Policy Analysis for California Education. Retrieved January 4, 2022, from https://files.eric.ed.gov/fulltext/ED594791.pdf

42 Polikof, M.S. (2019). *Gauging the revised California School Dashboard: Evidence from the 2019 PACE/USC Rossier Voter Poll.* Palo Alto, CA: Policy Analysis for California Education. Retrieved January 4, 2022, from https://files.eric.ed.gov/fulltext/ED594791.pdf

43 Adams, C.M., Forsyth, P.B., Ware, J., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*(7), 1-31. Retrieved January 18, 2022, from https://www.tcrecord.org/content.asp?contentid=20925

44 Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

45 Huck, C. & Zhang, J. (2021). Effects of COVID-19 pandemic on K-12 education: A systematic literature review. *Educational Research and Development Journal (24),*1, 53-84.

46 Haderlein, S.K., Saavedra, A.R., Polikoff, M.S., Silver, D., Rapaport, A., & Garland, S. (2021). Disparities in educational access in the time of COVID: Evidence from a nationally representative panel of American families. *AERA Open, 7*(1), 1-21. Retrieved February 8, 2022, from https://doi.org/10.1177/23328584211041350

47 Office of Elementary & Secondary Education (n.d). *State requests for waivers of ESEA provisions for SSA-administered programs.* Retrieved January 17, 2022, from https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/essa-state-plans-assessment-waivers/?utm_content=&utm_medium=e-mail&utm_name=&utm_source=govdelivery&utm_term=

48 Middleton, K.V. (2020). The longer-term impact of COVID-19 on K-12 student learning and assessment. *Educational Measurement: Issues and Practice, 39*(3), 41-44. Retrieved February 6, 2022, from https://doi.org/10.1111/emip.12368

49 For different perspectives, see for example, Harris, D.N. & Liu, L. (2021). *What get measured gets done: Principles for performance measurement in school accountability systems and how states can meet them.* Tulane University: Education Research Alliance NOLA. Retrieved April 28, 2022, from https://educationresearchallianceola.org/files/publications/061221-Harris-Liu-What-Gets-Measured-Gets-Done-Technical-Paper.pdf

Koretz, D. (2015). Adapting educational measurement to the demands of test-based accountability. *Measurement: Interdisciplinary Research and Perspectives, 13*(1), 1-25. Retrieved April 28, 2022, from https://doi.org/10.1080/15366367.2015.1000712

Polikoff, M.S., McEachin, A.J., Wrabel, S.L., & Duque, M. (2013). The waive of the future? School accountability in the waiver era. *Educational Researcher, (43)*1, 45-54. Retrieved April 28, 2022, from https://doi.org/10.3102/0013189X13517137

50 Bae, S. (2018). Redesigning systems of school accountability: A multiple measures approach to accountability and support. *Education Policy Analysis Archives, 26*(8). Retrieved April 25, 2022, from http://dx.doi.org/10.14507/epaa.26.2920

51 Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., et al. (1966). *Equality of educational opportunity* (No. OE-38001). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Education Statistics.

Benson, J. & Borman, G. (2010). Family, neighborhood, and school settings across seasons: When do socioeconomic context and racial composition matter for reading achievement growth of young children? *Teachers College Record, 112*(5), 1338-1390.

Borman, G. & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's Equality of Educational Opportunity data. *Teachers College Record, 112*(5), 1201-1246.

Wei, Y.D., Xiao, W., Simon, C.A., Liu, B. & Ni, Y. (2018). Neighborhood, race and educational inequality. *Cities, 73*, 1/13. Retrieved April 25, 2022, from https://doi.org/10.1016/j.cities.2017.09.013

52 Adams, C.M., Forsyth, P.B., Mwavita, M., Barnes, L.L., & Khojasteh, J. (2016). An empirical test of Oklahoma's A-F school grades. *Education Policy Analysis Archives, 24*(4). Retrieved January 3, 2022, from http://dx.doi.org/10.14507/epaa.v24.2127

Adams, C.M., Forsyth, P.B., Ware, J., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*(7), 1-31. Retrieved January 18, 2022, from https://www.tcrecord.org/content.asp?contentid=20925

53 Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., et al. (1966). *Equality of educational opportunity* (No. OE-38001). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Education Statistics.

Benson, J. & Borman, G. (2010). Family, neighborhood, and school settings across seasons: When do socioeconomic context and racial composition matter for reading achievement growth of young children? *Teachers College Record, 112*(5), 1338-1390.

Borman, G. & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's Equality of Educational Opportunity data. *Teachers College Record, 112*(5), 1201-1246.