16

Test-Based Accountability in K–12 Education

Andrew D. Ho
Harvard Graduate School of Education

Morgan S. Polikoff
University of Southern California

Ho, A. D., & Polikoff, M. S. (2025). Test-based accountability in K–12 education. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1113–1188). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0016

Over the past several decades, K–12 educational tests have served a variety of roles in formal and informal accountability policies at the student, teacher, school, district, state, and federal levels. The use of tests for accountability places a considerable burden on those seeking validity evidence (see Lane & Marion, this volume, for a discussion of validity and validation). While a canonical user of a test score may draw an inference about a single student, accountability systems often incorporate many actors, and many derived scores, at many levels of an educational system. Who is holding whom accountable? By what mechanism? Using which scores? For what purpose?

In the face of such complexity, the measurement field has sometimes treated accountability as a distinct second-stage consideration. A developer may first gather validity evidence for a student score interpretation and then leave those who wish to use scores in an accountability system to gather validity evidence for such uses. This two-stage formulation risks isolating measurement experts and measurement expertise in the first stage, when recent history has shown that the second stage often follows the first. A modern approach to analyzing tests from a measurement perspective is integrated. It anticipates the likely use of scores for accountability and aims to document and prevent the bias, variance, and unintended consequences that such uses can impart.

Our chapter extends the significant contribution by Koretz and Hamilton (2006) in the fourth edition of *Educational Measurement*. The first three editions of *Educational Measurement* had no stand-alone chapter on accountability. Koretz and Hamilton provided an extensive treatment of the validity of inferences under high-stakes conditions. Their treatment, further developed by Koretz (2008), convincingly demonstrates how validity evidence gathered under low-stakes administration conditions is inadequate to support the use of tests for accountability. We extend their framework and principles to the wide range of derived scores common in modern accountability systems.

Koretz and Hamilton (2006) also provided a comprehensive history of large-scale testing in the United States, including the increasing use of assessments for accountability from the 1980s through the initial years following the No Child Left Behind Act of 2001 (NCLB). We summarize and extend their history through three additional discernible national accountability policy periods, the Growth Model Pilot Program (GMPP) of 2005, the "waiver" period beginning in 2011, and the Every Student Succeeds Act of 2015 (ESSA). We emphasize how each period brought accompanying measurement challenges, including the measurement of growth, the measurement of college and career readiness, and score comparability across states and over time.

The COVID-19 pandemic closed U.S. schools in early 2020 and disrupted assessment and accountability systems worldwide. The pandemic coincided with the development of this chapter, and our perspective on pandemic-era accountability is limited. While we treat this recent period simplistically as a pause and return of ESSA accountability policies, we acknowledge that states varied considerably in their responses to the pandemic, and current accountability policies are even more varied across states than they were prior to the pandemic. We look forward to future research that summarizes the history and impact of this period of volatility in educational accountability policies.

Our chapter complements others in this volume. Our focus is on assessment for accountability in the United States, including the low-stakes, but not no-stakes, National Assessment of Educational Progress (NAEP). In contrast, Braun and Kirsch (this volume) focus on large-scale assessments in other countries, including cross-national monitoring assessments like the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). Our focus is on assessment for accountability across multiple levels, including the nation, state, district, school, classroom, teacher, and student levels. Brookhart and DePascale (this volume) cover student-level assessment for purposes like orienting diagnostic feedback and informing pedagogical decisions, and Margolis et al. (this volume) cover assessment for licensing and certification. Our focus is also on school environments in K–12 grade levels. Camara et al. (this volume) cover assessment in the context of higher education. Finally, our focus is on cognitive constructs related to academic content, whereas Kyllonen and Zu (this volume) cover issues related to assessments of social and emotional learning and nonacademic outcomes.

THE LOGIC AND HISTORICAL DEVELOPMENT OF K–12 TEST-BASED ACCOUNTABILITY

Accountability testing uses formal or informal policies and mechanisms to encourage or require actors to use test scores. Five questions can help to distinguish among different approaches to "assessment for accountability": (1) Who is holding whom accountable? (2) By what mechanism? (3) Using which scores? (4) For what purpose? (5) With what unintended negative consequences? Figure 16.1 provides an overview of these questions asked and answered at different levels of K–12 educational systems. We explain the vertical and horizontal dimensions of Figure 16.1 here and return to this framework throughout the chapter.

