Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 23-863

# State School Accountability Decisions and the Implications for School Ratings

Erica Harbatkin
Florida State University

Betsy Wolf
D.C. Office of the State Superintendent of Education

State school accountability systems have been a constant in educational policy since before federal law required them. Currently, state systems must include multiple measures of school quality but states otherwise have considerable leeway in measuring schools. We simulate school ratings using eight years of data to demonstrate how state decisions about (a) weighting multiple measures, and (b) how many years of data go into the rating, influence the list of schools identified for improvement. We evaluate the implications and tradeoffs inherent in these decisions. For example, weighting growth more than proficiency produces more valid measures of school quality, whereas weighting proficiency more identifies schools most in need of resources. We conclude with recommendations for state accountability systems.

VERSION: September 2026

Suggested citation: Harbatkin, Erica, and Betsy Wolf. (2026). State School Accountability Decisions and the Implications for School Ratings. (EdWorkingPaper: 23-863). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xt8e-0w18

**State School Accountability Decisions and the Implications for School Ratings**

Erica Harbatkin[1] & Betsy Wolf[2]

[1] Florida State University [2] D.C. Office of the State Superintendent of Education

Abstract: State school accountability systems have been a constant in educational policy since before federal law required them. Currently, state systems must include multiple measures of school quality but states otherwise have considerable leeway in measuring schools. We simulate school ratings using eight years of data to demonstrate how state decisions about (a) weighting multiple measures, and (b) how many years of data go into the rating, influence the list of schools identified for improvement. We evaluate the implications and tradeoffs inherent in these decisions. For example, weighting growth more than proficiency produces more valid measures of school quality, whereas weighting proficiency more identifies schools most in need of resources. We conclude with recommendations for state accountability systems.

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 in the United States 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 school proficiency rates to identify those in need of improvement. However, educational researchers have demonstrated that proficiency rates are an inadequate metric of school quality for at least two reasons. First, they rely on arbitrary thresholds that fail to capture distributional shifts in achievement and can misrepresent longer term achievement trends (Darling-Hammond, 2018; 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; Krieg & Storer, 2006; Reardon, 2019a).

The result has been longstanding inequity in school ratings, where schools serving greater percentages of economically disadvantaged, Black, and Hispanic student populations have been disproportionately labeled as "low performing" regardless of how much students learn while attending them (Burrell & Harbatkin, 2024; Diamond & Spillane, 2004; Harris, 2007; Hursh, 2007; Wright & Petrilli, 2017). The low-performing label has been associated with increased school closures in historically marginalized communities, staff demoralization, higher teacher turnover, and teacher recruitment challenges (Balfanz et al., 2007; Byrd-Blake et al., 2010; Clotfelter et al., 2004; Davis et al., 2015; Figlio & Rouse, 2006; Finnigan & Gross, 2007; Lee & Lubienski, 2017; Mintrop, 2003; Murillo & Flores, 2002; Rice & Malen, 2003). School ratings more broadly contribute to school segregation and resource inequity (Davis et al., 2015; Hasan & Kumar, 2019; Houston & Henig, 2023).

On the other hand, labeling schools as low performing has the potential to induce student achievement gains (Deming et al., 2016; Figlio & Rouse, 2006; Reback, 2008a; Saw et al., 2017; Winters & Cowen, 2012), and turnaround interventions implemented in struggling schools can be effective in improving student outcomes (Redding & Nguyen, 2020; Schueler et al., 2020). These seemingly contradictory findings highlight important tradeoffs in school accountability systems—school ratings and labels can improve student outcomes, but rating schools on factors outside of their control can lead to additional challenges for schools that are already struggling.

State policymakers had the opportunity to rethink their school accountability systems under the Every Student Succeeds Act (ESSA, 2015), which required states to construct a school quality index based on multiple metrics rather than proficiency rates alone. The metrics must include achievement (proficiency or another achievement measure),[1] 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)—though ESSA allows flexibility in defining and weighting these metrics. ESSA then requires states to use their school ratings to classify the bottom 5% of Title I schools as Comprehensive Support and Improvement (CSI) schools at least once every three years. States decide how many years of data to use in determining the CSI schools. Once CSI schools are identified, they must receive a needs assessment and evidence-based turnaround supports, though federal policy allows states broad autonomy in designing and implementing those supports. Therefore, ESSA opened the door to school accountability systems that allow for more nuance in school quality than proficiency rates alone.

That said, the nuance that comes with ESSA's added flexibility may also come at the expense of school ratings fluctuating to a greater extent from one year to the next. Accountability systems with

[1] Forty-four states use proficiency rates in full or as part of their ESSA achievement measure based on our coding of state ESSA plans, while another three use test levels (i.e., multiple achievement thresholds). Six use a scale score, either on its own or in combination with other achievement measures.

highly inconsistent year-to-year school ratings would not effectively serve their intended purpose of improving student outcomes for several reasons. First, incentives are effective only when the target is clear. If educational leaders are uncertain about where a school stands, they will struggle to adjust effort and resources toward school improvement goals. Second, inconsistent ratings would lead to unstable CSI designations. For example, whether a school is designated as “low performing” would depend in large part on the year the designations were determined. Third, the public would have trouble making meaning of highly inconsistent ratings in their decisionmaking.

The CSI requirements under ESSA will remain in place unless federal policymakers explicitly repeal the Elementary and Secondary Education Act (ESEA, 1965) or reauthorize it without school accountability provisions. Thus, even in an uncertain federal policy landscape, states are continuing to operate school accountability systems. Further, state accountability systems predate federal requirements for them, as nearly half of states had their own systems before NCLB was passed (Dee, 2012; Hanushek & Raymond, 2005). If the federal government does in fact retreat from K-12 educational policy, states will have increased autonomy to design and redesign their school accountability systems. The present is therefore a critical juncture for understanding the implications of these design choices. In sum, school accountability has been a mainstay of state educational policy for more than two decades, and is likely to continue in some form even as the federal policy landscape shifts with presidential administrations’ priorities.

In this article, we examine how states might redesign systems in line with ESSA or a similarly flexible federal policy, or for their own purposes. To do so, we simulate school ratings and CSI school lists emanating from those ratings by varying a) the respective weights of the proficiency and growth metrics in the school quality index, and b) the number of years of data used to calculate ratings and identify CSI schools. We then examine the simulation results against a conceptual framework that considers the validity, stability, and equity of school accountability systems.

To carry out our simulations, we draw on statewide administrative data from North Carolina, one of the nation's earliest adopters of a consequential school accountability system (Carnoy & Loeb, 2002). The earliest iteration focused on proficiency rates on state tests, and eventually began reporting growth measured by value added. Under Race to the Top, the state assigned schools to turnaround if they were in the bottom 5% on proficiency rates (or low graduation rates for high schools). With the adoption of ESSA, North Carolina added a multidimensional index to its accountability system. The U.S. Education Department approved the state's ESSA plan in June 2018, and the state identified its first round of CSI schools in 2018-19. Although states vary in their approach to measuring school performance under ESSA, North Carolina's ESSA index shares characteristics in common with many others; for example, it includes a growth component and college and career readiness measures (Education Commission of the States, 2021). However, we highlight that this is not intended as an evaluation of North Carolina's accountability system under ESSA. Instead, we draw on the state's data to develop our own school ratings based on the broader landscape of state approaches to school accountability ratings under ESSA, as we describe later.

Our findings show that decisions about how to weight metrics and how many years of data to use have tradeoffs. Weighting growth more than proficiency will produce a more valid but less stable indicator of school quality, whereas weighting proficiency more than growth will better identify schools most in need of resources but could unfairly sanction schools for factors outside of their control. Calculating school quality based on a three-year weighted mean rather than a single year of data will produce more valid school ratings with greater predictive validity, but stakeholders may perceive them as being unfair and not responsive to recent improvements. In what follows, we begin by describing our conceptual framework in the context of the broader literature. We turn next to a description of data and methods to simulate school ratings, followed by findings, and finally a discussion of implications for policy.

## Conceptual Framework for Analyzing School Accountability Systems

Our framework considers the validity, stability, and equity of school ratings, some of which overlap. In line with prior research, we also articulate two main purposes of school accountability systems: public and consequential accountability. School ratings broadly are a form of *public accountability* because states signal to the public that schools with higher ratings provide a higher quality education than schools with lower ratings (Hutt & Polikoff, 2020). While the stakes may be greater for schools at risk of a CSI label, all schools are subject to the public accountability that comes with school ratings because stakeholders can use the ratings for decision-making. For example, families use ratings to decide where to move or what school to send their children to (Hastings & Weinstein, 2008), and educators use ratings to decide where to work (Clotfelter et al., 2004; Feng et al., 2018). The labels that come from school ratings also have *consequential accountability* because the labels lead to state-mandated interventions or actions that aim to improve student achievement. Thus, the consequences of the labels may be positive (i.e., if they improve student outcome), negative (if they lead to negative side effects), or neutral. In what follows, we operationalize validity, stability, and equity for each purpose of school accountability. In doing so, we underscore that there is no singular agreed-upon definition of school quality, and different aims would motivate different approaches to optimizing school accountability systems.

### Validity

The validity of school ratings is critical because ratings are often used to make high-stakes decisions (e.g., sanctions, school closures, reconstitution), and lack of validity undermines public trust. Three types of validity are relevant to school accountability systems. Construct validity, whether the ratings measure what they purport to measure, is typically defined as the extent to which differences in ratings (and the CSI designation) reflect real differences in school quality, as opposed to noise or idiosyncratic measurement error (Polikoff et al., 2014). Predictive validity in this case is the extent to which the ratings and designations in one year predict ratings and designations in the next. Consequential validity deals with the outcomes, whether intended or unintended, of the accountability system itself.

***Construct Validity***

How school quality is defined differs across stakeholders. Some stakeholders—and particularly researchers—define school quality as the school's contribution to student learning. Other stakeholders might argue that school quality should be defined by student achievement, regardless of the school's contribution to student learning.

While both achievement and growth are to some extent confounded by factors outside of the school's control and include some degree of measurement error, researchers generally agree that student growth is a more valid indicator of the school's contribution to learning than proficiency rates (Ho, 2008; Kurtz, 2018). Proficiency rates and achievement levels penalize schools more heavily than growth measures for factors outside their control, such as student poverty rates (Raudenbush, 2004; Reardon, 2019a). Proficiency rates are overly sensitive to an arbitrary proficiency threshold (Ho, 2008; Holland, 2002; Ladd & Lauen, 2010) and more susceptible to gaming behavior (Ballou & Springer, 2016; Booher-Jennings, 2005; Darling-Hammond, 2006; Jennings & Sohn, 2014; Krieg, 2008; Reback, 2008).[2] On the other hand, families and policymakers may prefer achievement levels or proficiency rates if they are less concerned about the school's contribution to student learning than they are about average achievement within the school

[2] Growth measures are also subject to these criticisms and not all growth measures are equally valid. For example, many states use student growth percentiles, which have the advantage of being intuitive to the public but do not perform as well as value-added under nonrandom assignment conditions (Guarino et al., 2015). Additionally, many states dichotomize growth measures after calculating them; these dichotomized measures will also be sensitive to arbitrary "meets growth" thresholds.

(Abdulkadiroğlu et al., 2020; Figlio & Lucas, 2004; Glazerman & Dotter, 2017). Other stakeholders may value both achievement and growth—not simply one or the other (Houston & Henig, 2021, 2023). In this case, the multidimensional index will capture more of the measures that parents value, but the weights on each dimension will not necessarily match parents' priorities. In sum, the construct validity of school ratings is subjective and will vary across stakeholders.

Because there is no singular agreed-upon definition of school quality, construct validity requires clarity about what is being measured. If the purpose of school accountability is to rate schools based on their contribution to student learning and identify low-*performing* schools, ratings are more valid when they weight growth more heavily than proficiency. However, if the purpose is to rate schools based on achievement and target resources to the 5% of schools with the lowest *achievement*, then ratings that weight proficiency more heavily than growth might be more valid.

***Predictive Validity***

If a school rating is indeed a valid measure of school quality, then ratings in one year ought to be reasonably predictive of performance in future years. Unfortunately, defining predictive validity is hindered by the challenges described as part of construct validity above; without agreement on what a school rating ought to capture, there is no clear answer as to what it ought to predict performance *on*. However, a rating with predictive validity would at minimum be predictive of a school's future ratings on the same metric. This is related to stability, which we discuss in a subsequent section.

***Consequential Validity***

An accountability system with strong consequential validity would improve student achievement and outcomes (Brewer et al., 2015). The consequential validity of school ratings and the CSI designation under ESSA is critical yet under-researched, particularly in cases when few additional supports are provided to schools with low ratings or the CSI designation (Lane, 2020).

