ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-863

State Accountability Decisions under the Every Student Succeeds Act and the Validity, Stability, and Equity of School Ratings

Erica Harbatkin
Florida State University

Betsy Wolf
U.S. Department of Education

The Every Student Succeeds Act (ESSA) began a new wave of school accountability under which states draw on multiple measures to assess school quality. States have options in terms of how to weight components in their school quality indices and how many years of data to use to determine school ratings. In this study, we simulate school ratings using eight years of administrative data in North Carolina to demonstrate how state decisions about school ratings and identification influence school ratings and the list of schools identified for improvement. We then evaluate these decisions against a framework that considers the validity, stability, and equity of the ratings, underscoring the inherent tradeoffs that come with each. We show that while a system that weights proficiency more heavily than growth produces more stable school ratings, identifying schools based on multiple years of performance data instead of one more than offsets the loss of stability in shifting to a growth measure. We conclude with recommendations for state accountability systems under ESSA and for federal policymaking moving forward.

VERSION: October 2023

Suggested citation: Harbatkin, Erica, and Betsy Wolf. (2023). State Accountability Decisions under the Every Student Succeeds Act and the Validity, Stability, and Equity of School Ratings. (EdWorkingPaper: 23-863). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xt8e-0w18

State Accountability Decisions under the Every Student Succeeds Act and the Validity, Stability, and Equity of School Ratings

Erica Harbatkin[1] & Betsy Wolf[2]

[1] Florida State University [2] U.S. Department of Education

[2] This article was written by the author in her private capacity. No official support or endorsement by the U.S. Department of Education is intended or should be inferred.

Abstract: The Every Student Succeeds Act (ESSA) began a new wave of school accountability under which states draw on multiple measures to assess school quality. States have options in terms of how to weight components in their school quality indices and how many years of data to use to determine school ratings. In this study, we simulate school ratings using eight years of administrative data in North Carolina to demonstrate how state decisions about school ratings and identification influence school ratings and the list of schools identified for improvement. We then evaluate these decisions against a framework that considers the validity, stability, and equity of the ratings, underscoring the inherent tradeoffs that come with each. We show that while a system that weights proficiency more heavily than growth produces more stable school ratings, identifying schools based on multiple years of performance data instead of one more than offsets the loss of stability in shifting to a growth measure. We conclude with recommendations for state accountability systems under ESSA and for federal policymaking moving forward.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E150017 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those AERA or NSF.

Introduction

Historically, school ratings have served two purposes: to identify low-performing schools for state sanctions and targeted supports, and to communicate information about school quality to the general public (Figlio & Loeb, 2011). Since the passage of the No Child Left Behind Act, states have relied on proficiency rates for each district and school to identify those in need of improvement. However, educational researchers have demonstrated that proficiency rates are an inadequate metric of school quality for two reasons. First, they rely on arbitrary thresholds that fail to capture distributional shifts in achievement and can misrepresent longer term achievement trends (Ho, 2008). Second, they measure educational opportunity since birth, which does not isolate school contributions to learning and is strongly correlated with student race, ethnicity, and socioeconomic status (Heck, 2006; Kim & Sunderman, 2005; Reardon, 2019).

The result has been longstanding inequity in school ratings, where schools serving underserved student populations have been disproportionately labeled as "low-performing" regardless of how much students learn while attending these schools (Diamond & Spillane, 2004; Harris, 2007). The low-performing label has been associated with increased school closures in underserved communities, demoralization among school staff, higher teacher turnover, and teacher recruitment challenges (Byrd-Blake et al., 2010; Clotfelter et al., 2004; Mintrop, 2003; Murillo & Flores, 2002) There is also evidence that school ratings contribute to increased school segregation and resource inequities (Davis et al., 2015; Hasan & Kumar, 2019; Houston & Henig, 2023).

Conversely, labeling schools as low-performing has the potential to induce student achievement gains (Figlio & Rouse, 2006; Saw et al., 2017; Winters & Cowen, 2012), and the turnaround interventions implemented in struggling schools can be effective in improving

student outcomes (Redding & Nguyen, 2020; Schueler et al., 2020). These mixed results highlight important tradeoffs in school accountability systems—identifying schools as low performing can improve student outcomes, but rating schools on factors outside of their control can lead to additional challenges for schools that are already struggling and exacerbate inequities.

Under the Every Student Succeeds Act (ESSA), states had the opportunity to rethink how to calculate school ratings and define low-performing schools because they were required to construct a multidimensional school quality index rather than basing school ratings on proficiency rates alone. The law stipulates which components must be included in the school quality index, but allows flexibility in defining and weighting those components. The index must include proficiency, English learner progress; graduation rate for high schools; academic progress (often growth) for elementary and middle schools; and another indicator of school quality or student success (often chronic absenteeism). ESSA requires states to use their index to classify the bottom 5% of Title I schools as Comprehensive Support and Improvement (CSI) schools at least once every three years. CSI schools then must receive a needs assessment and evidence-based turnaround supports.

States developed their first ESSA indices and identified the first required cohort of CSI schools based on 2017-18 data. Absent COVID-19, states would have identified their second cohorts using 2020-21 data, but standardized tests were paused due to the pandemic. While testing administration resumed in spring 2021, the U.S. Department of Education also granted waivers from minimum participation targets and using the tests for accountability purposes. Given the disruption of the pandemic, there has been little research to date about how school ratings—and lists of CSI schools—have changed under ESSA.

In this article, we simulate school ratings by varying assumptions under two different "toggles." The first toggle involves the respective weights of the proficiency and growth components in the school quality index. The second toggle is number of years of data used to calculate the ratings and identify CSI schools. Then, we ask:

(1) How does varying the weights in the school quality index and the number of years of data affect school ratings?

(2) How does varying the weights in the school quality index and the number of years of data affect the list of CSI schools?

To answer these questions, we propose a framework to assess the design of school accountability systems that examines the validity, stability, and equity of the school ratings under each simulation.

The remainder of this paper proceeds as follows. First, we discuss the framework to assess school accountability design. Second, we describe the data and methods used to simulate the school ratings. Third, we present the results, showing how different weighting schemes and number of years of data affect school ratings and the CSI list, and discuss the trade-offs in terms of validity, stability, and equity. We conclude with a discussion of policy implications for state and federal policymakers designing school accountability systems under ESSA and beyond.

Framework for Assessing the Design of School Accountability Systems

Validity

The face validity of school ratings is critical because ratings are often used to make high-stakes decisions, and lack of face validity undermines public trust. While there are several components in school quality indices, proficiency rates and student growth are explicitly mentioned in ESSA and are typically weighted the most in state systems. Researchers generally

agree that student growth is a more valid measure of school quality than proficiency rates (Ho, 2008; Kurtz, 2018). Proficiency rates penalize schools for factors outside their control such as student poverty (Raudenbush, 2004; Reardon, 2019), are overly sensitive to students near an arbitrary proficiency threshold (Ho, 2008; Ladd & Lauen, 2010), and are more susceptible to gaming behavior (Booher-Jennings, 2005; Darling-Hammond, 2006; Reback, 2008). Therefore, if the goal of school ratings is to capture the contributions of the school to student learning, school quality indices that weight growth more heavily than proficiency are more valid.