The vertical dimension distinguishes among levels of the educational system, answering Question 1, "Who is holding whom accountable?" At higher levels of aggregation, national, state, and district leaders hold lower level leaders accountable to educational progress in the interest of the general public. At lower levels of aggregation, schools and teachers use tests to hold students accountable to learning. We reference "general public interest" and "general parent interest" at higher and lower levels of the educational system, respectively, to indicate that justifications for accountability emphasize more benefits to the general public at higher levels and more benefits for parents and their children in specific classrooms and schools at lower levels. These interests also interact, where parents serve in general public roles and can be a strong advocacy group at higher levels of educational systems.

The horizontal dimension of Figure 16.1 contains four columns, each of which addresses the remaining four questions in turn, about accountability mechanisms, score types, purposes, and unintended negative consequences, respectively. The first column lists accountability decisions and mechanisms that test scores serve at different levels

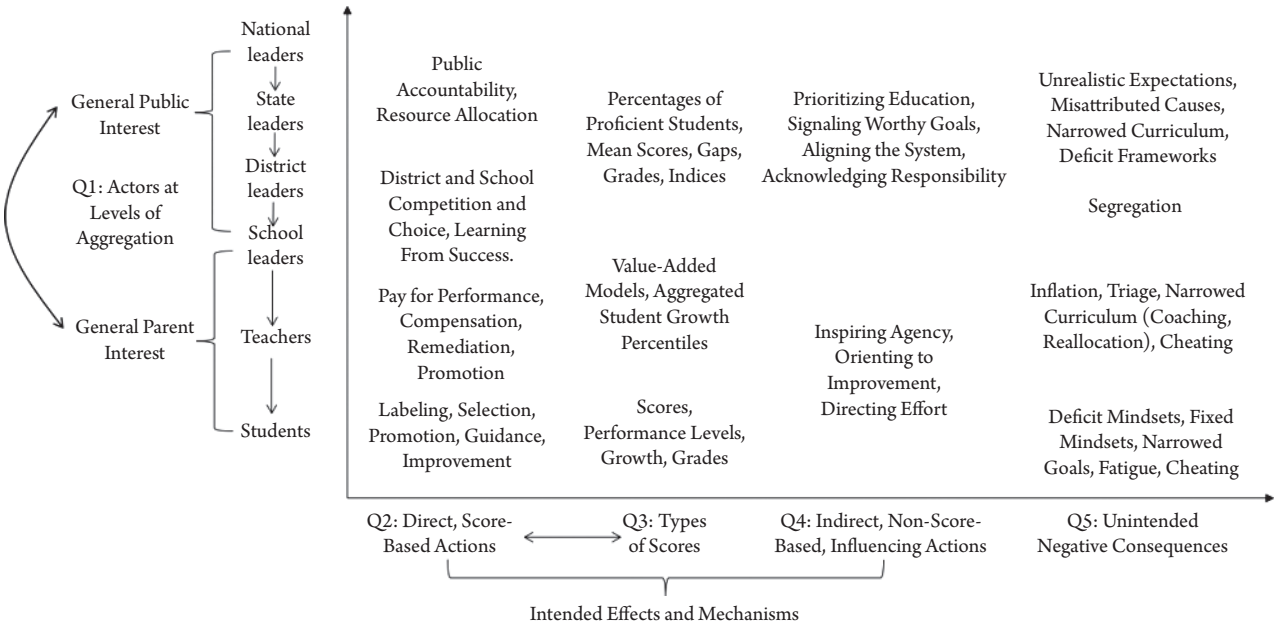

FIGURE 16.1
A Framework for K–12 Test-Based Accountability Purposes and Mechanisms

of the system. For example, at higher levels of aggregation, publicly reported aggregate scores for the nation, states, districts, and schools can act as a form of "public accountability" (Hutt & Polikoff, 2020), where transparency about aggregate scores and progress can theoretically inform the general public and inspire educational improvement. Policies can also establish funding or program enrollment for higher scoring or lower scoring schools, teachers, and students. These policies have higher or lower stakes depending on the consequence of the policy, the individuals or groups it targets, and the relative weight of test scores in the policy decision.