While previous federal reforms came with significant resources for designated schools, ESSA requires only that states designate a small portion of federal Title I funds for CSI schools. In practice, CSI schools do not receive the additional funding needed to make meaningful improvements and some designated schools actually end up spending *less* per pupil when they are designated as CSI due to scaled back state and local dollars (Atchison & Blair, 2025; Blair et al., 2025; Hyslop & Zhou, 2024). An accountability system that exacerbates resource gaps between lower and higher rated schools could produce negative effects on student achievement and lead to a lack of consequential validity.

An accountability system that improves student outcomes—either through ratings alone or through the turnaround interventions that come with a CSI designation—would have greater consequential validity. However, to the extent that research on the consequential validity of school accountability systems focuses primarily on effects on student achievement and does not examine effects on other characteristics of school quality that families may care about (e.g., afterschool care, sports, arts, etc.), the consequential validity of a system is unclear. For example, an accountability system may serve to improve student achievement while also redirecting resources away from arts programs (Gara et al., 2022); in this case, the consequential validity of the system would not be straightforwardly positive or negative, and different stakeholders might reach different conclusions.

**Stability**

Our framework defines stability as the consistency of year-to-year school ratings and CSI designations. Stability has dual relevance—on its own merits and because it has implications for predictive validity. Stability on its own is desirable in part for three reasons. First, the public makes decisions based on school ratings, for example, to decide where to live or send children to school. Families using ratings for these purposes would struggle to make sense of rapidly changing ratings. Second, school improvement and planning require transparent and consistent data around school performance, and volatile ratings from year to year would undermine planning processes. We note, however, that no system will (or should) produce perfectly stable ratings. School performance will meaningfully fluctuate from year to year due to educator turnover, curricular changes, and other system changes, and school ratings should pick up this variation. On the other hand, if these meaningful year-to-year shifts in school performance are overshadowed by random noise, then the school ratings lack predictive validity (Kane & Staiger, 2002; Linn & Haug, 2002). Third, the CSI designations that come from the ratings often come with additional resources for schools, and because states aim to allocate scarce resources for school improvement efficiently, consistent ratings from year to year would better allow states to direct resources to schools that stand to benefit the most from them.

A stable list of CSI schools is important because it would identify roughly the same schools regardless of which year CSI identification occurred, given that ESSA requires identification once every three years. By contrast, an accountability system that is not stable would classify CSI schools based on one-year fluctuations in student achievement, leading states to invest scarce resources in schools that potentially would not have the greatest level of need in a subsequent year. Thus, a relatively stable CSI list should be a priority for states, though no system will (or should) produce perfectly stable ratings.

We know from prior research that accountability systems that rate schools based on multiple years of data produce more stable school ratings than those using only a single year of data (Kane & Staiger, 2002). Weighting proficiency rates more heavily than student growth also increases stability. This is because proficiency rates are so heavily interrelated with student background characteristics (including poverty), which are generally persistent from year to year (Boyd et al., 2008; Kane & Staiger, 2002; McEachin & Polikoff, 2012). Herein lie some of the tensions between validity and stability in the framework.

**Equity**

We consider two dimensions of equity. The first is disproportionality. Here, following Heck (2006), we are interested in the extent to which school ratings are systematically and disproportionately lower for schools serving greater percentages of economically disadvantaged, Black, and Hispanic student populations. Broadly, disproportionality may stem from two different causes—a failure to accurately measure school quality (i.e., a deficiency with the rating itself), or a broader societal failure to equip schools and neighborhoods that are home to the nation's most vulnerable students with the resources they need. To the extent that disproportionality comes from the rating system holding schools responsible for factors outside of their control, it would lack fairness (McEachin & Polikoff, 2012; Mintrop, 2004; Polikoff et al., 2014). Accountability systems that disproportionately identify, label, and sanction such schools could further marginalize 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).

Thus, ratings that are disproportionately lower for schools serving marginalized populations could imply an unfair rating system. However, disproportionality in ratings could also be an accurate reflection of the inequities flowing in from outside the school. For example, schools serving more disadvantaged student populations grapple with high teacher turnover, particularly among highly effective teachers (Boyd et al., 2005; Simon & Johnson, 2015), disadvantaged students are consistently exposed to lower quality teachers than more advantaged students (Goldhaber et al., 2018), and there is evidence that at least some of the disproportionality in ratings can be explained by differences in teacher qualifications (Burrell & Harbatkin, 2024). Further, accountability systems may exacerbate teacher retention challenges in schools that are labeled as low performing (Clotfelter et al., 2004). There is also strong evidence that schools serving more disadvantaged students lack the necessary resources to make meaningful improvements (Atchison & Blair, 2025); in other words, they do not have the resources they need to deliver the education they are being held accountable for. This tension reiterates the importance of consequential validity; it's not just the ratings that matter but also what happens as a result of those ratings.

We therefore also examine equity of access, which considers the extent to which the accountability system could plausibly narrow resource gaps between economically disadvantaged and advantaged students or between Black and Latino students and White students. By this dimension, an accountability system would be more equitable if it assigned lower ratings to the schools with the greatest level of need *and* provided additional resources to those schools as a result of those ratings. Here, "need" could be determined based on achievement levels, greater economic needs, or limited available resources.

Equity of access overlaps with consequential validity. An accountability system that identifies CSI schools with the greatest needs and provides them with additional resources and interventions has a greater chance of achieving consequential validity. Yet even when schools receive additional resources, the resources may not result in improved student outcomes and in some cases can even have negative effects, resulting in a lack of consequential validity (Atchison, 2020; Atchison et al., 2025; Henry & Harbatkin, 2020). In practice, school accountability systems may narrow opportunity gaps to some extent, but CSI schools generally do not receive the level of funding they would need to meaningfully improve student learning and experiences (Atchison & Blair, 2025; Blair et al., 2025). This underscores that simply identifying schools with the highest need will not necessarily increase equity of access. Equity of access requires sufficient resources and supports to be paired with any low ratings and designations.

It is important to note that disproportionality and equity of access can be in conflict with one another. A rating system that produces demographically disproportionate ratings could be equitable on the access dimension if additional resources were allocated to schools with lower ratings. Disproportionality may also accurately reflect broader inequities rather than an unfair rating system. In addition, there are multiple ways of conceptualizing equity, and different states will have different priorities depending on factors such as student demographics, geographic locations, and state politics. The two dimensions of equity are not comprehensive. They are intended as a foundation from which state policymakers can consider the equity outcomes that are most relevant to their state.

## 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 exclude nontraditional schools (e.g., magnet, special education, alternative, vocational, hospital), and restrict the analytic dataset to the 1,898 traditional public schools that were open for all eight years of the study period from 2010-11 through 2017-18. Our analytic sample therefore reflects the full population of traditional public schools in North Carolina that were open for the full study period. Using these data, we develop ESSA-compliant school quality indices and then simulate (1) school ratings and (2) CSI schools (bottom 5%) based on the ratings.

There are two relevant points regarding our study window. First, the study time period includes the December 2015 passage of ESSA, which shifted states' approach to measuring school quality. States under ESSA were required to adopt multidimensional metrics of school quality in time for designations no later than the 2018-19 school year, though some adopted earlier. North Carolina identified its first cohort of CSI schools beginning in 2018-19 using data from 2017-18 (i.e., the last year in this panel). Before ESSA's implementation, the state identified its turnaround schools using proficiency and graduation rates rather than a multidimensional ESSA index. Our results are intended to speak broadly to states considering their accountability system designs, and not specifically to North Carolina. Second, the study time period precedes the COVID-19 pandemic, which complicated progress toward improvement goals nationally.

Table 1 provides school-level means for the analytic sample, overall, then by proficiency rate, and finally by the ESSA rating produced by the most common weighting scheme ("modal index") across states in the U.S. This multidimensional "modal index" is based on our coding of ESSA plans for all 50 states and the District of Columbia and described in more detail in the following section. The proficiency rate groupings broadly reflect a No Child Left Behind (NCLB) accountability system, where ratings and designations were based on achievement levels such as proficiency rates. In the columns, we show means for the bottom 5%, middle 5% (i.e., percentile 47.5–52.5), and top 5% of schools based on proficiency rates and ESSA index scores, respectively, in addition to the overall means. Panel A shows that the bottom 5% of schools by the ESSA index have slightly higher proficiency rates than the bottom 5% of schools determined by only proficiency rates, which is expected because the ESSA index includes multiple metrics.

Moving to Panel B, urban schools are overrepresented and suburban school underrepresented in the bottom 5% on both indices, though to a lesser extent on the ESSA-aligned index. Although much of the discourse around school accountability centers on urban schools, rural schools can struggle with inadequate local educator labor markets, geographic isolation, and smaller budgets and economies of scale—all of which hinder student achievement (Harbatkin, 2022; Rauscher, 2020). Relative to public schools in the United States, greater percentages of those in North Carolina are rural (55% versus 30% of all U.S. schools) and serve diverse student populations (U.S. Department of Education, 2024). North Carolina's diversity of school settings makes it a useful state context to study school accountability systems under ESSA, and our findings may be reasonably generalized to states with diversity in locales and student populations. Findings are less generalizable for states with predominantly White students.

TABLE 1

Panel C shows that the bottom 5% of schools on the ESSA-aligned index have slightly lower percentages of economically disadvantaged and Black students, respectively, and higher percentages of White students than the bottom 5% based on proficiency ratings. However, using either proficiency rates alone or the ESSA index, the bottom 5% of schools serve disproportionately more economically disadvantaged, Black, and Hispanic students and fewer White and Asian students than the average school in North Carolina. The bottom 5% of schools using both proficiency rates and the ESSA index also have smaller average enrollments than the average school, 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).

Panel D shows that the lowest performing schools have lower average enrollments, and the highest performing schools have larger average enrollments under both indices. Finally, Panel E shows that the bottom 5% of schools employ more novice teachers and offer lower pay than the average school, consistent with a large literature showing that students of color and economically disadvantaged students learn from less qualified teachers than their White and more advantaged peers (e.g., Simon & Johnson, 2015). To some extent, these differences in pay are likely attributable to the higher performing schools having more experienced teachers—though the teacher experience figures suggest experience is unlikely to explain the full difference in pay between the lowest and highest proficiency schools in particular. For example, teacher pay supplements—or pay above the state-minimum salary schedule—in the highest proficiency schools are 45% higher on average than supplements in the lowest proficiency schools, whereas teacher experience is only 36% higher in the highest proficiency schools. Under the ESSA-aligned index, the experience and pay differences match more closely, with both teacher experience and pay being about one-third higher in the highest rated schools than in the lowest rated schools.

**Indices**

We develop three ESSA-compliant school quality indices to use in our analyses. The goal is to start from a weighting scheme that is representative of the modal state with respect to proficiency and growth in particular, and then to vary the weights so that one index weights proficiency more heavily one weights growth more heavily than the modal state. To do so, we coded all state ESSA plans and selected weights that represented the approximate midpoint of state weights. Each index uses the same metrics: proficiency rates calculated by the state (math and ELA), student achievement growth measured as value-added (Education Value-Added Assessment System, or EVAAS, in this case, which is calculated as part of the state's accountability system),[3] graduation rate for high schools, English learner proficiency, and chronic absenteeism.[4] We convert scores on each metric to percentile ranks in a given year in order to place them on a consistent scale before combining them into a single index score.[5] In creating our three indices, we toggle only the weights on proficiency and growth while holding the other metrics constant.

The first index follows the most common ESSA weighting scheme ("modal index") across states based on our coding of all state ESSA plans. Here, proficiency and growth are weighted

[3] Although many states began calculating growth more recently, North Carolina was an early adopter of growth measures and reporting. The state has been using value-added in some capacity since 2005 and adopted statewide teacher value-added in 2011. The state has been calculating a composite growth index that combines grades and subjects into a single scale for the duration of the study.

[4] We choose chronic absenteeism because it is the most prevalent school quality and student success indicator used in state ESSA plans. 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. To calculate English learner proficiency and chronic absenteeism, we had access to student-level data, which we then aggregated to the school level.

[5] We chose to use percentile ranks because it is a simple, easily replicable approach to placing each metric on the same scale. This choice, which creates a uniform distribution of index scores, has implications for the results (Atchison et al., 2023). For example, research comparing use of this approach with a "natural distribution" approach that preserves the shape of each underlying variable's distribution shows that using a percentile approach can result in the stability of the CSI designation being lower for elementary and middle schools and higher for high schools than a system using a natural distribution. On the other hand, the stability of the index score itself may be higher (Atchison et al., 2023).

similarly at the elementary/middle school levels (35% proficiency and 40% growth), and proficiency is weighted twice as much as growth for high schools (30% proficiency and 15% growth). The second weighting scheme ("higher proficiency") weights proficiency more heavily than growth, with 60% proficiency and 15% growth for elementary and middle schools, and 45% proficiency and 0% growth for high schools. The third ("higher growth") weights growth more heavily than proficiency, with 60% growth and 15% proficiency for elementary and middle schools, and 30% growth and 15% proficiency for high schools. Appendix Table A-1 provides the weights for each index.