On the other hand, if the goal is to identify schools with the lowest achievement levels in order to target scarce resources to the most underserved schools, then school quality indices that weight proficiency more heavily than growth might be more valid—though in this case "low performing" would be a misleading term. Ammar and colleagues (2000) have characterized this distinction as the difference between low *performance* (schools with low achievement) and low *performing* (schools that are not effective at improving student achievement). Thus, the extent to which school ratings are valid is dependent on the main purpose of the school accountability system. The validity of school ratings should therefore be assessed by the extent to which they support the main purpose of the accountability system. To support the purpose, differences in ratings must reflect real differences in the construct of interest, as opposed to noise or idiosyncratic measurement error. This is the rating's "construct validity" (Polikoff et al., 2014). Therefore, the validity of school ratings also depends on the validity of the components used in the school quality index.

Stability

Our framework defines stability as the consistency of year-to-year school ratings. While school quality varies over time due to educator turnover, curricular changes, and other system

changes, large fluctuations in a short timeframe imply that ratings are driven by nonpersistent factors such as noise—which also presents validity concerns—and one-time shocks that may not be relevant in the future (Kane & Staiger, 2002; Linn & Haug, 2002). Moreover, stability matters for determining the CSI list. A CSI list that is stable from year to year would identify roughly the same schools regardless of which year CSI identification occurred—which is important because ESSA requires identification once every three years. On the other hand, an accountability system that classifies schools as low performing due to a one-year fluctuation in student achievement would sanction schools based on random noise and lead states to invest scarce resources in schools that potentially do not have greatest level of need. Thus, an accountability system that produces a stable CSI list should be a priority for states.

Accountability systems that rate schools based on multiple years of data will produce more stable school ratings than those that use only a single year of data. Weighting proficiency rates more heavily than student growth will also increase stability, but only because proficiency rates are so heavily impacted by student demographics, which are persistent from year to year (Boyd et al., 2008; Kane & Staiger, 2002; McEachin & Polikoff, 2012). Herein lies some of the tensions between validity and stability in the framework.

Equity

Our framework includes a focus on equity, which we define as the extent to which school ratings are systematically and disproportionately lower for schools serving underserved student populations (Heck, 2006). It is likely unfair and inaccurate to systematically label schools serving underserved students as low-performing, and this labeling itself can contribute to educational inequality. Accountability systems that disproportionately identify, label, and sanction schools serving underserved student populations could further marginalize these

populations by closing neighborhood schools, isolating communities due to segregation, and demoralizing staff and students (Davis et al., 2015; Lipman, 2017; Schueler & West, 2022; Trujillo, 2013). In addition, schools designated as "low performing" may engage in more triage behaviors in response to accountability threats while higher performing schools make efforts toward broader improvements (Diamond & Spillane, 2004; Jennings & Sohn, 2014). For example, students in low-performing schools may experience reduced educational and enrichment opportunities and a rigid focus on test score outcomes in mathematics and literacy.

Prior research has found that accountability systems based solely on proficiency rates are highly inequitable and disproportionately identify low-performing schools as those serving predominantly low-income, Black, or Hispanic students (Balfanz et al., 2007; Kim & Sunderman, 2005). Thus, a school quality index that places more weight on proficiency would yield less equitable ratings than an index that places more weight on growth. However, there is little evidence about whether drawing on multiple years of data to rate schools would improve or worsen the equity of school ratings, though it may be the case that if the consistently lowest performing or lowest achieving schools serve the most disadvantaged students, identifying on multiple years of data may in fact generate an even less equitable CSI list than identifying on a single year.

Data and Methods

Sample and Data

We draw on eight years of North Carolina administrative data maintained by the Education Policy Initiative at Carolina (EPIC) at the University of North Carolina at Chapel Hill. We restrict the analytic dataset to the 1,898 schools that were open for all eight years of the study

period. Using these data, we develop ESSA-compliant school quality indices and then simulate the school ratings and the bottom 5% of schools (CSI schools) given each index.

Table 1 provides school-level descriptive statistics for the analytic sample, first overall and then by whether schools are in the bottom 5% of proficiency rates, which was how schools were identified in North Carolina prior to ESSA. As expected, the bottom 5% of schools serve a disproportionate share of economically disadvantaged families and Black students. Moreover, city and rural schools are disproportionately represented in the bottom 5% because there are a greater share of high-poverty schools in these locales, which is consistent with prior literature (Clotfelter & Ladd, 1996; Harris, 2007). The overidentification of rural schools is especially relevant to states like North Carolina, where more than 40% of schools are rural and underresourced (Oakes et al., 2021; U.S. Department of Education, 2017). Finally, the bottom 5% have smaller enrollments, on average, than other schools statewide, which is consistent with prior literature showing that small schools tend to be overrepresented at the bottom of the distribution due to measurement issues (Kane & Staiger, 2002). ESSA gives states the opportunity to address the shortcomings of ratings based purely on proficiency rates.

TABLE 1

Constructing ESSA-Compliant School Quality Indices

We construct three different ESSA-compliant school quality indices by varying the first toggle in our study, the weights of the proficiency rates and student growth components included in the index. Each index uses the same set of components: proficiency rates, student achievement growth (Education Value-Added Assessment System, or EVAAS, in this case), graduation rate for high schools, English learner proficiency, and chronic absenteeism.[1] In creating our three different

[1] We choose chronic absenteeism because it is the most prevalent school quality and student success indicator used in state ESSA plans. Proficiency rates, growth, and graduation rates (cohort four-year graduation rates) are available from the state as

indices, we toggle only the weights on proficiency and growth while holding the weights on the other components constant.

The first index follows the most common weighting scheme ("modal index") across states in the U.S. based on our coding of ESSA plans for all 50 states and the District of Columbia. In this scheme, proficiency and growth are weighted similarly at the elementary and middle school level (35% proficiency and 40% growth), and proficiency is weighted twice as much as growth for high schools (30% and 15%, respectively). The second weighting scheme ("higher proficiency") weights proficiency more heavily than growth, with proficiency accounting for 60% of school ratings for elementary and middle schools and 45% for high schools. The third weighting scheme ("higher growth") weights growth more heavily than proficiency, with growth accounting for 60% of school ratings for elementary and middle schools and 30% for high schools. Appendix Table A-1 shows the three weighting schemes for all components.