The second column describes the types of scores that usually support these decisions and mechanisms. Emphasizing score types and uses is particularly essential for accountability testing because intended uses are often at different levels of aggregation, or scores are "adjusted" or "combined" in various ways to serve accountability purposes. These "derived scores" have properties that can both serve intended purposes and create unintended consequences (Haertel & Ho, 2016). Chapter 13 of the *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA] et al., 2014) also discusses these issues in the context of "accountability indices."

The third column of Figure 16.1 describes "indirect" or "influencing" purposes of accountability testing that also serve as broad justifications for accountability. Haertel (2013) distinguished these indirect purposes of testing as mechanisms, including incentives and messaging, that influence actors without directly using their test scores. For example, a goal of test-based graduation policies is to direct student and teacher effort, and a goal of school accountability policies is to signal worthy curricular goals

and levels of proficiency. Both intended effects occur as the result of the policy, not as a result of interpretations or uses of any particular scores themselves. Indirect purposes are usually but not necessarily explicit in the logic of test-based accountability policies. Validity evidence supporting indirect test uses includes quasi-experimental analyses that estimate the effect of the test-based accountability policy over a counterfactual policy with less testing or accountability.

The final column of Figure 16.1 lists common unintended negative consequences of test-based accountability policies at different levels of educational systems. These can include common responses like inflation that derive from incomplete or unrealistic goal structures (Koretz, 2008). Additional unintended consequences of accountability testing include "deficit interpretations" of students or groups where users interpret low scores or averages as irreparable or innate student or group deficits for themselves or for others. For example, recent empirical evidence suggests that "achievement gap" framing affects teacher support for policy solutions and general public beliefs about racial stereotypes and causes of disparities (Quinn, 2020; Quinn et al., 2019). Professional testing standards are explicit about the importance of anticipating, documenting, and minimizing unintended negative consequences in test-based accountability (AERA et al., 2014).

Avoiding unintended consequences requires resolving a central tension in test-based accountability: on the one hand, between incentivizing achievement of students, teachers, and school leaders to meet test-based goals and, on the other hand, formalizing the responsibility of school, district, state, and federal actors to provide the capacity and opportunity for this achievement. Elmore (2004) laid out five guiding principles for designing fair accountability policies that attempt to resolve this tension: (a) there should be empirical evidence that goals are achievable, (b) measures should be sufficiently accurate and precise for their purposes, (c) students should be held accountable only to what they have had the opportunity to learn, (d) schools should be held accountable only to the value they add, and (e) the performance that a policy requires should be reciprocated by the responsibility of the system to provide that capacity.

Implementing these principles is challenging in part because test-based measures cannot on their own distinguish between a low score that reflects a lack of individual commitment to achievement and a low score that reflects a lack of systematic support for achievement. A lack of individual commitment might suggest that stakes should be higher, whereas a lack of systematic support suggests that support should be stronger. This fundamental ambiguity in test score interpretation enables actors at different levels of the accountability system to use the same test score data to ask more from each other. Higher level actors use policy and authority to require more of lower level actors. Lower level actors can use mechanisms of public accountability (Hutt & Polikoff, 2020) to require more assistance or support from higher level actors.

These mechanisms are not equal in power. Because of this asymmetry, particular care is required to ensure that the system is beneficent in providing opportunities to accelerate learning without reinforcing existing inequalities and power structures. Toward

the end of this chapter, we review recent movements that attempt to rebalance power toward holding systems accountable to learners, including systems of equity indicators (National Academies of Sciences, Engineering, and Medicine [NASEM], 2019), sociocultural assessment (Shepard, 2021), and indigenous data sovereignty and policy (Walter et al., 2021).

Balance between stakes and supports among actors, mechanisms, scores, and purposes has evolved through the history of U.S. federal educational policy. In the remainder of this section, we review five phases in rough chronological order: the Tyler rationale, program evaluation and monitoring, programmed instruction, minimum competency testing, and the standards movement. Our review draws from similar reviews of history relevant to accountability testing, including those by Haertel and Herman (2005), Koretz and Hamilton (2006), and Madaus et al. (1983). Like Shepard et al. (in press), we also update this history by describing recent waves of federal accountability policy, including NCLB and ESSA.