**Simulating School Ratings and CSI Schools**

In addition to varying the weighting scheme, we also vary the years of data used to calculate ratings and determine CSI schools. Most states use only one year of data to calculate school ratings and identify CSI schools—after calculating a rating for each school in the CSI identification year, they select the bottom 5% of schools in that year (McNeill et al., 2021; Meyers et al., 2023). 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 generate more stable ratings and designations (Kane & Staiger, 2002; McEachin & Polikoff, 2012). We therefore toggle the number of years of data used to simulate school ratings and identify CSI schools in each study year and create two versions of each index:

- **Single year (status quo):** Use one year of data to calculate school ratings and determine CSI schools, the most common approach in state ESSA plans.
- **3-year weighted mean:** Calculate school ratings and determine CSI schools based on a weighted mean of the scores across three years, where the current year is weighted most heavily (times 3), one year prior is weighted next most heavily (times 2), and two years prior is weighted the least (times 1).

**Applying the Conceptual Framework to Assess Accountability Systems**

***Validity***

Although we do not directly assess the construct and consequential validity of each index in our study, we are able to assess the predictive validity, or the extent to which indices are predictive of future school ratings. Specifically, for each index, we estimate a series of bivariate regressions predicting the one-year index score one, two, and three years in the future. For example, the equations using the modal index and predicting one year into the future would take the form:

$$1YearModalIndex_{it} = \beta_0 + \beta_1 1YearModalIndex_{it-1} + \varepsilon_{it} \quad (1a)$$
$$1YearModalIndex_{it} = \beta_0 + \beta_1 3YearModalIndex_{it-1} + \varepsilon_{it} \quad (1b)$$

predicting the modal index score for school *i* in year *t* as a function of the modal index constructed using a single year of data in year *t*-1 (1a). We next estimate a parallel equation replacing the modal index constructed using a single year of data with the modal index constructed using the three-year weighted mean on the right side of the equation (1b). We run these same equations for the other two index weighting structures (i.e., high proficiency, high growth). Next, we repeat the same set of equations replacing the one-year lagged metric on the right side of the equation with a two-year lag, and then a three-year lag. For each set of models, we descriptively examine the difference between $\beta_1$ and the model $R^2$ in Equation 1a and Equation 1b. Intuitively, this approach provides evidence on the indices' predictive validity, defined as the extent to which one versus three years of data is predictive of the following year's index score within a weighting scheme.

As discussed in our conceptual framework, construct validity depends on the purpose of the school accountability system. If the purpose is to isolate the school's contribution to student learning, school quality indices that weight growth more heavily than proficiency necessarily have greater construct validity. If the purpose is to target resources to schools with the lowest achievement, school quality indices that weight proficiency more heavily than growth have greater construct validity. Consequential validity is greater if the accountability system targets resources to schools in a way that facilitates school improvement and improves student outcomes. All forms of validity are greater when school ratings and CSI schools are identified based on true school performance, and not idiosyncratic error.

***Stability***

We estimate the stability of school ratings by calculating Pearson's correlation coefficients between ratings in consecutive years overall and within quintiles of performance to understand whether there is differential volatility at different points in the distribution. We note that quintiles are a somewhat arbitrary choice, but we use quintiles because they seem to better reflect the reality of a family's choice set than the full distribution of schools. For example, families with the resources to attend a 95th percentile school are likely not deciding between schools in the 95th and fifth percentiles. To them, the stability of the top 20% may be more relevant than the overall stability. To a median income family, the middle three deciles are likely more relevant. Families with the fewest resources may have limited choices, all near the bottom of the distribution—whether there is a meaningful difference between the first and 19th percentile, then, will have implications for their choice. In sum, the public is using ratings to make decisions within a constrained choice set, so the stability of the ratings within those constraints may have greater practical relevance than the overall stability.

Statistically, the correlations within the quintiles will be lower than correlations overall because there is less variation. The overall correlations do not share the problem of arbitrary cutpoints, though they will be less reflective of the choice set of a family evaluating their schooling options. Therefore, the two correlations should be interpreted in tandem with one another and could be considered as a sort of upper and lower bound; the within-quintile correlations may underestimate the signal in a given family's choice set to the extent that families have more heterogeneous school markets, whereas the overall correlations will likely overestimate the signal in a given family's choice set. To determine the stability of the CSI list, we then calculate the share of schools identified in a given year that would have also been identified in the prior and subsequent years, respectively, under each simulation.[6,7]

***Equity***

To explore implications for the equity of school ratings, we examine average student compositions, novice teacher rate, and teacher pay supplement by rating quintile and index. Disproportionality would be apparent in ratings that differed systematically by student demographics. As noted above, disproportionality may stem from an unfair rating system or from broader inequities that lead to lower educational quality in schools serving more marginalized student populations. In the former case, disproportionality would reflect inequity of the ratings themselves. In the latter case, disproportionate ratings would be an accurate reflection of reality. Ratings that shed light on educational inequalities are not inequitable for simply pointing it out; indeed, that is a core purpose of public accountability. Rather, ratings are unfair if they penalize accountable for factors outside of their control. Our analyses cannot disentangle these two possibilities; instead, they shed light on the disproportionality of the ratings and designations to elicit consideration of the implications for equity.

To explore the equity of the CSI designation, we disaggregate schools into three groups (i.e., lowest 25%, middle 50%, and highest 25%) for each of the five variables (i.e., percent economically disadvantaged, Black, and Hispanic, and novice teacher rate and teacher pay supplement).[8] We then calculate the share of schools within each of the three groups that would be designated as a CSI

[6] In addition, in alignment with McEachin & Polikoff (2012), we calculate the number of years each school spends in the bottom 5% for each index. We present these results in the appendix (Figure A-4).
[7] It is useful to note that due to the way North Carolina identified its lowest performing schools under Race to the Top and a subsequent state-initiated turnaround (largely based on proficiency), some of the schools in this analysis received turnaround services. That said, the research on the effects of these interventions (Heissel & Ladd, 2018; Henry & Guthrie, 2019; Henry & Harbatkin, 2020) suggests that any resulting bias in the stability measures from accountability-system-induced school improvement would be minimal.
[8] We also run analyses breaking the sample into three equally sized groups and present the tercile analysis in an appendix (Figure A-1).

school based on (a) each of the three weighting schemes, and (b) one versus three years of data. By comparing the share of the middle 50% identified as CSI with the share of the top and bottom quartiles identified, we reveal differences between the average school and the tails of the distribution, where consequential accountability is most salient (Davis et al., 2015; Hasan & Kumar, 2019). Greater disproportionalities across demographic quartiles again could stem from an unfair accountability system or from true disparities in access to a quality education. Meanwhile, an equitable CSI list by the access dimension would identify more schools serving students with greater economic needs and with fewer resources *and* provide these schools with additional supports and funding to narrow resource gaps. We do not observe whether schools were provided with additional supports in this study.

## Findings

### Stability

Figure 1 shows scatterplots of school ratings in consecutive years, with rows 1–3 toggling through the three ESSA school quality indices and Panels A–B toggling the number of years of data used to construct them. Overall correlations are reported at the bottom right-hand corner of each graph, and correlations within quintiles are reported at the top. Across the full distribution of schools, the year-to-year correlation using the modal index is .700 when calculated based on a single year of data and .922 when calculated using three years. In other words, the percentile rank in year $t$ explains 49% of the variation in the percentile rank in year $t$+1 using a single year of data compared with 85% using three years of data.

Comparing Panels A and B within quintiles, 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 the variation in the subsequent ratings for schools in the lowest quintile of performance compared with only 4-18%

when using only one year of data.[9] For schools in the highest quintile, using three years of data explains 26–55% of the variation in subsequent ratings compared with 6-19% when using a single year. And for schools in the three middle quintiles, using three years of data explains 7–32% of the variation in subsequent ratings compared with 1–8% when using one year. Therefore, small shifts in simulated ratings appear to reflect idiosyncratic year-to-year variation more than persistent signals of quality, but ratings include more signal when they draw on three years of data rather than one. However, we note that one-year changes are centered around zero for each of the three indices and number of years of data, as shown in Appendix Figure A-3.

Figure 1 also highlights that ratings are more stable over time under the high proficiency index (Row 2) and less stable under the high growth index (Row 3). This is unsurprising as proficiency rates are strongly correlated with persistent student characteristics, such as family income, whereas growth measures are more volatile over time (Boyd et al., 2008; Kane & Staiger, 2002; McEachin & Polikoff, 2012). Consistent with prior research and statistical expectations, there is also greater stability in the tails of the distribution.

FIGURE 1

Table 2 provides results for the stability of the CSI list for each ESSA index (panels) and the number of years of data used to create the list (rows). As expected, using three years of data produces more stable CSI lists than using a single year. Across all three weighting schemes, using three years of data generates a CSI list that is two to nearly three times more stable than when using a single year. When using only one year of data, 13–25% of CSI schools in year *t* would also be identified in both the previous and subsequent years. This finding means when using only one year of data, the majority of CSI schools would not have been identified in consecutive years within the

[9] These and all subsequent percentages reported can be calculated by squaring the Pearson's *r* coefficients shown in the figures. For example, the *r* for the lowest quintile using the high growth index in Row 3 is 0.198, and $0.198^2=0.039$, or about 4% (the lower end of the reported range using one year of data), while the *r* for the lowest quintile using the high proficiency index in Row 2 is 0.424, and $0.424^2=0.180$, or 18% (the upper end of the reported range).

ESSA three-year accountability cycle. As expected, the higher proficiency index (Panel B) generates the most stable CSI list, and the higher growth index (Panel C) generates the least stable list. Under the most stable conditions (i.e., three-year weighted mean, higher proficiency index), about half of schools would be identified in each of the three consecutive years. However, using three years of data instead of one 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 as CSI schools for three consecutive years using one year of data and the higher growth index (Panel C, row 1), 34% would be identified for three consecutive years using three years of data—a 36% improvement over the stability of the CSI list using only one year of data and the higher proficiency index (Panel B, row 1).

TABLE 2

**Equity**

Following with our conceptual framework, we consider two dimensions of equity, beginning with disproportionality and then moving to equity of access. Figure 2 shows school characteristics (panels) by rating quintile (bar clusters) for each of the three ESSA indices (bar colors).[10] All three indices show disproportionality on demographic measures, with the share of economically disadvantaged students (Panel A), and, to a lesser extent, Black and Hispanic students (Panels B and C), decreasing monotonically as school ratings increase. Meanwhile, the novice teacher rate decreases and the teacher pay supplement size increases as school ratings increase, underscoring that higher-rated schools are able to offer higher teacher pay and likely face fewer recruitment and retention challenges. This latter finding also shows that the higher performing schools have more qualified teachers, which likely stems from broader inequities that may be contributing to lower performance. To the extent that the more qualified teachers at the higher-rated schools also

[10] For simplicity of exposition, Figure 2 shows ratings using just one year of data; however, the findings are substantially similar using three-year weighted means (Figure A-1).

contribute to a more effective education, the demographic disproportionality in ratings would reflect these broader inequities rather than unfair ratings.

Panels D and E shed light on equity of access, suggesting that all three indices could improve equity of access if they result in policymakers taking action to reduce resource gaps between the lowest and highest rated schools. Because the higher proficiency index most sharply distinguishes the schools with the lowest levels of teacher experience and pay, it could be the most effective option for increasing equity of access if policymakers responded to it by narrowing the teacher experience and pay gap between the schools in the lowest and highest rating quintiles. However, improving equity of access depends on the political will to narrow these resource gaps. In the absence of that political will, the accountability systems would remain inequitable regardless of which ESSA index is selected.

FIGURE 2

Figure 3 illustrates the equity implications for the CSI list, with bar heights representing the share of schools in each grouping (lowest 25%, middle 50%, and highest 25%) that would be identified for CSI by ESSA index and number of years of data.[11] The left column shows the share of schools in the bottom 5% based on a single year of data while the right shows the share of schools in the bottom 5% based on the three-year weighted mean. The horizontal dashed line at 0.05 (or 5%) provides a benchmark for assessing disproportionality and equity of access. A system without disproportionality would yield three sets of bars that are exactly at the 0.05 line. Meanwhile, overidentifying the schools with greatest needs, such as those in the highest quartile of novice teacher rates and the lowest quartile of teacher pay, could improve equity of access, if substantial additional resources were provided to the schools.