Simulating School Ratings and the CSI List

We then vary the second toggle in our study, years of data used to calculate the school ratings and determine the CSI list. Most states use only one year of data to calculate school ratings and identify CSI schools—after calculating an index score for each school in the CSI identification year, they select the bottom 5% of schools in that year (McNeill et al., 2021). But ESSA does not require the CSI list to be generated based on a single year of data, and there is evidence that drawing on multiple years of data can lead to a more stable CSI list with better

school-level measures. For the other metrics of English learner proficiency and chronic absenteeism, we had access to student-level data, and aggregated the student-level data to the school level. Chronic absenteeism is defined as students are absent for 10% or more of the days a student was enrolled in the school. English learner proficiency is defined as percent of students classified as English learners who scored at proficient or above on the end-of-grade reading (grades 3-8) or English 2 end-of-course (grade 10) exam.

validity (Kane & Staiger, 2002; McEachin & Polikoff, 2012). We therefore toggle the number of years of data used to identify the CSI list using four different decision rules:

- **Single-year (status quo):** Use a single year of data to calculate school ratings, the most common approach cited in state ESSA plans (McNeill et al, 2020).
- **3-year weighted mean:** Calculate school ratings based on the three-year weighted mean index score, where the current year is weighted most heavily (times 3), one year prior is weighted next heavily (times 2), and two years prior is weighted the least heavily (times 1).
- **3-of-3-year:** Calculate school ratings using one year of data for three consecutive years, but only classify schools that fall in the lowest performing group for *three consecutive years*.[2]
- **2-of-3-year:** Calculate school ratings using one year of data for three consecutive years, but only classify schools that fall in the lowest performing group for at least *two of three consecutive years*.

We also simulate school ratings more generally using the single-year rule and the 3-year weighted mean rule.

Applying the Framework to Assess Accountability System Design

We next apply our framework to assess the stability, equity, and validity of the simulated school accountability systems. The following sections provide additional detail about how we operationalize stability and equity in our analyses. We do not describe an operationalization of validity because we apply our framework drawing on existing literature on the validity of proficiency- versus growth-based measures.

[2] Because the bottom 5% changes from year to year, classifying a minimum of 5% of schools as mandated by ESSA requires identifying more than the bottom 5% in any given year. Here and in the case of the 2-of-3-year rule, we set thresholds based on the number needed to reach 5% after three years. We provide these thresholds in Appendix Table A-2.

Stability

We estimate the stability of school ratings by calculating simple correlation coefficients between the ratings in consecutive school years within quintiles of performance to understand whether there is differential volatility at different points in the distribution. Then, to measure the stability of the CSI list, we calculate the share of schools identified in a given year that would have also been identified in the prior year and in the subsequent year, respectively, under each simulation. Schools that spend consecutive years in the bottom 5% are more likely to have persistent needs as opposed to being identified as a CSI school due to random noise or one-year shocks. In addition, ESSA requires identification of CSI schools once every three years, so ideally, a similar set of schools would be identified regardless of when the identification year happened to occur within a three-year period.

Equity

To investigate the equity of school ratings, we examine student compositions (i.e., percent economically disadvantaged and Black, respectively) in schools by school rating quintile. Then, to estimate equity of the CSI list, we break the list of schools into three quantiles based on student composition (i.e., 25% of schools with largest share of economically disadvantaged or Black students, middle 50%, and 25% of schools with smallest share of economically disadvantaged or Black students, respectively). We then calculate the share of schools within each of the three student composition quantiles that would be identified as CSI under each simulation. A perfectly equitable accountability system would identify exactly 5% of CSI schools in each of the quantiles. Thus, the extent to which the share of schools identified in a given quantile deviates from 5% represents the extent to which the CSI list is inequitable.

Findings

We discuss the results from our simulations according to our framework of stability, equity, and validity. Within each of the three categories, we first answer RQ1, discussing the results for school ratings in general. We then answer RQ2 by discussing the results for the CSI list in particular.

Stability

Figure 1 shows scatterplots of school ratings in consecutive years for each weighting scheme, with the modal school quality index in Panel A, the higher proficiency index in Panel B, and the higher growth index in Panel C. The first column provides the school rating percentile based on a single year and the second based on a three-year weighted mean. Looking within the first column, the tighter clustering of schools around the linear fit line in the higher proficiency index (Panel B) compared with the greater dispersion in the higher growth index (Panel C) underscores that weighting proficiency more heavily produces more stable year-to-year ratings than weighting growth more heavily. This is unsurprising as proficiency rates are strongly correlated with persistent student characteristics, such as family income, whereas growth measures are noisy measures and therefore more volatile over time. There is also greater stability in the tails of the distribution. For example, when using only one year of data, previous year ratings explain 4–18% of variation in the subsequent ratings for schools in the lowest quintile of performance, 6–19% for schools in the highest quintile, and only 1–8% for schools in the three middle quintiles.

Next, comparing columns, school ratings based on three years of data are three to four times more stable than those based on a single year across all three weighting schemes. When using three years of data, previous year ratings explain 20–52% of variation in the subsequent

ratings for schools in the lowest quintile of performance, 26–55% for schools in the highest quintile, and 7–32% for schools in the three middle quintiles. Therefore, all simulated ratings reflect idiosyncratic year-to-year variation more than persistent signals of quality, but include more signal when they draw on three years of data rather than one.

FIGURE 1

We next assess the stability of the CSI list by examining the bottom 5% of schools under each simulation. We calculate the share of CSI schools that would have also been identified in the prior year, subsequent year, and both, respectively. Table 2 shows the results for each weighting scheme (panels) and the number of years of data used to create the list (rows). As expected, using multiple years of data produces more stable CSI lists than using only one year of data. In particular, the 3-of-3-year rule produces the most stable list, followed by a slight loss in stability when using the 3-year weighted mean rule. Using the 2-of-3-year rule substantially reduces stability, and using only one of data results in the greatest loss in stability and the most volatile list of CSI schools.

As expected, the higher proficiency index (Panel B) generates the most stable list (25–58% identified in each of three years) and the higher growth index (Panel C) generates the least stable list (13–34%). However, using three years of data more than offsets the loss of stability associated with shifting from a high proficiency to a high growth index. For example, while only 13% of schools would be identified for three consecutive years using the single-year status quo rule with the higher growth weighting scheme (Panel C, row 1), 39% would be identified using the 3-of-3-year rule (Panel C, row 2)—this is more than a 50% improvement over the stability of the list using only one year of data and the higher proficiency index (Panel B, row 1).

TABLE 2

Equity

Figure 2 shows the percentage of students who were economically disadvantaged or Black by school rating quintile and school quality index. The figure highlights that all three indices generate inequitable ratings, with the share of economically disadvantaged (Panel A) and Black (Panel B) students decreasing monotonically as school ratings increase. The higher proficiency index is the most inequitable among the three indices, and the higher growth index is the least inequitable. For example, under the higher proficiency index, 75% of students in the lowest rated 20% of schools were economically disadvantaged, compared with 39% in the top 20% of schools. Under the higher growth index, 65% of students in the lowest rated 20% of schools were economically disadvantaged compared with 47% in the top 20% of schools. The findings follow the same pattern for the percentages of Black students.