The Tyler Rationale

The Tyler rationale (Tyler, 1949) forms the basis of many modern models for improving curriculum and instruction. Tyler drew lessons from the "Eight-Year Study" (E. Smith & Tyler, 1942) conducted from 1933 to 1941. The original study investigated whether students in 30 high schools with relaxed college-preparatory course requirements had different outcomes than students in high schools that maintained the strict curricular sequences of that era.

Tyler (1949) distilled four principles from the study: (a) define appropriate objectives, (b) design educational experiences aligned to objectives, (c) organize educational experiences for effectiveness, and (d) evaluate whether objectives have been attained. As Haertel and Herman (2005) noted, Tyler and his colleagues also articulated clearly that tests served multiple purposes beyond evaluation. These purposes included focusing student and teacher effort and providing "a sound basis for public relations" (E. Smith & Tyler, 1942, p. 10) through evidence about outcomes. The use of tests to focus the system and shape public perceptions would grow in the decades to come (Haertel, 2013).

Programmed Instruction and Criterion-Referenced Testing

From the 1950s through the 1970s, Tyler's principles undergirded the rise of programmed instruction, a model that shifted Tyler's emphasis on measurable objectives from the program level to the classroom level. Psychologists at the time envisioned a well-crafted sequence of curricular topics, with tests that guided and sometimes determined subsequent instruction. Propelled by behaviorist theories (Skinner, 1953), these systems included tests to check understanding and provide immediate feedback to reinforce desired behaviors.

To support these uses, tests had to support inferences about what students could do. Glaser (1963) coined the term *criterion referenced* to describe tests designed to assess

"the degree to which the student has attained criterion performance" (p. 6). This set up a useful contrast to *norm-referenced tests* that have a primary purpose of differentiating among test takers relatively, often for the purpose of selection into competitive programs. Although well-designed tests can support scores that have criterion-referenced and norm-referenced interpretations (AERA et al., 2014), essentially all modern accountability tests claim to be criterion referenced to emphasize that the primary scores of interest indicate what students know and can do against an absolute standard. The term also suggests the appealing theoretical possibility that all test takers can achieve a given criterion level.

To support or operationalize decisions in programmed instruction, practitioners and systems use established cut scores or set them on the basis of what students know and can do. Bloom's popular "mastery learning" model (1968), for example, included end-of-unit tests with cut scores that distinguish between students who are and those who are not ready for the next unit. Cut scores on criterion-referenced tests should enable teachers to describe a student who scores at or above a cut score in terms of their absolute knowledge and skills, without reference to any norm or reference group.

Tests in this era became much more focused on fine-grained objectives that developed along continua. Score reporting turned to support interpretations of specific criteria, knowledge, skills, and abilities. Programmed instruction also raised the stakes on potential instructional decisions for students, including remediation and repetition of previous instruction. And the use of tests to reform and improve instruction, both directly by providing data and informing decisions and indirectly by focusing the system, continued to rise.

Testing for Monitoring and Program Evaluation

As programmed instruction was increasing the role of testing in classroom instruction, the 1960s saw an increase in the role of testing in federal programs and surveys. The Elementary and Secondary Education Act of 1965 (ESEA) included a required, annual, measurement-based evaluation of Title I programs under the Title I Reporting and Evaluation System. The National Science Foundation began to require the use of educational tests in evaluations of funded curricula. And the headline findings from the Equality of Educational Opportunity Survey (Coleman et al., 1966) were drawn from standardized achievement tests. Extending the Tyler rationale, educational testing began to be a default requirement for federal programs, a signal of rigor, transparency, and public accountability.