[11] Findings are very similar when we use terciles instead of quartiles (Figure A-2).

The large discrepancies in bar heights across the low, middle, and high groupings on each of the five variables highlight that there is disproportionality in each of the simulations because they all overidentify underserved schools for CSI. Schools with the highest percentages of economically disadvantaged students (Panel A), Black students (Panel B), and novice teachers (Panel D) in particular are heavily overidentified as CSI schools. It is important to note that we cannot disentangle the extent to which the demographic disproportionality emanates from an unfair rating system vs. from broader societal inequities leading to lower quality education in schools serving more disadvantaged student populations. However, again, the overrepresentation of schools with high novice teacher rates provides evidence that all of these ratings are indeed capturing some true differences in school quality. The higher proficiency index demonstrates greater disproportionality than the higher growth index, but all indices yield substantial disproportionality along these dimensions. Additionally, the close similarities between columns show that the number of years of data is not consequential for disproportionality; using three years of data instead of one does not meaningfully improve or impair the disproportionality of the CSI list.

Similar to the findings for the equity in access of school ratings, to the extent that under-resourced schools are identified for CSI and provided with meaningful additional resources, any index could, in theory, improve equity of access. Panels D and E show that the schools employing the highest rates of novice teachers and offering the lowest teacher pay are disproportionately likely to be identified for CSI. The CSI designation could come with turnaround supports that serve to close the teacher pay gap and improve teacher recruitment and retention in CSI schools.

FIGURE 3

Our findings that all three indices produce demographically disproportionate ratings and CSI designations may be surprising given the literature showing that growth-based indices are less correlated with student demographics than proficiency-based ones (McEachin & Polikoff, 2012;

Reardon, 2019b). Our findings do not conflict with the literature, as the ratings and designations using the higher growth index are not as strongly associated with demographics as those using the higher proficiency index. Yet, the higher growth index still produces disproportionate ratings for at least four reasons. The first is that it still includes proficiency; ESSA requires that achievement is included and has a "substantial weight" in the index calculation. Second, it is not the case that growth is completely uncorrelated with demographics (Ehlert et al., 2014). Students with differing levels of opportunity to learn outside of the school building may learn at different rates (Amrein-Beardsley, 2008), which would affect growth. Third, schools with high (or low) concentrations of student disadvantage may be impacted by negative (positive) peer effects (Hanushek et al., 2003; Rothstein, 2006; Sacerdote, 2011), which could lead to index scores that systematically vary by student prior achievement. Finally, it is likely that the growth measures are to some degree capturing true differences in school quality that are correlated with student demographics.

**Validity**

In this section, we describe our findings on predictive validity, for which we present empirical analyses, and deal with consequential and construct validity as part of the discussion in the next section. Table 3 summarizes the results of the models predicting one-year ratings one (Column 1, labeled *t*+1), two (Column 2, labeled *t*+2), and three (Column 2, labeled *t*+2) years into the future based on the past rating(s). Across all three weighting schemes (i.e., modal, high proficiency, and high growth), the index based on the three-year weighted mean has greater predictive validity than the index based on a single year of data.

The simulated scores are scaled as percentiles, so the coefficients in Panel A1 indicate that a one percentile point increase on the modal index using a single year of data in year *t* corresponds to increases of 0.7 percentile points in *t*+1, 0.62 percentile points in *t*+2, and 0.56 percentile points in *t*+3. The year *t* index here explains 49% of the variation in year *t*+1, 39% in *t*+2, and 31% in *t*+3.

Panel A2 shows that the relationship between the three-year weighted mean index and future performance is steeper (demonstrated by the coefficient estimates) and stronger (demonstrated by the standard errors and $R^2$). Specifically, a one percentile point increase on the modal index using a single year of data in year *t* corresponds to increases of 0.89 percentile points in *t*+1, 0.79 percentile points in *t*+2, and 0.66 percentile points in *t*+3. The year *t* index here explains 79% of the variation in year *t*+1, 62% in *t*+2, and 44% in *t*+3. These patterns hold across all three weighting schemes. In sum and as expected, the index using three years of data has greater predictive validity than the index using a single year of data.

TABLE 3

Comparing across weighting schemes, the results follow the same pattern as our stability findings, with the high proficiency index having the greatest predictive validity and the high growth index having the lowest. This finding is unsurprising given that proficiency rates measure educational opportunity since birth, which generally does not substantially change from one year to the next, whereas student growth is more volatile (Boyd et al., 2008; Kane & Staiger, 2002; McEachin & Polikoff, 2012). We discuss implications of our empirical findings on construct and consequential validity in the discussion.

## 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. We explore states' first accountability systems under ESSA and find that initial plans present challenges in terms of the stability, equity, and validity of school ratings and the identification of the bottom 5% (or CSI) schools.

While the future of federal educational policy is uncertain, two things about school accountability remain steadfast. First, states will continue to maintain school accountability systems whether or not it is required by federal law. Twenty-two states had school accountability systems prior to NCLB's passage (Hanushek & Raymond, 2005), and even after NCLB was replaced with more flexible rules many states maintained the systems and interventions they developed under NCLB (Meyers et al., 2023; VanGronigen & Meyers, 2019). Second, educational policy swings like a pendulum; even if federal school accountability policy weakens under one administration, it is likely to come back under another.

Our findings point to several implications for states refining their accountability systems. First, states should draw on multiple years of data to both calculate school ratings and identify their CSI schools. Using multiple years of data will lead to more stable ratings and CSI designations with greater predictive validity—and therefore result in a better signal of school quality. A large literature on teacher value-added reaches the same conclusion—that using multiple years of data increases precision and reduces statistical uncertainty and is needed for high-stakes decisions (e.g., Goldhaber & Hansen, 2012; McCaffrey et al., 2009).

Second, states should reconsider how school ratings are reported and how families can interpret them. Our findings suggest that a rank-ordering of all schools is not sufficiently stable to support public decision-making. For example, a 100-point school rating implies that movement from the 20th to 40th percentile reflects meaningful school improvement, when in reality, our analyses suggest that much of that movement is driven by factors outside the school's control and statistical noise. On the other hand, ratings can reliably differentiate between schools in the 20th percentile versus the 80th percentile—though this is not the choice presented to most families in deciding where to send their children to school due to constraints in school options. More stability in the tails also suggests that CSI schools may have consistently lower ratings. However, even in the best-case scenario in our simulations, only half of schools on the CSI list would be identified as CSI schools across three consecutive years.

Third, states should be intentional about the weighting schemes used in their school quality indices, and each state's index should support its accountability system's goal because the framing of the ratings matter for their construct validity. A school quality index similar to our higher growth index will better characterize *performance*, whereas a school quality index similar to our higher proficiency index will better characterize *achievement*. Relatedly, federal lawmakers and policymakers should consider the purpose of school accountability systems when reauthorizing the Elementary and Secondary Education Act (ESEA). The term "low performing" appears throughout the language of ESSA, even as it calls for states to capture *achievement* as opposed to school *performance*. If the goal is to identify the schools with the greatest needs for supports, the law should specify metrics that identify opportunity and resource gaps, and characterize these schools as being under-resourced instead of low performing. If the goal is to identify schools that are underperforming in terms of improving student achievement, the law should call for accountability systems that reliably capture school contributions to student learning.

Intrinsic in our findings is that accountability systems involve tradeoffs. One important tradeoff is related to the multiple-metrics approach that became universal under ESSA. A system that accounts for multiple metrics can encourage educational leaders to direct resources to students across the full distribution of achievement and based on outcomes beyond test scores. This advantage is near-universally considered to be an improvement over the NCLB era's hyperfocus on proficiency. However, using multiple metrics has implications for the validity, stability, and equity of the broader index. A growth component may contribute less demographically disproportionate ratings but also less resource equity, whereas a proficiency component may lead to more disproportionate ratings but greater stability. A decision to weight one component more than another reflects an implied judgment about the underlying properties of each component.

Our findings shed light on some of these tradeoffs, which we summarize in Table 4, with Panel A comparing the higher proficiency and higher growth indices[12] and Panel B comparing decisions to use one vs. three years of data. In sum, the higher proficiency index has greater construct validity for identifying schools with low *achievement* and greater predictive validity over time, and is better at identifying the most under-resourced schools in the state—but it also systematically assigns lower ratings to schools serving underserved student populations and would sanction schools for factors outside their control. That said, imposing interventions in schools with low proficiency but higher growth could serve to undermine rather than support school improvement if those interventions lack coherence with ongoing strategies that are already showing promise. The higher growth index has greater construct validity for identifying low-*performing* schools and would result in less disproportionate ratings and CSI designations—but it produces less stable ratings from year to year and may not narrow resource gaps as much as the higher proficiency index if resources are tied to CSI designations.

With respect to years of data, the three-year weighted mean performs better for all indices, although it may be less politically palatable if stakeholders perceive the inclusion of past data as less fair. Importantly, using three years of data in an index does not necessarily mean using three-year averages for school improvement planning. Educators ought to continue to draw on granular, student-level data to assess student needs and guide instructional decisions.

TABLE 4

Our analyses also underscore a broader point: there is no agreed-upon definition of school quality, and any accountability system is necessarily value-laden. For example, although we employ multidimensional metrics in our indices, each of our indices still places the greatest weight on student test scores, as required by ESSA. Yet some stakeholders do not view test scores as the most

[12] We do not show the modal index because it represents the midpoint between the higher proficiency and higher growth options.

important educational outcome, regardless of how they are operationalized. Focusing almost exclusively on test scores while ignoring other school factors and family constraints in school choice thus privileges one conception of school quality. In addition, regardless of intent, school accountability systems have consistently produced substantially higher ratings for schools serving the greatest shares of economically advantaged and White students, increasing demand for these schools and exacerbating access barriers for families with less economic wealth and families of color. Conversely, schools serving the largest shares of economically disadvantaged students and students of color have received systematically lower ratings, contributing to declines in public funding and support, school closures, increased teacher turnover, and other destabilizing pressures.

Finally, our findings highlight the need for additional research on the consequential validity of school accountability systems, that is, whether systems lead to improved student outcomes or instead contribute to further disinvestment in already under-resourced schools. State and federal policymakers are ultimately accountable for policies as implemented, not merely policies as intended. Accordingly, our results call for further examination of whether school accountability systems are producing their desired effects. If we do not pay attention to consequential validity, accountability systems risk doing more harm than good. CSI designations are often accompanied by additional resources, yet in practice, these supports have been insufficient to substantially improve schools (Atchison & Blair, 2025; Blair et al., 2025; Hyslop & Zhou, 2024). If we are serious about school improvement, then consequential validity must be tracked, and accountability systems should be adjusted when they fail to produce the desired outcomes.

Under any school accountability system, policymakers will inevitably face tradeoffs in terms of stability, equity, and validity. This paper provides a starting point for understanding the implications of school accountability decisions in terms of each, and to spur further conversation about how to improve the way that we rate and talk about schools.

**References**


Abdulkadiroğlu, A., Pathak, P. A., Schellenberg, J., & Walters, C. R. (2020). Do Parents Value School Effectiveness? *American Economic Review*, *110*(5), 1502–1539. https://doi.org/10.1257/aer.20172040

Amrein-Beardsley, A. (2008). Methodological Concerns About the Education Value-Added Assessment System. *Educational Researcher*, *37*(2), 65–75. https://doi.org/10.3102/0013189X08316420

Atchison, D. (2020). The Impact of Priority School Designation Under ESEA Flexibility in New York State. *Journal of Research on Educational Effectiveness*, *13*(1), 121–146. https://doi.org/10.1080/19345747.2019.1679930

Atchison, D., & Blair, D. (2025). *School Accountability, Underfunding, and Inefficiency*. American Institutes for Research. www.air.org/sites/default/files/2025-06/School-Accountability-Underfunding-and-Inefficiency-Brief-June-2025.pdf

Atchison, D., Hyland, K., Ozek, U., Le Floch, K., & Hurlburt, S. (2023). *Identification of Comprehensive Support and Improvement Schools in Ohio*. American Institutes for Research. https://www.air.org/sites/default/files/2024-04/23-20936_Ohio_Measures_Report_Final.pdf

Atchison, D., Özek, U., Le Floch, K. C., & Hurlburt, S. (2025). Effects of Comprehensive Support and Improvement School Designation Under the Every Student Succeeds Act on Student Outcomes: Evidence from Ohio. *Journal of Research on Educational Effectiveness*, *0*(0), 1–35. https://doi.org/10.1080/19345747.2025.2537131

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

Ballou, D., & Springer, M. G. (2016). Has NCLB Encouraged Educational Triage? Accountability and the Distribution of Achievement Gains. *Education Finance and Policy*, *12*(1), 77–106. https://doi.org/10.1162/EDFP_a_00189

Blair, D., Atchison, D., & Le Floch, K. C. (2025). Investing in improvement: A critical look at funding for low-performing schools. *Educational Policy*, 08959048251318626. https://doi.org/10.1177/08959048251318626

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*.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the short careers of high-achieving teachers in schools with low-performing students. *American Economic Review*, *95*(2), 166–171.