FIGURE 2

Figure 3 shows the equity implications for the CSI list under each simulation. The figure disaggregates schools into quantiles (lowest 25%, middle 50%, highest 25%) of economic disadvantage and Black student shares, respectively, and shows the share of schools in each quantile that would be identified as CSI. A perfectly equitable system would identify exactly 5% of schools in each quantile. The large disparities in bar heights across the three quantiles highlight that none of the simulations are equitable on this measure. The schools with the greatest share of economically disadvantaged (Panel A) and Black (Panel B) students are disproportionately identified for CSI while the schools with the smallest shares of these populations are very rarely identified. As expected, the disparities are starkest using the higher proficiency index and somewhat less pronounced using the higher growth index. The higher proficiency index would identify 15–16% of schools with the greatest economic disadvantage, 3% of schools in the middle half of economic disadvantage, and only 0.01% of schools with the

least economic disadvantage. The higher growth index would identify about 10–11% with the greatest economic disadvantage schools, about 5% of the middle group, and 0.06–0.07% of schools with the lowest economic disadvantage. Panel B shows that Black share follows a similar pattern.

The close similarities across panels also show that the number of years of data used is not consequential; using multiple years versus only one year of data to identify CSI schools does not meaningfully improve or decrease equity.

FIGURE 3

Validity

While we do not present a direct test of validity, we discuss implications of the simulations for validity. The instability of school ratings raises concerns for using ratings as a public signal of school quality. Increasing the number of years of data used to calculate the ratings and focusing on schools in the highest or lowest quintiles of performance might increase stability—and therefore result in a more predictive signal of school quality—but a rank ordering of all schools is not sufficiently valid to support public decisionmaking.

If the main purpose of school ratings is to identify the schools in need of improvement, then validity is increased when using multiple years of data. In addition, the extent to which the lowest achieving or performing schools are reliably identified depends on whether the components themselves have construct validity. For example, validity is compromised when proficiency rates are a poor indicator of the lowest performing schools and when growth metrics do not reliably capture how much students are learning. Finally, to be valid, the labeling of the CSI list must match the reason that schools are identified. Schools that are identified using the higher proficiency index should be adequately labeled to reflect schools with the lowest achievement as opposed to those with the lowest performance.

Discussion

States had the opportunity to rethink their school accountability systems under ESSA, and there is little research on the implications of the decisions states made as part of this process. Most states on their ESSA plans weight proficiency and growth similarly for elementary and middle schools, split high school weights relatively evenly across proficiency, growth, and graduation rates, and use a single year of data to calculate school ratings. This status quo accountability system presents challenges in terms of the stability, equity, and validity of school ratings and the identification of the bottom 5% (or CSI) schools.

Our findings point to several considerations for states refining their accountability systems under ESSA. First, we suggest that states should draw on multiple years of data to identify their CSI schools. Using multiple years of data will lead to a more stable (and arguably more valid) CSI list. A large literature on value-added measurements comes to the same conclusion—that multiple years of data increases precision and reduces statistical uncertainty, and is needed for high-stakes decisions (Goldhaber & Hansen, 2012; McCaffrey et al., 2009).

Second, states should be intentional about the weighting schemes in their school quality indices, and the index should support the accountability system's goals. If the goal is to identify the 5% lowest performing schools, our recommendation is to construct a school quality index similar to our higher growth index and use three years of data to identify CSI schools. This system has advantages in terms of being among the most stable options— without much loss relative to the higher proficiency index—and is the most equitable option we studied. It is also the most valid option in terms of focusing on outcomes that schools control (i.e., growth). However, our equity analyses show that this system will still lead to inequitable school ratings that disproportionately identify CSI schools as those serving underserved student populations.

On the other hand, if the goal is to identify the schools with the lowest achievement levels, our recommendation is to use the higher proficiency index and three years of data. This option is the most stable from year to year. However, because proficiency rates are more susceptible to gaming and can fail to capture meaningful distributional shifts in achievement, states might also want to consider determining low achievement based on average test scores in addition to proficiency rates and target schools with the lowest achievement levels for the most intensive supports. However, this option is among the least equitable, and a focus on proficiency would also mean that the "low performing" label that comes with CSI would be misleading. States could avoid such labels, and instead underscore opportunity gaps and the efforts they are making to close these gaps and accelerate student learning. Still, any rating system that places schools on a nominal scale has the potential to lead to further disinvestment in these schools, undermining the opportunity to improve student outcomes.

Third, our findings suggest that school rankings themselves are generally not reliable indicators of future rankings. As ESSA does not require reporting a school's rank order on ESSA's meaningful differentiation index, states might want to reconsider how ratings are reported and how families can interpret them. For example, selecting a school that that is ranked at the 60^{th} percentile over a school that is ranked at the 50^{th} percentile is akin to making a decision based on noise. School rankings may be more valid and stable for schools within the top or bottom quintiles of performance, but their degree of instability indicates that caution is needed, and the ratings themselves may not be sufficiently valid or stable to support decisionmaking.

Our findings also have implications for national policy as lawmakers move toward Elementary and Secondary Education Act (ESEA) reauthorization. The term "low-performing school" appears throughout the language of ESSA, even as it calls for a states to measure academic

achievement using proficiency rates. Upon ESEA reauthorization, policymakers should consider the goal of the federal school accountability system much like states need to consider their goals in complying with that law. If the goal is to identify the schools with the greatest needs for supports, the law should not characterize these schools as low performing. If the goal is to name and shame, ESEA should call for accountability systems that best capture school contributions to student learning.

Under any accountability system, decision makers will inevitably face tradeoffs in terms of stability, equity, and validity. States must decide how to balance all three. This paper provides a starting point to understanding the implications of these accountability decisions, and to spur further conversation about how to improve the status quo.

References

Ammar, S., Bifulco, R., Duncombe, W., & Wright, R. (2000). Identifying low-performance public schools. *Studies in Educational Evaluation*, *26*(3), 259–287. https://doi.org/10.1016/S0191-491X(00)00019-5

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 High Schools? *American Educational Research Journal*, *44*(3), 559–593. https://doi.org/10.3102/0002831207306768

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). Measuring effect sizes: The effect of measurement error. *National Conference on Value-Added Modeling*.

Byrd-Blake, M., Afolayan, M. O., Hunt, J. W., Fabunmi, M., Pryor, B. W., & Leander, R. (2010). Morale of Teachers in High Poverty Schools: A Post-NCLB Mixed Methods Analysis. *Education and Urban Society*, *42*(4), 450–472. https://doi.org/10.1177/0013124510362340

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. *Holding Schools Accountable: Performance-Based Reform in Education*, 23–64.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management*, *23*(2), 251–271. https://doi.org/10.1002/pam.20003

Darling-Hammond, L. (2006). No Child Left Behind and high school reform. *Harvard Educational Review*, *76*(4), 642–667.