Planning for NAEP also began in the 1960s, and the first NAEP assessments were administered in 1969–1970 (Jones & Olkin, 2004). Early NAEP testing did not report scale scores and was instead inspired by the same objectives-oriented theories that guided programmed instruction. In a 2004 reflection, Cronbach compared the reporting goals to those of opinion polling, where results consisted of numerous percentages of correct answers to specific questions, disaggregated by groups. He also notes, "There was interest in making the results comprehensible to teachers in terms that they could

put to use in their teaching" (Cronbach, 2004, p. 141). However, the reports were unwieldy and made it challenging to measure progress over time. We reserve a fuller review of NAEP and its significant and growing role in educational accountability for the section "NAEP and a Theory of Public Accountability". Here, we observe the growing pains of early NAEP as an example of a continuing tension between designing diagnostic tests for instructional guidance and designing monitoring tests for large-scale educational progress.

Minimum Competency Testing

In the 1970s, concern about youth unemployment, low graduation standards, and intractable achievement gaps helped to spur a "back to basics" movement (Haertel & Herman, 2005). In response, district and state policies known as *minimum competency testing* began to spread, typically taking the form of making high school graduation and, in some cases, grade promotion contingent on students scoring above selected score thresholds on standardized tests. Statewide minimum competency testing requirements existed in at least 29 states by the end of the decade.

Minimum competency testing built on the theoretical foundations of educational psychologists from previous decades, including the use of tests that measured specific competencies and the use of a criterion-referenced cut score for "minimum competency." However, a dramatic extension from previous decades was the use of large-scale educational policy to articulate a clear and high-stakes decision rule, in this case with direct consequences for students. The rhetoric of the movement has largely faded in favor of an emphasis on high standards and higher-order thinking skills. Nonetheless, the movement helped to advance a more substantial state role in educational testing, as well as the use of criterion-referenced tests and cut scores for high-stakes decisions (Koretz & Hamilton, 2006).

With low cut scores and numerous opportunities to retake exams, Catterall (1989) summarized the scholarly consensus that "the competency test was legislated as a lion but implemented as a lamb" (p. 4). As we review federal and state accountability policies for schools and students over the next decades, this pattern recurs: initially unrealistic political rhetoric and goals are followed predictably by increasing flexibility.

The Birth of the Standards Movement

The 1980s began with a more centralized and data-driven effort than the previous decade to signal a crisis that required national educational reform. The influential 1983 report *A Nation at Risk* (National Commission on Excellence in Education, 1983) used a range of indicators, including international comparisons from the International Association for the Evaluation of Educational Achievement (Bloom, 1973), results from the SAT (formerly Scholastic Aptitude Test), and NAEP results to argue that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people" (National Commission on Excellence in Education, 1983, p. 5). The report called out the shortcomings of

minimum competency tests specifically, arguing that "the 'minimum' tends to become the 'maximum' thus lowering educational standards for all" (p. 20). The report instead called for "more rigorous and measurable standards, and higher expectations, for academic performance" (p. 27). The report specified that tests "should be administered at major transition points from one level of schooling to another" and "should include other diagnostic procedures that assist teachers and students to evaluate student progress" (p. 28).

The coming years continued to clarify the logic of modern large-scale accountability testing: aligned tests, higher standards, and transparent reporting helped to focus and improve teaching and learning (Resnick & Resnick, 1992; Smith & O'Day, 1991). The state and federal roles in accountability testing also continued to rise. Responding to requests from some governors in 1984, Secretary of Education Terrel Bell prepared a one-page "wall chart" of state educational outcomes, including graduation rates and SAT and ACT (formerly American College Testing) scores. Once planned for release in a small press conference in the secretary's office, the wall chart resulted in the largest press conference in the Department of Education's history (Ginsburg et al., 1988). Later that year, the Council of Chief State School Officers passed by a vote of 27 to 12 a report supporting improved measurement of educational indicators across states, with a closer 20–19 vote that avoided the deferral of a key section of the report about comparable outcomes (Toch, 1984). The report noted that, "on a technical level, state-to-state comparisons are problematical for a variety of reasons . . . on a political level, however, the attention given to the Secretary's 'wall chart' makes inevitable future state to state comparisons on outcome measures" (Council of Chief State School Officers, 1984, p. 3). This articulates a theory of public accountability that we develop in the next section, where transparent reporting of comparable outcomes signals worthy goals, creates competition, and draws public attention, even without explicit consequences or rewards for scores.