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2015). Consequential Validity of Accountability Policy: Public Understanding of Assessments. *Educational Policy*, *29*(5), 711–745. https://doi.org/10.1177/0895904813518099

Burrell, N., & Harbatkin, E. (2024). Beyond the school building: Examining the association between of out-of-school factors and multidimensional school grades. *Education Policy Analysis Archives*, *32*. https://doi.org/10.14507/epaa.32.8497

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

Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305–331. https://doi.org/10.3102/01623737024004305

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.

Darling-Hammond, L. (2018). From “separate but equal” to “No Child Left Behind”: The collision of new standards and old inequalities. In *Thinking about schools* (pp. 419–437). Routledge.

Davis, T., Bhatt, R., & Schwarz, K. (2015). School segregation in the era of accountability. *Social Currents*, *2*(3), 239–259. https://doi.org/https://doi.org/10.1177/2329496515589852

Dee, T. (2012). *School Turnarounds: Evidence from the 2009 Stimulus* (Working Paper No. 17990). National Bureau of Economic Research. https://doi.org/10.3386/w17990

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School Accountability, Postsecondary Attainment, and Earnings. *The Review of Economics and Statistics*, *98*(5), 848–862. https://doi.org/10.1162/REST_a_00598

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.

Education Commission of the States. (2021, December). *50-State Comparison: States’ School Accountability Systems*. https://www.ecs.org/50-state-comparison-states-school-accountability-systems/

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri. *Statistics and Public Policy*, *1*(1), 19–27. https://doi.org/10.1080/2330443X.2013.856152

Feng, L., Figlio, D., & Sass, T. (2018). School accountability and teacher mobility. *Journal of Urban Economics*, *103*, 1–17.

Figlio, D., & Loeb, S. (2011). Chapter 8: School accountability. In *Handbook of the Economics of Education* (Vol. 3, pp. 383–417). Elsevier. doi.org/10.1016/B978-0-444-53429-3.00008-9

Figlio, D. N., & Lucas, M. E. (2004). What’s in a Grade? School Report Cards and the Housing Market. *American Economic Review*, *94*(3), 591–604. https://doi.org/10.1257/0002828041464489

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

Finnigan, K. S., & Gross, B. (2007). Do Accountability Policy Sanctions Influence Teacher Motivation? Lessons From Chicago’s Low-Performing Schools. *American Educational Research Journal*, *44*(3), 594–630. https://doi.org/10.3102/0002831207306767

Gara, T. V., Farkas, G., & Brouillette, L. (2022). Did consequential accountability policies decrease the share of visual and performing arts education in U.S. public secondary schools during the No Child Left Behind era? *Arts Education Policy Review*, *123*(4), 218–235. https://doi.org/10.1080/10632913.2020.1854911

Glazerman, S., & Dotter, D. (2017). Market Signals: Evidence on the Determinants and Consequences of School Choice From a Citywide Lottery. *Educational Evaluation and Policy Analysis*, *39*(4), 593–619. https://doi.org/10.3102/0162373717702964

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has It Always Been This Way? Tracing the Evolution of Teacher Quality Gaps in U.S. Public Schools. *American Educational Research Journal*, *55*(1), 171–201. https://doi.org/10.3102/0002831217733445

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2015). A Comparison of Student Growth Percentile and Value-Added Models of Teacher Performance. *Statistics and Public Policy*, *2*(1), 1–11. https://doi.org/10.1080/2330443X.2015.1034820

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, *18*(5), 527–544. https://doi.org/10.1002/jae.741

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. https://doi.org/10.1002/pam.20091

Harbatkin, E. (2022). Staffing for School Turnaround in Rural Settings. *Leadership and Policy in Schools*, *0*(0), 1–23. https://doi.org/10.1080/15700763.2022.2058963

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*.

Hastings, J. S., & Weinstein, J. M. (2008). Information, School Choice, and Academic Achiement: Evidence from Two Experiments. *Quarterly Journal of Economics*, *123*(4), 1373–1414. https://doi.org/10.1162/qjec.2008.123.4.1373

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

Heissel, J. A., & Ladd, H. F. (2018). School turnaround in North Carolina: A regression discontinuity analysis. *Economics of Education Review*, *62*, 302–320. https://doi.org/10.1016/j.econedurev.2017.08.001

Henry, G. T., & Guthrie, J. E. (2019). The Effects of Race to the Top School Turnaround in North Carolina. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://edworkingpapers.com/ai19-107

Henry, G. T., & Harbatkin, E. (2020). The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Intervention. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2020.1814464

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

Holland, P. W. (2002). Two Measures of Change in the Gaps between the CDFs of Test-Score Distributions. *Journal of Educational and Behavioral Statistics*, *27*(1), 3–17.

Houston, D. M., & Henig, J. R. (2021). The Effects of Student Growth Data on School District Choice: Evidence from a Survey Experiment. *American Journal of Education*, *127*(4), 563–595. https://doi.org/10.1086/715003

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

Hursh, D. (2007). Exacerbating inequality: The failed promise of the No Child Left Behind Act. *Race Ethnicity and Education*, *10*(3), 295–308. https://doi.org/10.1080/13613320701503264

Hutt, E., & Polikoff, M. S. (2020). Toward a Framework for Public Accountability in Education Reform. *Educational Researcher*, *49*(7), 503–511. https://doi.org/10.3102/0013189X20931246

Hyslop, A., & Zhou, Z. (2024, February 1). *When Equity Is Optional: Does Accountability Drive School Spending? | All4Ed*. https://all4ed.org/publication/when-equity-is-optional-does-accountability-drive-school-spending/

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. https://doi.org/https://doi.org/10.3102/0013189X034008003

Krieg, J. M. (2008). Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act. *Education Finance and Policy*, *3*(2), 250–281. https://doi.org/10.1162/edfp.2008.3.2.250

Krieg, J. M., & Storer, P. (2006). How Much Do Students Matter? Applying the Oaxaca Decomposition to Explain Determinants of Adequate Yearly Progress. *Contemporary Economic Policy*, *24*(4), 563–581. https://doi.org/10.1093/cep/byl003

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

Lane, S. (2020). Test-Based Accountability Systems: The Importance of Paying Attention to Consequences. *ETS Research Report Series*, *2020*(1), 1–22. https://doi.org/10.1002/ets2.12283

Lee, J., & Lubienski, C. (2017). The Impact of School Closures on Equity of Access in Chicago. *Education and Urban Society*, *49*(1), 53–80. https://doi.org/10.1177/0013124516630601

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

Meyers, C. V., Brandt, W. C., & VanGronigen, B. A. (2023). State ESSA Plans and Comprehensive Support and Improvement (CSI) Status. *Educational Policy*, *37*(5), 1277–1314. https://doi.org/10.1177/08959048221087209

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

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work)*. Teachers College Press.

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.

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

Rauscher, E. (2020). Does Money Matter More in the Country? Education Funding Reductions and Achievement in Kansas, 2010–2018. *AERA Open*, *6*(4), 2332858420963685. https://doi.org/10.1177/2332858420963685

Reardon, S. F. (2019a). 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/

Reardon, S. F. (2019b). Educational Opportunity in Early and Middle Childhood: Using Full Population Administrative Data to Study Variation by Place and Age. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *5*(2), 40–68. https://doi.org/10.7758/rsf.2019.5.2.03

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

Rice, J. K., & Malen, B. (2003). The Human Costs of Education Reform: The Case of School Reconstitution. *Educational Administration Quarterly*, *39*(5), 635–666. https://doi.org/10.1177/0013161X03257298

Rothstein, J. M. (2006). Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions. *The American Economic Review*, *96*(4), 1333–1350. https://doi.org/https://doi.org/10.1257/aer.96.4.1333

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Economics of education* (Vol. 3, pp. 249–277). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-53429-3.00004-1

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

Simon, N. S., & Johnson, S. M. (2015). Teacher Turnover in High-Poverty Schools: What We Know and Can Do. *Teachers College Record*, *117*(3). (Teachers College, Columbia University. P.O. Box 103, 525 West 120th Street, New York, NY 10027. Tel: 212-678-3774; Fax: 212-678-6619; e-mail: tcr@tc.edu; Web site: http://www.tcrecord.org).

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

VanGronigen, B. A., & Meyers, C. V. (2019). How State Education Agencies Are Administering School Turnaround Efforts: 15 Years After No Child Left Behind. *Educational Policy*, *33*(3), 423–452. https://doi.org/10.1177/0895904817691846

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. https://doi.org/https://doi.org/10.3102/0162373712440039

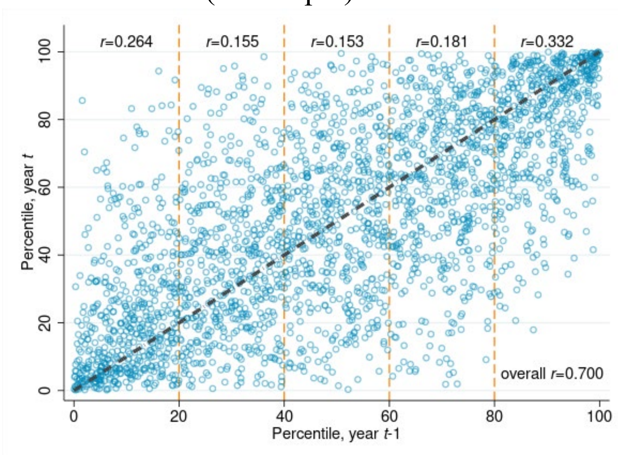
Wright, B. L., & Petrilli, M. J. (2017, November 14). Rating the Ratings: An Analysis of the 51 ESSA Accountability Plans. *The Thomas B. Fordham Institute*. https://fordhaminstitute.org/national/research/rating-ratings-analysis-51-essa-accountability-plans

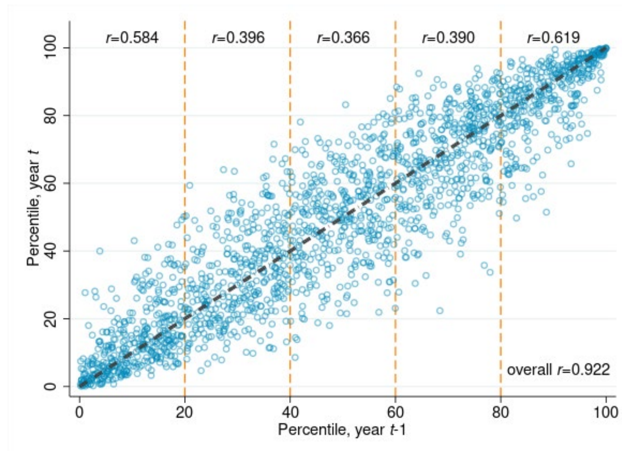
## Figures

**Figure 1. Correlations of school ratings in consecutive years by ESSA index and years of data**

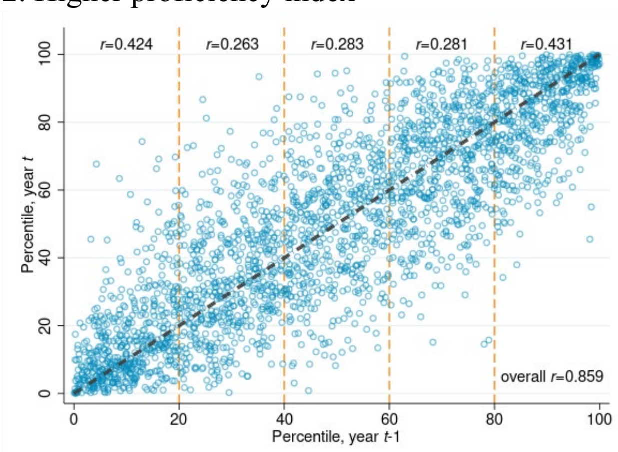
Panel A. Single year (status quo) Panel B. Three-year weighted mean