Davis, T., Bhatt, R., & Schwarz, K. (2015). School segregation in the era of accountability. *Social Currents*, *2*(3), 239–259.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, *106*(6), 1145–1176.

Figlio, D., & Loeb, S. (2011). Chapter 8—School accountability. In *Handbook of the Economics of Education* (Vol. 3, pp. 383–417).

Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, *90*(1), 239–255. https://doi.org/10.1016/j.jpubeco.2005.08.005

Harris, D. N. (2007). High-Flying Schools, Student Disadvantage, and the Logic of NCLB. *American Journal of Education*, *113*(3), 367–394. https://doi.org/10.1086/512737

Hasan, S., & Kumar, A. (2019). Digitization and divergence: Online school ratings and segregation in America. *Available at SSRN 3265316*.

Heck, R. H. (2006). Assessing School Achievement Progress: Comparing Alternative Approaches. *Educational Administration Quarterly*, *42*(5), 667–699. https://doi.org/10.1177/0013161X06293718

Ho, A. D. (2008). The Problem With "Proficiency": Limitations of Statistics and Policy Under No Child Left Behind. *Educational Researcher*, *37*(6), 351–360. https://doi.org/10.3102/0013189X08323842

Houston, D. M., & Henig, J. R. (2023). The "Good" Schools: Academic Performance Data, School Choice, and Segregation. *AERA Open*, *9*, 23328584231177664. https://doi.org/10.1177/23328584231177666

Jennings, J., & Sohn, H. (2014). Measure for Measure: How Proficiency-based Accountability Systems Affect Inequality in Academic Achievement. *Sociology of Education*, *87*(2), 125–141. https://doi.org/10.1177/0038040714525787

Kane, T. J., & Staiger, D. O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*, *16*(4), 91–114. https://doi.org/10.1257/089533002320950993

Kim, J. S., & Sunderman, G. L. (2005). Measuring Academic Proficiency under the No Child Left behind Act: Implications for Educational Equity. *Educational Researcher*, *34*(8), 3–13.

Kurtz, M. D. (2018). Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects? *Statistics and Public Policy*, *5*(1), 1–8. https://doi.org/10.1080/2330443X.2018.1438938

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, *29*(3), 426–450. https://doi.org/10.1002/pam.20504

Linn, R. L., & Haug, C. (2002). Stability of School-Building Accountability Scores and Gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29–36.

Lipman, P. (2017). The landscape of education "reform" in Chicago: Neoliberalism meets a grassroots movement. *Education Policy Analysis Archives*, *25*, 54–54. https://doi.org/10.14507/epaa.25.2660

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. https://doi.org/10.3102/0013189X12453494

McNeill, S., Henry, G. T., Covelli, L., Redden, A. R., & Crutchfield, A. N. (2021, March). *The comprehensive support and improvement of lowest performing schools under the Every Student Succeeds Act: An evaluation of state plans*. Annual Conference of the Association of Education Finance and Policy (AEFP). https://aefpweb.org/proposals/44/44513

Mintrop, H. (2003). The Limits of Sanctions in Low-Performing Schools. *Education Policy Analysis Archives*, *11*(0), Article 0. https://doi.org/10.14507/epaa.v11n3.2003

Murillo, E. G., & Flores, S. Y. (2002). Reform by shame: Managing the stigma of labels in high stakes testing. *The Journal of Educational Foundations*, *16*(2), 93.

Oakes, J., Cookson, P., George, J., Levin, S., & Carver-Thomas, D. (2021). Adequate and Equitable Education in High-Poverty Schools: Barriers and Opportunities in North Carolina. Research Brief. In *Learning Policy Institute*. Learning Policy Institute. https://eric.ed.gov/?id=ED614422

Polikoff, M. S., McEachin, A. J., Wrabel, S. L., & Duque, M. (2014). The Waive of the Future? School Accountability in the Waiver Era. *Educational Researcher*, *43*(1), 45–54. https://doi.org/10.3102/0013189X13517137

Raudenbush, S. W. (2004). Schooling, Statistics, and Poverty: Can We Measure School Improvement? In *Educational Testing Service*. https://eric.ed.gov/?id=ED486417

Reardon, S. F. (2019). Affluent schools are not always the best schools. *The Educational Opportunity Project at Stanford University*. https://edopportunity.org/discoveries/affluent-schools-are-not-always-best/

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5), 1394–1415. https://doi.org/10.1016/j.jpubeco.2007.05.003

Redding, C., & Nguyen, T. D. (2020). The Relationship Between School Turnaround and Student Outcomes: A Meta-Analysis. *Educational Evaluation and Policy Analysis*, 0162373720949513. https://doi.org/10.3102/0162373720949513

Saw, G., Schneider, B., Frank, K., Chen, I.-C., Keesler, V., & Martineau, J. (2017). The impact of being labeled as a persistently lowest achieving school: Regression discontinuity evidence on consequential school labeling. *American Journal of Education*, *123*(4), 585–613. https://doi.org/10.1086/692665

Schueler, B. E., Asher, C. A., Larned, K. E., Mehrotra, S., & Pollard, C. (2020). Improving low-performing schools: A meta-analysis of impact evaluation studies. *American Educational Research Journal*, *59*(5), 00028312211060855. https://doi.org/10.3102/00028312211060855

Schueler, B. E., & West, M. R. (2022). Federalism, Race, and the Politics of Turnaround: U.S. Public Opinion on Improving Low-Performing Schools and Districts. *Educational Researcher*, *51*(2), 122–133. https://doi.org/10.3102/0013189X211053317

Trujillo, T. M. (2013). The Disproportionate Erosion of Local Control: Urban School Boards, High-Stakes Accountability, and Democracy. *Educational Policy*, *27*(2), 334–359. https://doi.org/10.1177/0895904812465118

U.S. Department of Education. (2017). *National Center for Education Statistics (NCES) Common Core of Data (CCD) Public Elementary/Secondary School Universe Survey* [dataset]. https://nces.ed.gov/pubs2018/2018052/tables/table_04.asp