In addition to higher performance standards as exemplified by "proficiency," reports from this era emphasized complex, higher-order thinking skills and attempted to distinguish clearly between performance standards and content standards. Content standards referred to "the knowledge, skills, and other understandings that schools should teach in order for students to attain high levels of competency in challenging subject matter," and performance standards "define various levels of competence in the challenging subject matters set out in the standards" (National Council on Education Standards and Testing, 1992, p. 13). The National Council on Education Standards and Testing was convened to explore the possibility of a national system of standards and tests, with candidate sets of content standards like those by the National Council of Teachers of Mathematics in 1989 and the American Association for the Advancement of Science in 1993.

An ambitious effort known as the New Standards Project was a joint effort of the University of Pittsburgh's Learning Research and Development Center and the National Center on Education and the Economy. The project attempted to operationalize higher

order content standards in the form of performance assessments. The project had 17 states and 6 urban districts signed on at its peak. It ultimately faced a number of challenges, including state resistance to standards that were not "home grown" and apparent underestimation of the amount of time and money it takes to produce performance assessments that support accountability decisions at scale (Hoff, 2001). Koretz et al. (1994) summarized similar tensions among affordability, score precision, and instructional relevance in Vermont's portfolio assessment system.

The 1990s also saw the rise of explicit school-level stakes. Whereas minimum competency testing had direct consequences on students, the logic of standards-based reform focused on incentives at the school level. The popular Quality Counts series published by *Education Week* reviewed state educational reform plans annually, including a grade for states on multiple indicators related to standards-based reform. By 1999, the report found that 36 states had committed to transparent reporting of report cards for schools, with 19 issuing overall performance-based ratings (Jerald, 2000). In addition, 16 states had the legal authority to reconstitute schools that failed performance measures, and 13 states offered monetary rewards to successful schools. These state systems established the precedent for the explicit school sanctions that would be required of all states by NCLB in the years to come.

The NCLB Era

The NCLB Act, passed in 2001, was a bipartisan triumph of the standards movement in education. NCLB represented a reaction to federal discontent over the progress of school reform throughout the 1990s, most specifically the slow and weak implementation of the prior ESEA reauthorization, the Improving America's Schools Act (McDonnell, 2005). NCLB also grew out of state-level reforms, primarily in Texas, that seemed to be boosting student performance (Deming et al., 2016; Grissmer & Flanagan, 1998).

The NCLB suite of reforms was premised on a straightforward theory of change (Porter & Polikoff, 2007). First, establish clear, grade-level content standards in the core academic subjects (mathematics and English language arts, though science was also included in the law). Second, create aligned assessments to measure student performance against those standards and inform the public and educators of that performance. Third, through accountability policies, hold schools and districts accountable for student performance against the standards and offer interventions in schools where the standards are not met. Fourth, disaggregate all results for key student groups of interest to ensure high average performance does not mask low-performing student groups. With these policies in place, the thinking went, teachers (supported by school and district leaders) would be incentivized to align their instruction with the challenging academic content standards, student opportunity to learn would improve, and achievement would rise. In schools where achievement was not rising, public school choice (enabled by the public release of school performance data) would put additional market pressure on schools to improve, while supplemental education services would provide additional support to students and teachers. The design of the accountability

systems broadly aligned with this theory of change, though we discuss the effectiveness of these policies in the section "The Impact of Accountability Policies".

School accountability under NCLB was based on a straightforward set of measures: student proficiency rates on state tests, student participation rates on state tests, and an additional state-selected indicator (high school graduation rates and, typically, attendance rates for elementary and middle schools). For student proficiency, states were required to establish proficiency rate trajectories for each tested subject and grade. These trajectories began from baseline proficiency rates set in 2002 based on the distribution of student performance in the state to the target of 100% proficiency by 2014. Each year's target was called an annual measurable objective (AMO), and meeting all AMOs was necessary to make adequate yearly progress (AYP) and avoid consequences. In other words, the NCLB accountability system represented a conjunctive approach to combining multiple measures—schools had to cross each AMO bar overall, and for all student groups—high performance in one area could not offset low performance in another.