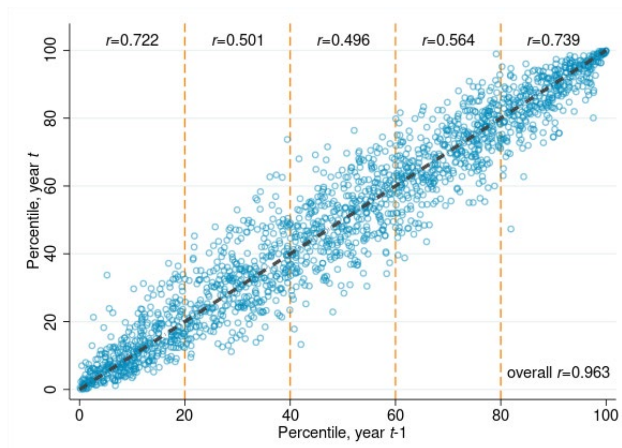
1. Modal index (status quo)



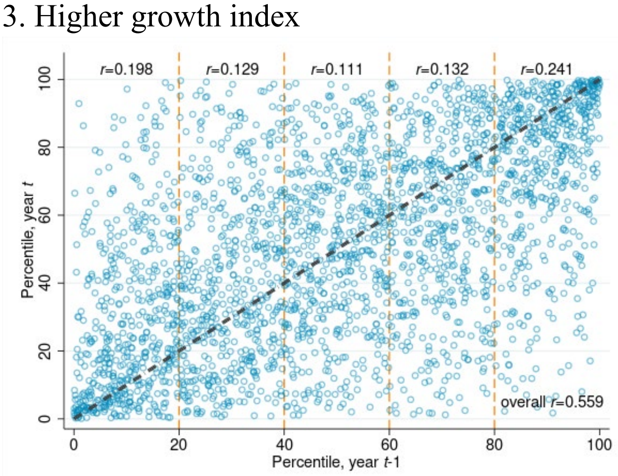



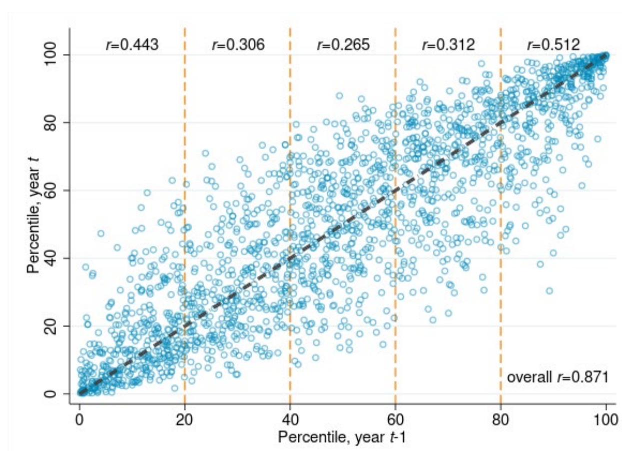

2. Higher proficiency index



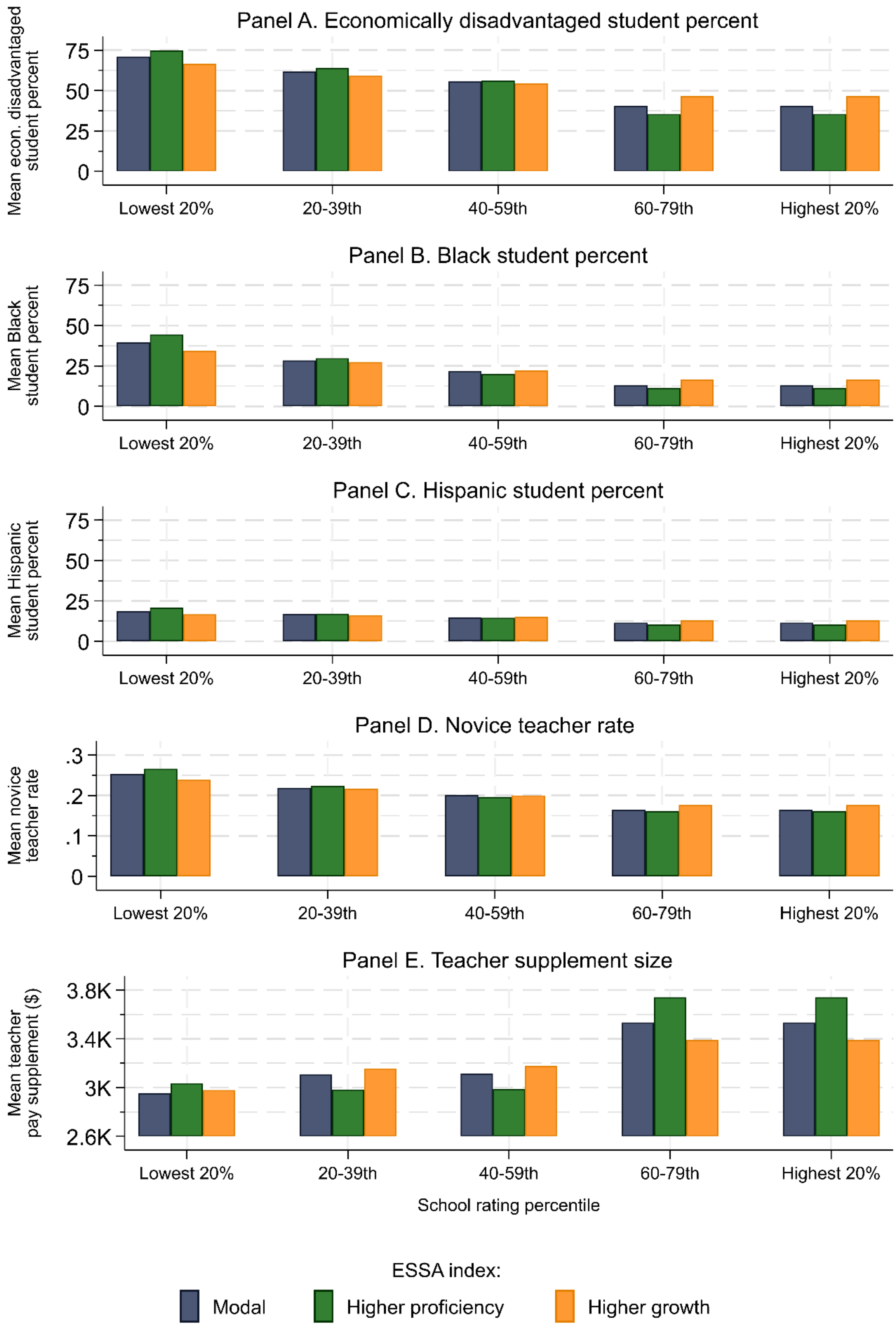

3. Higher growth index



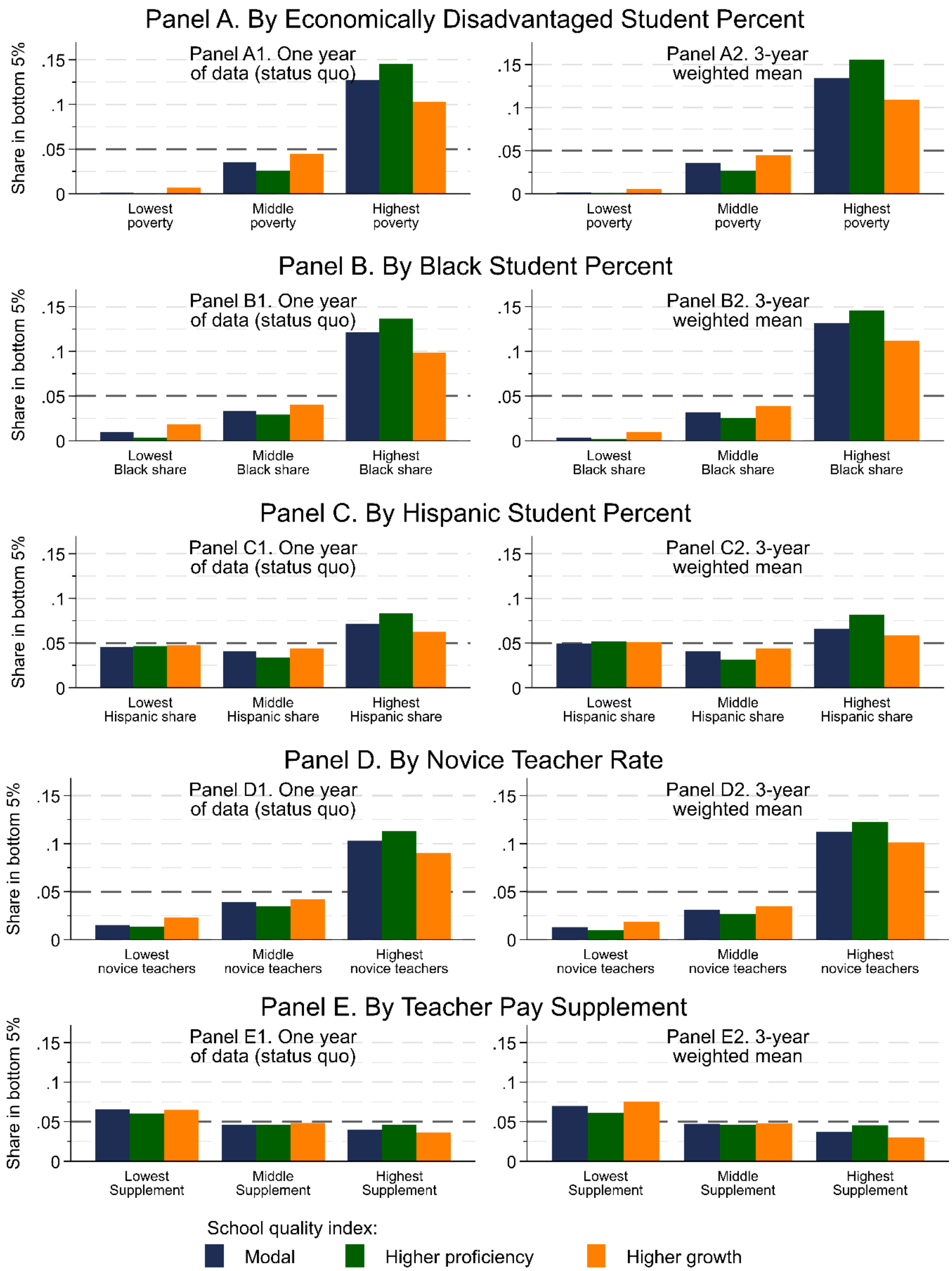

Note: Scatterplots of school ratings percentiles in adjacent years. Points represent random sample of 20% of schools by years for visibility. Correlations are calculated using the full sample.

**Figure 2. School characteristics by rating quintile and ESSA index**



Note: Bar heights represent mean (A) percent economically disadvantaged students, (B) percent Black students, (C) percent Hispanic students, (D) novice teacher rate, and (E) teacher pay supplement size within each quintile of school ratings and ESSA index. Teachers in North Carolina are paid on a state salary schedule based on experience, education, and National Board Certification, but may receive a local pay supplement in addition to the state-determined amount. Thus, the pay supplement is the average amount over and above the salary determined by the state salary schedule.

**Figure 3. Share of CSI schools by groupings based on school characteristics by ESSA index and years of data**



Note: Bar heights represent the share of schools within each grouping of (A) economically disadvantaged student percent, (B) Black student percent, (C) Hispanic student percent, (D) novice teacher rate, and (E) teacher pay supplement size, identified in the bottom 5% based on each ESSA index and number of years of data. Groupings are schools with the lowest 25%, middle 50%, and highest 25% on each school characteristic. “Lowest” groups contain schools with the lowest values on the observed measure and “Highest” groups contain schools with the highest values. For Panels A–D, the low groups represent the *less* disadvantaged schools (i.e., serving fewer students from historically marginalized backgrounds or employing fewer inexperienced teachers) and the high groups represent the *more* disadvantaged schools (i.e., serving more students from historically marginalized backgrounds or employing fewer inexperienced teachers). For Panel E, the low quantile represents the *more* disadvantaged schools (i.e., schools offering the lowest teacher pay) while the high quantile represents the *less* disadvantaged schools (i.e., schools able to offer the highest teacher pay). The horizontal line at 0.05 denotes the 5% threshold—the share of schools that would be identified in each quantile if the subgroups were designated as CSI proportionally.