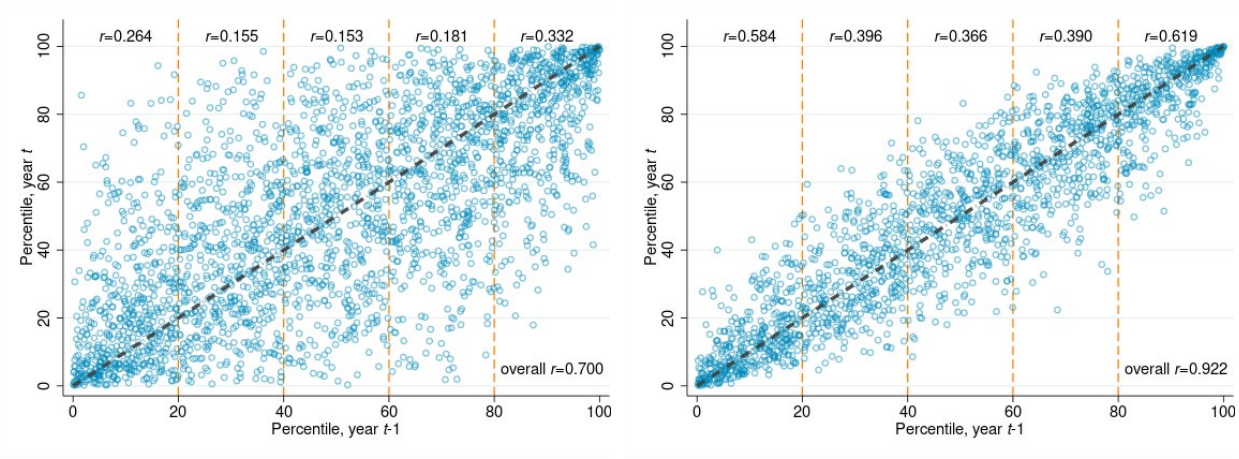
Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and Student Proficiency in America's Largest School District. *Educational Evaluation and Policy Analysis*, *34*(3), 313–327.

Figures

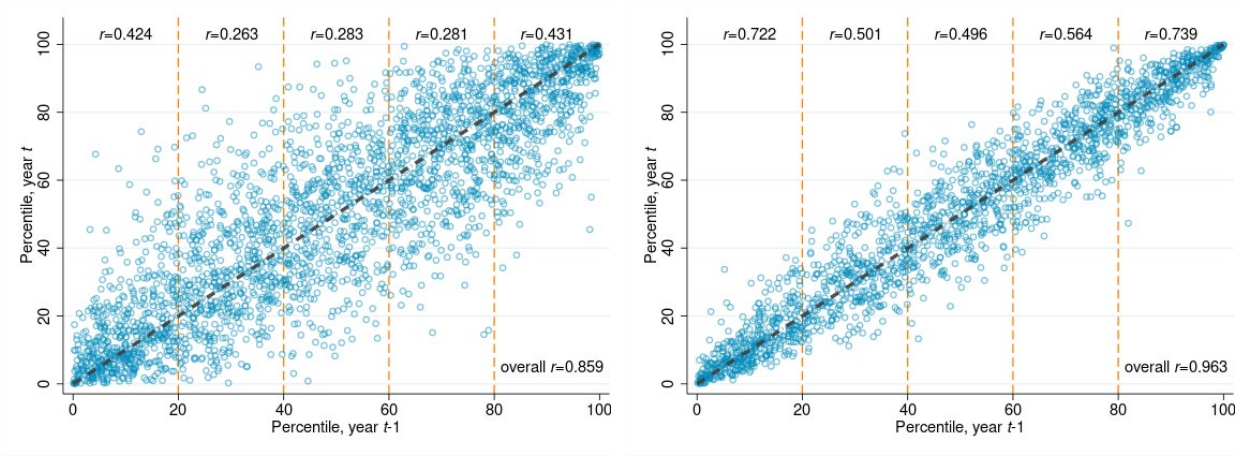
Figure 1. Correlations of school ratings in adjacent years by weighting scheme and identification rule

Panel A. One-year percentile | Panel B. Three-year weighted mean percentile

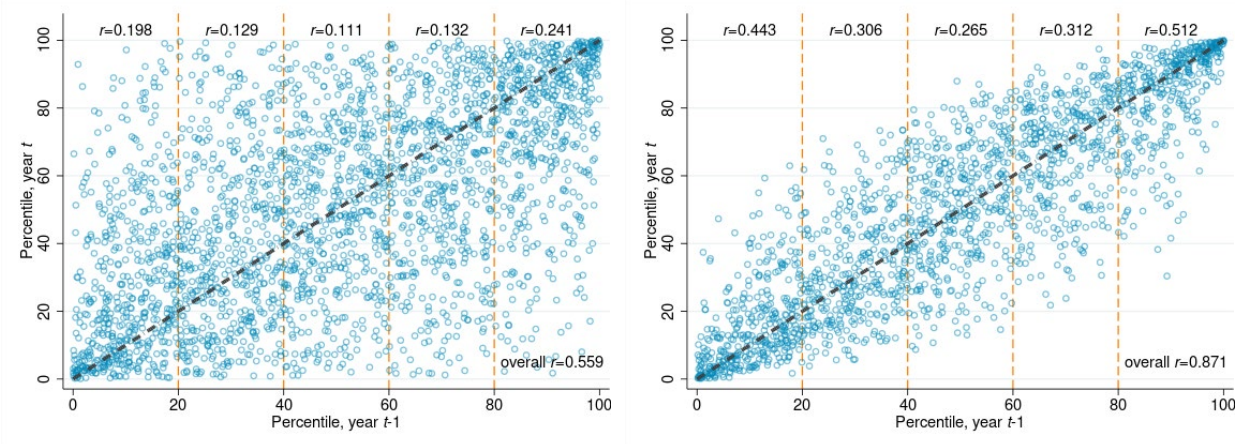
1. Modal index



2. Higher proficiency index

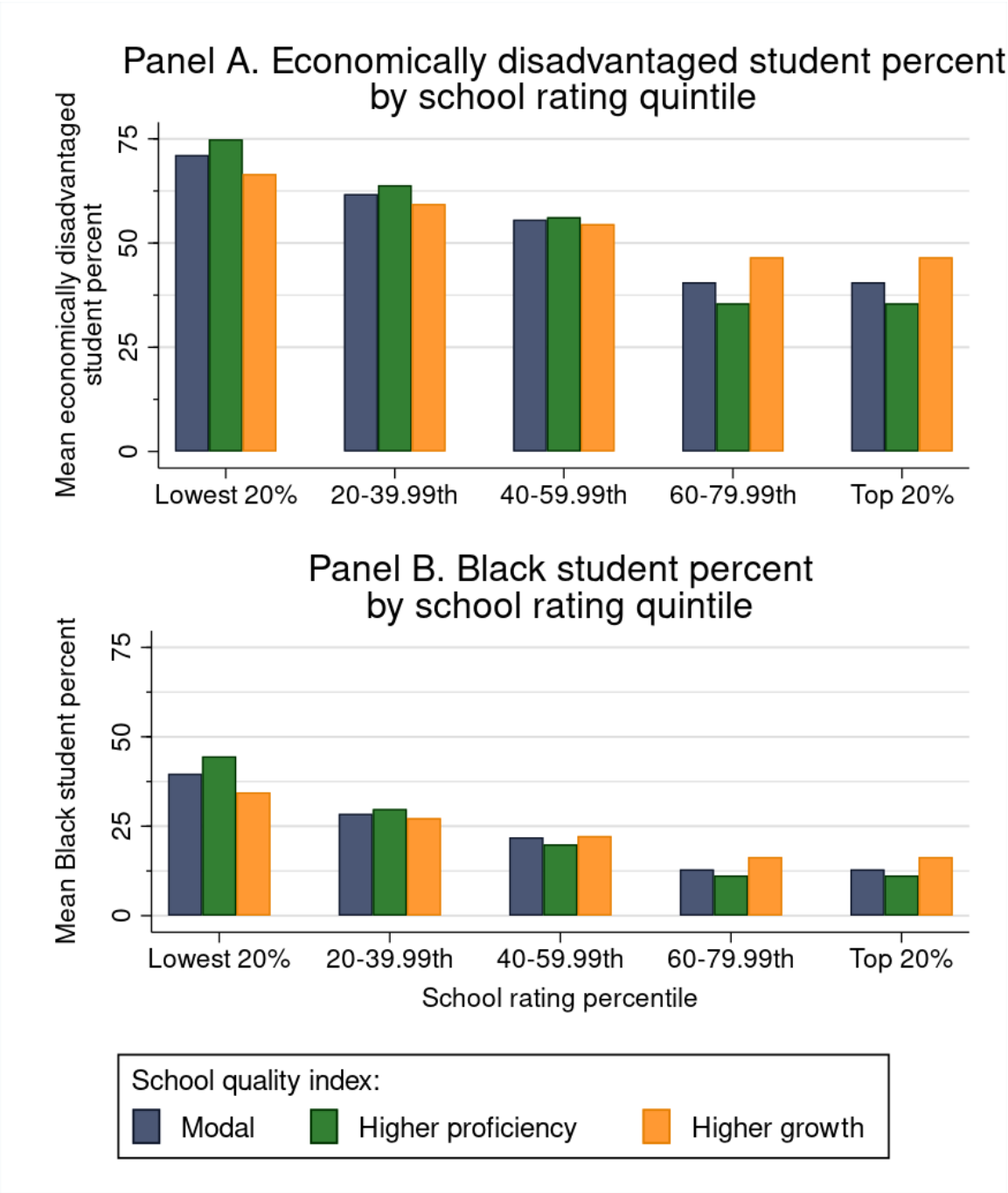


3. Higher growth index



Note: Scatterplots of school quality index percentile in adjacent years. Markers represent random sample of 20% of school-years. Correlations are calculated on full sample.

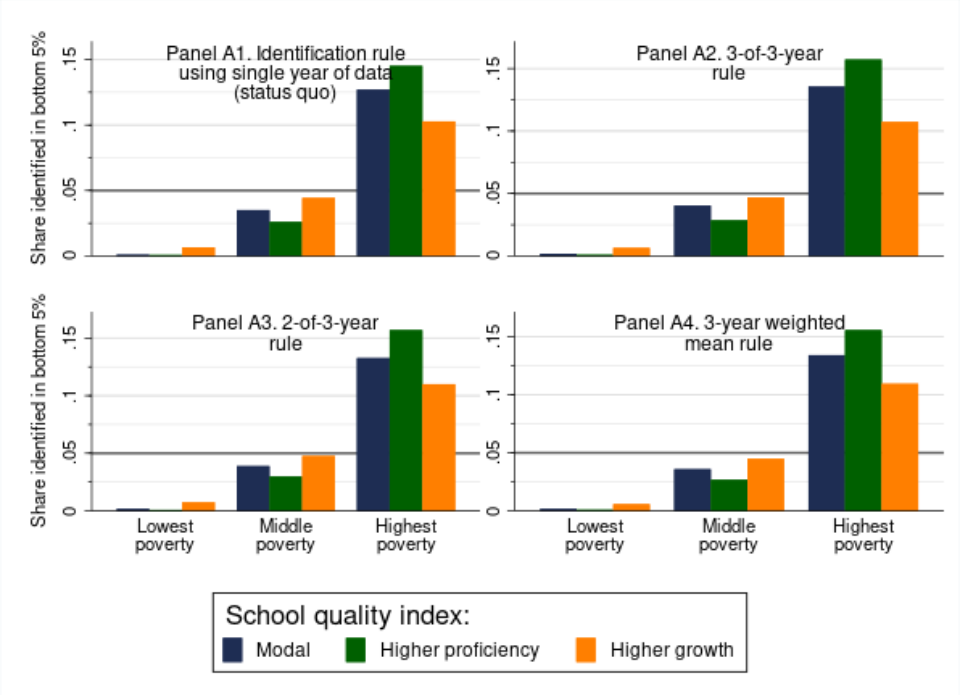
Figure 2. School demographics by rating quintile, by school quality index weighting scheme



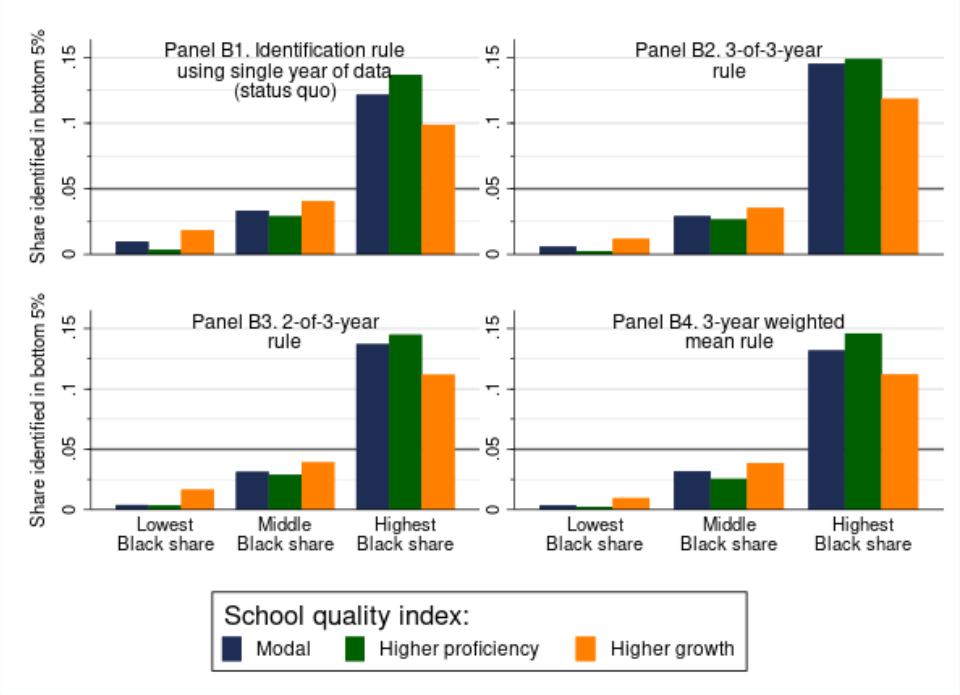
NOTE: Bar heights represent mean percent of economically disadvantaged (Panel A) and Black (Panel B) students within each quintile of school ratings based on each of three school quality indices. Quintiles based on school accountability index score, where "Lowest 20%" denotes the lowest scoring 20% of schools and "Top 20%" is the highest scoring 20% of schools.