## Tables

**Table 1. School-level sample statistics overall and by proficiency rate and ESSA rating**

| | | By Proficiency Rate | | | By ESSA Rating from Modal Index | | |
|---|---|---|---|---|---|---|---|
| | All Schools | Bottom 5% | Middle 5% | Top 5% | Bottom 5% | Middle 5% | Top 5% |
| ***A. Proficiency rate (%)*** | | | | | | | |
| | 60.4 | 31.6 | 60.7 | 88.6 | 39.3 | 60.4 | 79.5 |
| | | | (0.64)$^{***}$ | (0.53)$^{***}$ | | (0.76)$^{***}$ | (0.65)$^{***}$ |
| ***B. Locale (%)*** | | | | | | | |
| City | 26.7 | 33.8 | 20.6 | 50.1 | 20.6 | 25.4 | 38.6 |
| | | | (2.27)$^{***}$ | (2.50)$^{***}$ | | (2.16)$^{*}$ | (2.30)$^{***}$ |
| Suburb | 9.0 | 2.8 | 8.0 | 9.8 | 4.1 | 10.4 | 10.5 |
| | | | (1.15)$^{***}$ | (1.24)$^{***}$ | | (1.33)$^{***}$ | (1.33)$^{***}$ |
| Town | 9.0 | 4.6 | 8.6 | 3.1 | 9.6 | 9.5 | 7.6 |
| | | | (1.28)$^{**}$ | (0.99) | | (1.51) | (1.44) |
| Rural | 55.3 | 58.8 | 62.8 | 37.0 | 65.7 | 54.7 | 43.3 |
| | | | (2.51) | (2.50)$^{***}$ | | (2.50)$^{***}$ | (2.50)$^{***}$ |
| ***C. Student Demographics (%)*** | | | | | | | |
| Economically disadvantaged | 55.6 | 82.5 | 57.1 | 22.6 | 76.3 | 54.8 | 35.2 |
| | | | (0.79)$^{***}$ | (0.86)$^{***}$ | | (0.90)$^{***}$ | (0.93)$^{***}$ |
| White | 53.3 | 11.7 | 60.9 | 68.6 | 26.5 | 55.7 | 69.2 |
| | | | (0.92)$^{***}$ | (0.79)$^{***}$ | | (1.22)$^{***}$ | (1.02)$^{***}$ |
| Black | 53.3 | 62.0 | 18.4 | 9.8 | 46.0 | 22.1 | 11.1 |
| | | | (1.03)$^{***}$ | (0.90)$^{***}$ | | (1.17)$^{***}$ | (1.00)$^{***}$ |
| Asian | 24.0 | 1.3 | 1.6 | 8.6 | 1.3 | 1.9 | 5.1 |
| | | | (0.12)$^{*}$ | (0.41)$^{***}$ | | (0.14)$^{***}$ | (0.31)$^{***}$ |
| Hispanic, any race | 15.1 | 18.7 | 14.1 | 9.0 | 18.7 | 14.9 | 10.3 |
| | | | (0.71)$^{***}$ | (0.65)$^{***}$ | | (0.71)$^{***}$ | (0.63)$^{***}$ |
| Indigenous | 2.2 | 3.4 | 0.9 | 0.6 | 3.9 | 1.3 | 0.6 |
| | | | (0.44)$^{***}$ | (0.44)$^{***}$ | | (0.53)$^{***}$ | (0.49)$^{***}$ |
| Pacific Islander | 1.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| | | | (0.02) | (0.02) | | (0.01) | (0.01)$^{*}$ |
| ***D. Enrollment (100s)*** | | | | | | | |
| Enrollment | 6.3 | 4.9 | 6.0 | 7.3 | 4.9 | 6.4 | 7.3 |
| | | | (0.15)$^{***}$ | (0.21)$^{***}$ | | (0.14)$^{***}$ | (0.19)$^{***}$ |
| ***E. Teacher Resources*** | | | | | | | |
| Novice teacher (%) | 20.4 | 30.7 | 19.2 | 17.1 | 28.2 | 20.2 | 15.7 |
| | | | (0.57)$^{***}$ | (0.61)$^{***}$ | | (0.60)$^{***}$ | (0.55)$^{***}$ |
| Teacher experience (years) | 11.0 | 8.6 | 11.2 | 11.7 | 9.1 | 11.0 | 12.2 |
| | | | (0.18)$^{***}$ | (0.20)$^{***}$ | | (0.19)$^{***}$ | (0.18)$^{***}$ |
| Teacher pay supplement[a] ($) | 3,189 | 2,941 | 2,922 | 4,263 | 2,800 | 3,125 | 3,733 |
| | | | (94) | (102)$^{***}$ | | (88)$^{*}$ | (93)$^{***}$ |

Note: All cells contain means. Standard errors in parentheses and significance stars are from *t*-tests comparing group in column (i.e., middle 5% or top 5%) with bottom 5% group. $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$

[a] In North Carolina, teachers are paid on a state salary schedule based on experience, education, and National Board Certification, but may receive a local pay supplement in addition to the state-determined amount. Thus, the pay supplement is the average amount over and above the salary determined by the state salary schedule.

**Table 2. Count and percent of CSI schools that would have been identified as CSI schools in adjacent years by ESSA index and years of data**

Panel A. Modal index (status quo)

| Years of data | CSI in year *t* | *t*-1 | | *t*+1 | | *t*-1 and *t*+1 | |
|---|---|---|---|---|---|---|---|
| | N | N | % | N | % | N | % |
| Single year (status quo) | 376 | 140 | 37.2 | 148 | 39.4 | 68 | 18.1 |
| 3-year weighted mean | 376 | 241 | 64.1 | 238 | 63.3 | 160 | 42.6 |

Panel B. Higher proficiency index

| Years of data | CSI in year *t* | *t*-1 | | *t*+1 | | *t*-1 and *t*+1 | |
|---|---|---|---|---|---|---|---|
| | N | N | % | N | % | N | % |
| Single year (status quo) | 376 | 166 | 44.2 | 172 | 45.7 | 94 | 25.0 |
| 3-year weighted mean | 376 | 264 | 70.2 | 274 | 72.9 | 192 | 51.1 |

Panel C. Higher growth index

| Years of data | CSI in year *t* | *t*-1 | | *t*+1 | | *t*-1 and *t*+1 | |
|---|---|---|---|---|---|---|---|
| | N | N | % | N | % | N | % |
| Single year (status quo) | 376 | 105 | 27.9 | 123 | 32.7 | 47 | 12.5 |
| 3-year weighted mean | 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).

**Table 3. Predictive validity of ratings by ESSA index and years of data**

| | $t$+1 | $t$+2 | $t$+3 |
|---|---|---|---|
| **Panel A. Modal** | | | |
| *A1. One-year* | 0.700 | 0.620 | 0.560 |
| SE | (0.006) | (0.007) | (0.009) |
| $R^2$ | 0.490 | 0.385 | 0.314 |
| *A2. Three-year* | 0.891 | 0.787 | 0.663 |
| SE | (0.004) | (0.006) | (0.008) |
| $R^2$ | 0.794 | 0.620 | 0.439 |
| **Panel B. High proficiency** | | | |
| *B1. One-year* | 0.859 | 0.799 | 0.752 |
| SE | (0.004) | (0.006) | (0.007) |
| $R^2$ | 0.737 | 0.638 | 0.566 |
| *B2. Three-year* | 0.952 | 0.895 | 0.819 |
| SE | (0.003) | (0.004) | (0.006) |
| $R^2$ | 0.907 | 0.802 | 0.671 |
| **Panel C. High growth** | | | |
| *C1. One-year* | 0.559 | 0.461 | 0.385 |
| SE | (0.007) | (0.008) | (0.009) |
| $R^2$ | 0.313 | 0.212 | 0.149 |
| *C2. Three-year* | 0.832 | 0.679 | 0.503 |
| SE | (0.005) | (0.007) | (0.009) |
| $R^2$ | 0.692 | 0.462 | 0.253 |

Note: Each subpanel (i.e., A1, A2, B1, B2, C1, C2) and column provides results from a separate regression. Within each subpanel, first row provides $\beta_1$ estimate from Equation 1, second row provides standard error on $\beta_1$ estimate, and third row provides model $R^2$.

**Table 4. Summary of tradeoffs in accountability system designs**

*Panel A. Higher proficiency vs. higher growth index*

| | | Higher proficiency | | Higher growth |
|---|---|---|---|---|
| **Validity** | | | | |
| Construct | +<br>– | Stronger for achievement<br>Weaker for school quality | –<br>+ | Weaker for achievement<br>Stronger for school quality |
| Predictive | + | Stronger at predicting future ratings | – | Weaker at predicting future ratings |
| Consequential | +<br>– | Possibility of improving outcomes in most underresourced schools if the system includes sufficient additional resources<br>Sanctions would disproportionately fall on high-needs schools vs. lowest performing schools | –<br>+ | Would not target the most underresourced schools with additional resources<br>Sanctions would be better targeted to low performing schools |
| **Stability** | + | Ratings and CSI designations are more stable from year to year (but not stable enough for high-stakes decisions) | – | Ratings and CSI designations are less stable from year to year (and not stable enough for high-stakes decisions) |
| **Equity** | | | | |
| Disproportionality | – | Systematically assigns lower ratings to underserved schools | – | Systematically assigns lower ratings to underserved schools but to a lesser degree |
| Equity of access | + | Stronger if CSI designations result in narrowed resource gaps | – | Weaker because CSI designations would not identify the most underresourced schools to narrow resource gaps |

*Panel B. Index based on single year of data vs. three-year weighted mean*

| | | One year of data | | Three-year weighted mean |
|---|---|---|---|---|
| **Validity** | | | | |
| Construct | –<br>+ | Measurement error from single year will account for more of rating, but…<br>Rating will represent most recent year's performance, which might be perceived as more valid | +<br>– | Noise is averaged out over three years<br>Including prior years of data may be perceived as less valid as a measure of the current year |
| Predictive | – | Less predictive of future ratings and CSI designations | + | More predictive of future ratings and CSI designations |
| Consequential | – | Potentially weaker, statistical noise will play a larger role in which schools are targeted so the system may be less effective at improving the highest-needs schools | + | Potentially stronger, will target schools with consistently greater need so the system may be more effective at improving the highest-needs schools |
| **Stability** | – | Year-to-year ratings and CSI designations are less stable (and not stable enough for high-stakes decisions) | + | Year-to-year ratings and CSI designations are more stable (but not stable enough for high-stakes decisions) |
| **Equity** | | | | |
| Disproportionality | + | Disproportionality is similar. However, may be perceived as more fair because ratings are based on most recent year of performance | – | Disproportionality is similar. However, may be perceived as less fair because ratings are not based entirely on most recent year of performance |
| Equity of access | – | Potentially weaker if resources are allocated to lower-needs schools due to measurement error | + | Potentially stronger if resources are allocated to the very highest needs schools |