Figure 3. Share of CSI schools by quantiles of school demographics, by school quality index weighting scheme and identification rule

Panel A. Over economically disadvantaged student share quantiles



Panel B. Over Black student share quantiles



Tables

Table 1. School-level sample statistics overall and by proficiency category

	All schools	Bottom 5% of schools	Remaining 95% of schools
Overall Achievement	**(%)**	**(%)**	**(%)**
Proficiency rate	60.4	30.7	61.8
	(17.8)	(12.7)	(16.8)
Grade level	**(%)**	**(%)**	**(%)**
Elementary	55.4	54.4	55.5
	(49.7)	(49.8)	(49.7)
Middle	25.0	26.0	25.0
	(43.3)	(43.9)	(43.3)
High	19.6	19.6	19.5
	(39.7)	(39.8)	(39.7)
Locale	**(%)**	**(%)**	**(%)**
City	26.7	33.4	26.4
	(44.2)	(47.2)	(44.1)
Suburb	9.0	2.6	9.3
	(28.7)	(15.8)	(29.1)
Town	9.0	4.4	9.2
	(28.6)	(20.5)	(28.9)
Rural	55.3	59.7	55.1
	(49.7)	(49.1)	(49.7)
Student demographics	**(%)**	**(%)**	**(%)**
Economically disadvantaged	55.6	82.2	54.4
	(21.7)	(15.7)	(21.2)
White	53.3	11.0	55.2
	(26.3)	(11.6)	(25.2)
Black	24.0	63.3	22.2
	(21.8)	(22.7)	(20.0)
Hispanic	15.1	18.3	14.9
	(12.0)	(16.9)	(11.7)
Asian	2.2	1.2	2.3
	(4.2)	(2.1)	(4.2)
American Indian	1.4	3.4	1.3
	(6.8)	(11.7)	(6.5)
Pacific Islander	0.1	0.1	0.1
	(0.3)	(0.5)	(0.3)

	All schools	Bottom 5% of schools	Remaining 95% of schools
Enrollment	**Mean**	**Mean**	**Mean**
Enrollment (100s)	6.3	4.8	6.3
	(3.6)	(2.3)	(3.6)
School *N*	15,184	662	14,522

Note: School-level means with standard deviations in parentheses. Sample includes all 1,898 public schools observed in all eight years.

Table 2. Count and percent of CSI schools in a given year that would have been identified as CSI schools adjacent years, by weighting scheme and years of data

Panel A. Modal index

	CSI in year *t*	*t*-1		*t*+1		*t*-1 and *t*+1	
Years of data	N	N	%	N	%	N	%
Single-year status quo rule	376	140	37.2	148	39.4	68	18.1
3-of-3-year rule	398	256	64.3	261	65.6	183	46.0
2-of-3-year rule	386	228	59.1	227	58.8	138	35.8
3-year weighted mean rule	376	241	64.1	238	63.3	160	42.6

Panel B. Higher proficiency index

	CSI in year *t*	*t*-1		*t*+1		*t*-1 and *t*+1	
Years of data	N	N	%	N	%	N	%
Single-year status quo rule	376	166	44.2	172	45.7	94	25.0
3-of-3-year rule	389	282	72.5	291	74.8	227	58.4
2-of-3-year rule	387	252	65.1	255	65.9	171	44.2
3-year weighted mean rule	376	264	70.2	274	72.9	192	51.1

Panel C. Higher growth index

	CSI in year *t*	*t*-1		*t*+1		*t*-1 and *t*+1	
Years of data	N	N	%	N	%	N	%
Single-year status quo rule	376	105	27.9	123	32.7	47	12.5
3-of-3-year rule	383	226	59.0	229	59.8	149	38.9
2-of-3-year rule	388	195	50.3	199	51.3	107	27.6
3-year weighted mean rule	376	216	57.5	226	60.1	129	34.3

NOTE: Restricted to four-year period from 2014–2017, which are the years for which we can observe a CSI list for both *t*-1 and *t*+1 for all identification rules. "CSI in year *t* N" provides the number of school-years that would be classified as CSI in year *t*. The "N" and "%" columns provide the number and percentage, respectively, of those school-years that would have also been classified in *t*-1 (the prior year), *t*+1 (the subsequent year), and in both *t*-1 and *t*+1 (*both* the prior and subsequent years)

Appendix

Table A-1. Weighting schemes for ESSA-compliant school quality index

	Elementary and middle schools			High schools		
	Modal	Higher proficiency	Higher growth	Modal	Higher proficiency	Higher growth
Proficiency rate	0.35	0.6	0.15	0.3	0.45	0.15
Student growth[1]	0.4	0.15	0.6	0.15	0	0.3
Graduation rate				0.2	0.2	0.2
EL proficiency[2]	0.1	0.1	0.1	0.1	0.1	0.1
Chronic absenteeism	0.15	0.15	0.15	0.25	0.25	0.25
	1.0	1.0	1.0	1.0	1.0	1.0

[1] Student growth is calculated from the Education Value-Added Assessment System, or EVAAS.
[2] We do not have access to data from the English language proficiency exam taken by English learners (ELs). As a proxy, we use the percent of students classified as ELs who scored at proficient or above on the EOG reading (grades 3-8) or English 2 EOC (grade 10). We set the minimum cell size for ELs with ELA proficiency scores to five students. For schools that do not have at least five students, we reallocate the EL weight proportionally across the other index components.

Table A-2. Percentile low-performing threshold by school year and index for 2-of-3-year and 3-of-3 year rules

	2-of-2-year rule			3-of-3-year rule		
	Modal index	Higher proficiency	Higher growth	Modal index	Higher proficiency	Higher growth
2012-13	11	9.5	13	17	12	20
2013-14	10.5	8.5	12	15	11.5	19
2014-15	11	9	12.5	15.5	11	19
2015-16	10	8.5	12	16	11.5	17.5
2016-17	9.5	8.5	11.5	13.5	11	16.5
2017-18	10	8.5	11.5	14	10.5	17

Note: Numbers in cells represent percentile thresholds on each school quality index. We calculate these thresholds as the minimum threshold needed in each year to achieve a low-performing list that contains 5% of schools as required by ESSA. They are higher than 5% because these two rules require schools to be classified as low performing in consecutive years. The threshold will therefore be higher than 5% to the extent that ratings vary from year to year. In other words, an index with no movement in and out of the bottom 5% for three consecutive years would yield a low-performing threshold of 5%. An index with more instability such as the higher growth index will require a higher threshold than an index with less instability such as the higher proficiency index.