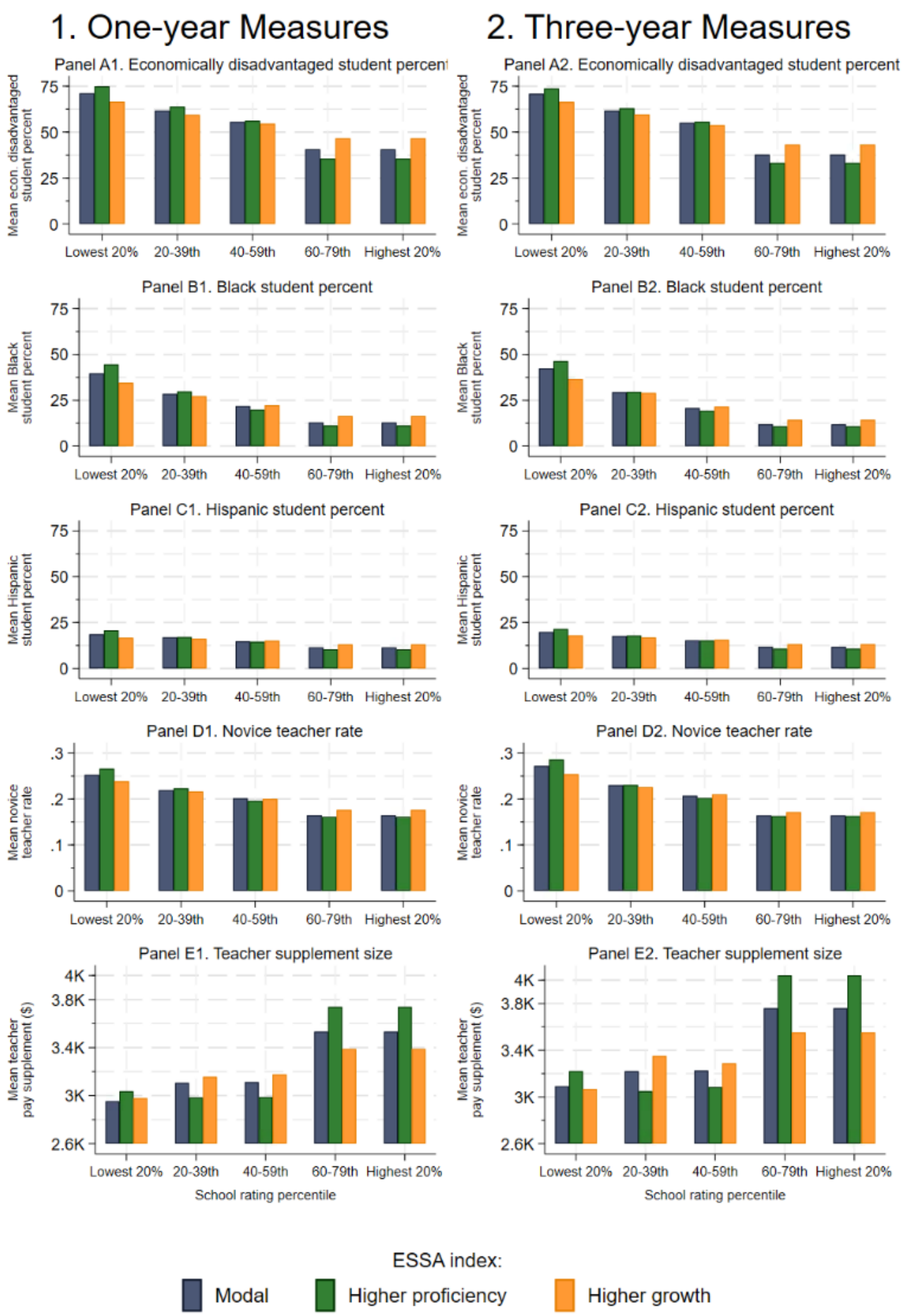
## Appendix

**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 |

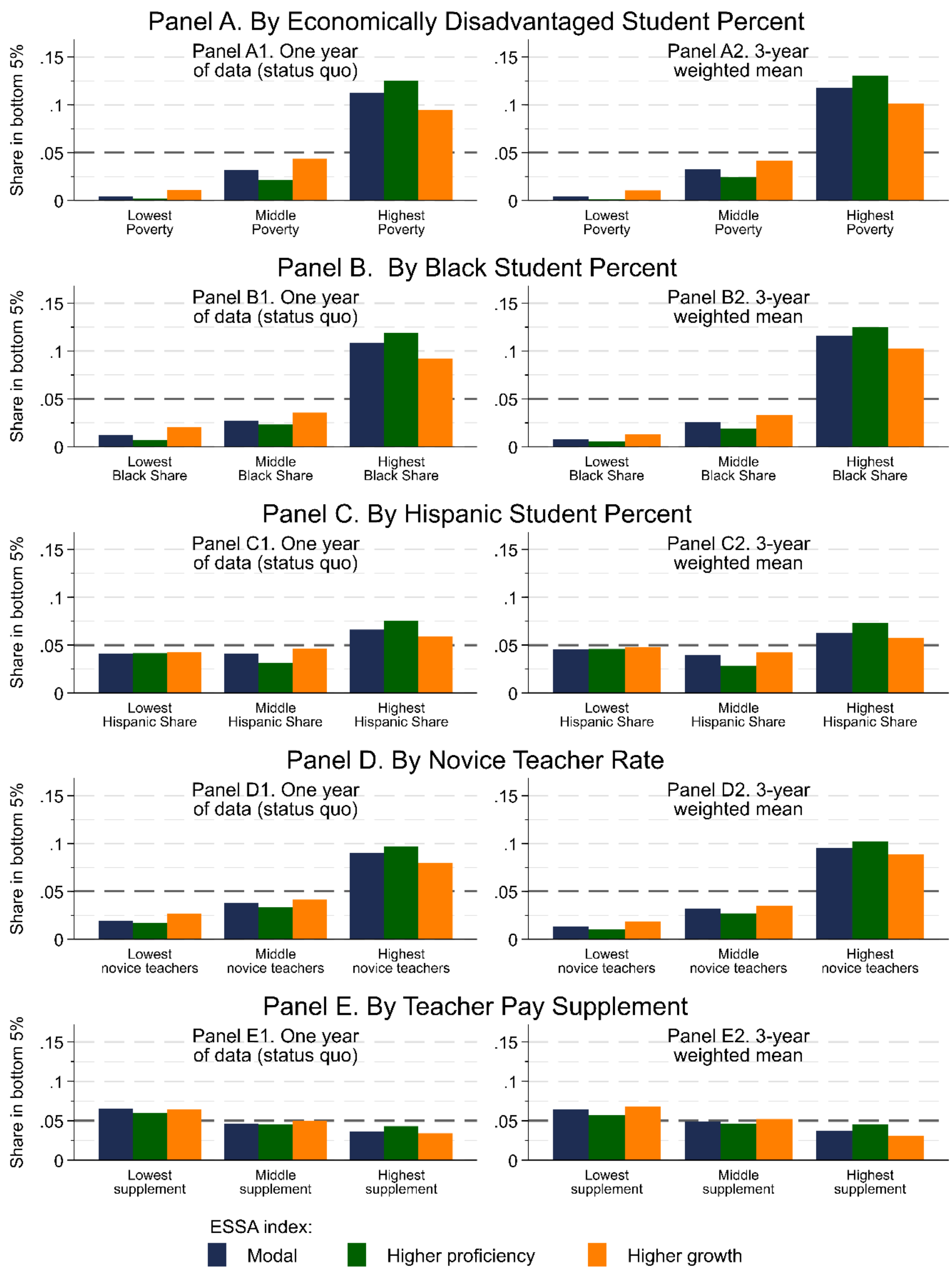
[1] Student growth is calculated from the Education Value-Added Assessment System, or EVAAS.
[2] We do not have access to data from the English language proficiency exam taken by English learners (ELs). As a proxy, we use the percent of students classified as ELs who scored at proficient or above on 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 designated, we reallocate the EL weight proportionally across the other index components.

**Figure A-1. School characteristics by rating quintile and ESSA index, one and three-year indices**




Note: Left panel (Panel 1) replicates Figure 2 using one-year indices for comparison, with right panel (Panel 2) displaying parallel calculations using percentiles based on three-year weighted mean.

**Figure A-2. Share of CSI schools by terciles of school demographics, by school quality index weighting scheme and identification rule**



Note: Graph based on terciles, in parallel to Figure 3 in the main text showing bottom quartile, middle two quartiles, and top quartile. Bar heights represent the share of schools within each grouping of (A) economically disadvantaged student percent, (B) Black student percent, (C) Hispanic student percent, (D) novice teacher rate, and (E) teacher pay supplement size, identified in the bottom 5% based on each ESSA index and number of years of data. Groupings are schools with the lowest 25%, middle 50%, and highest 25% on each school characteristic. "Lowest" groups contain schools with the lowest values on the observed measure and "Highest" groups contain schools with the highest values. For Panels A–D, the low groups represent the *less* disadvantaged schools (i.e., serving fewer students from historically marginalized backgrounds or employing fewer inexperienced teachers) and the high groups represent the *more* disadvantaged schools (i.e., serving more students from historically marginalized backgrounds or employing fewer inexperienced teachers). For Panel E, the low quantile represents the *more* disadvantaged schools (i.e., schools offering the lowest teacher pay) while the high quantile represents the *less* disadvantaged schools (i.e., schools able to offer the highest teacher pay). The horizontal line at 0.05 denotes the 5% threshold—the share of schools that would be identified in each quantile if the subgroups were designated as CSI proportionally.

**Figure A-3. Distribution of One-Year Changes by Index**

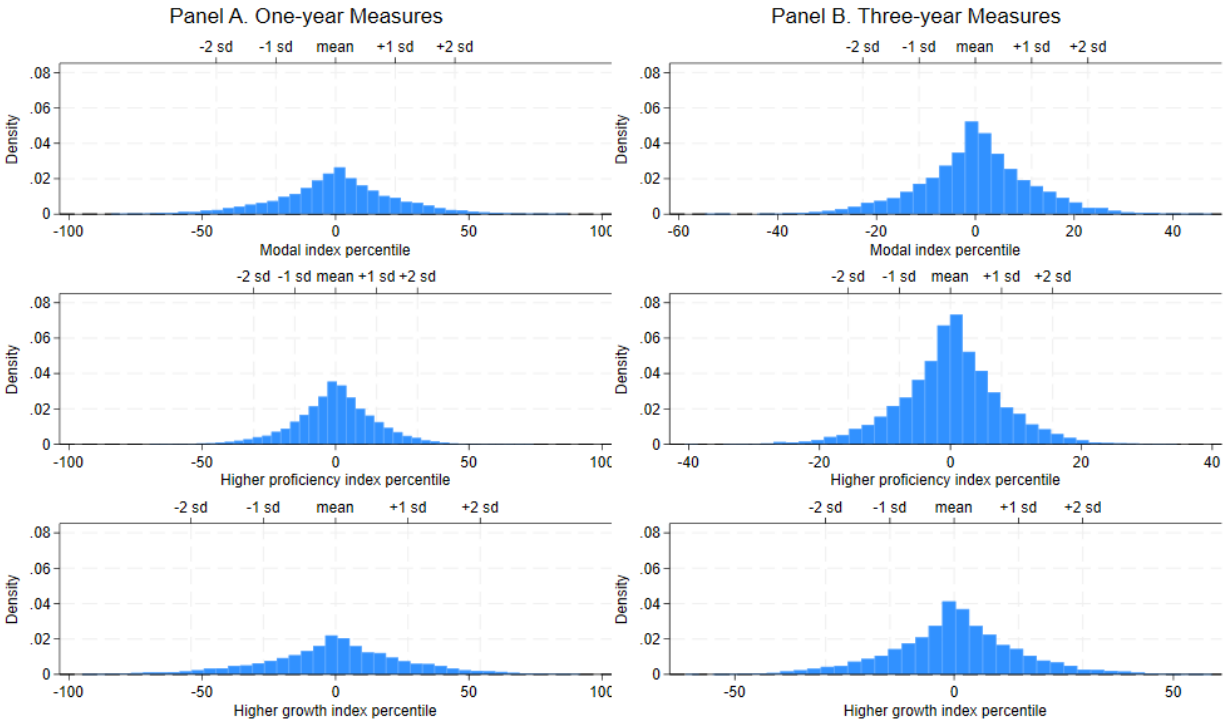
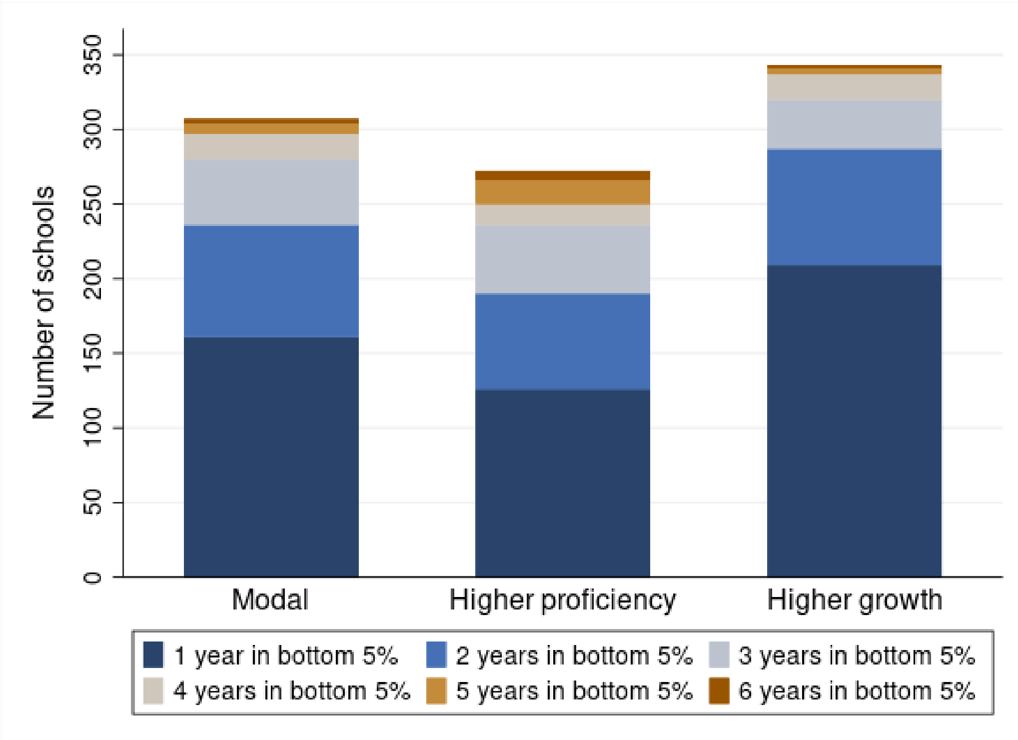


**Figure A-4. Number of years in the bottom 5% by index weighting scheme and identification rule**

*Panel A. Decision rule using single year of data (status quo)*




*Panel B. Decision rule using three-year weighted mean*