The consequences for failing to meet AYP escalated with the number of years of failure, beginning with labeling, then public school choice, then supplemental tutoring, and finally the possibility of state intervention (including school closure, charter conversation, state takeover, or other significant turnaround efforts). States were allowed to set different trajectories for different grade levels and subject areas, but they were required to have the same targets for all student groups (NCLB required states to disaggregate data for, at minimum, student groups based on race/ethnicity, socioeconomic status, disability status, and English proficiency). To ensure that schools were not selectively excluding low-performing students from taking tests, the law also required 95% of each student group to participate in each exam.

While these test-based performance measures seem straightforward and comparable across states, in practice, state decisions about the construction of their systems dramatically affected how many and what kinds of schools would be subject to accountability pressure. These decisions are described well elsewhere (Davidson et al., 2015; Polikoff & Wrabel, 2013; Porter et al., 2005), but we briefly summarize some of the relevant key decision points here.

One decision point states were allowed to make under NCLB was about student group size. This was the minimum number of students required in a particular student group for that student group to count toward accountability. The modal subgroup size under NCLB was 30 (Polikoff & Wrabel, 2013), meaning that if a school had 29 low-income students, that student group would not count as its own group for accountability purposes, whereas if it had 30 or more, it would. (The students would count in the overall school average regardless.) However, minimum group sizes ranged considerably across states (from a low of 5 to a high of 100; Davidson et al., 2015). Smaller student group sizes can dramatically affect the number of students who are included in student group accountability measures. An analysis of California data (Polikoff et al., 2018), for instance, found that moving from their old group size of 100 to a group size of 30

increased the average number of accountable student groups in schools from 3.65 to 5.44. This had the effect of dramatically reducing the number of students who were in groups not reported or used for accountability (for instance, under the previous system, 80% of multiple-race students were in schools where that group was not large enough to be reported, but under the new system it was just 28%). Of course, there are trade-offs of changes in student group size rules: With smaller minimum group sizes, reliability will be lower (year-to-year fluctuations will be higher), but there will be greater incentive for educators to attend to student group performance if the school is accountable for more student groups.

Second, as described by Polikoff and Wrabel (2013), the law allowed states to use a variety of alternative methods in determining whether schools' and districts' proficiency rates were above the performance target. By far the best known of these was the Safe Harbor provision, where schools were given credit for meeting a proficiency target if the percentage of students scoring below *Proficient* decreased by 10% (not 10 percentage points) from the previous year. These alternative methods had the effect of sometimes dramatically lowering the proficiency threshold. For instance, consider a school with 100 students, of whom 50 are *Proficient* (50% proficiency rate) against an AYP target of 75%. This school would need to get just 5 more students over the proficiency threshold in the next year to meet the AYP target, even though the proficiency rate of 55% would fall far below the actual target for that year.

Third, as described in detail by Davidson et al. (2015), states also could use statistical adjustments to their accountability results in making accountability decisions. Specifically, states were permitted to adjust targets based on confidence intervals, in essence allowing states to say, "we will only hold your school accountable if we are extremely confident your school's performance is below the proficiency threshold." Confidence intervals were defined in terms of sampling variance of proportions, but were applied in an unusual way. Instead of indicating the precision of the parameter estimate, they were used to lower targets in proportion to imprecision. Importantly, confidence intervals and Safe Harbor could be layered on top of one another. In the previous example of the 100-student school in California, it would actually only need to get 2 more students over the proficiency line, because California applied a 75% confidence interval to the 5-student Safe Harbor target.

Fourth, states had leeway over a variety of technical decisions about which students would count for accountability and how to aggregate their scores. For example, states applied variable rules around the inclusion of students with certain kinds of disabilities and varied in their selection of student racial/ethnic groups (Davidson et al., 2015). States could set their own rules about student enrollment, with some states requiring students to be enrolled in a school for an entire year to count for accountability purposes and others requiring just a few months' enrollment (the former approach excluding highly mobile students). And states could decide how to aggregate results across grades and subjects, whether by first aggregating results and then comparing these to AYP targets or by comparing results to targets separately by grade and subject.

Finally, states were allowed to set nonlinear performance trajectories from the 2002 starting point to the 2014 100% proficiency target (Porter et al., 2005). Some states set "straight-line" targets, expecting a constant annual increase in the percentage of students scoring *Proficient*. Other states set more back-loaded targets, with smaller annual increases in the earlier years of the law and larger increases in the later years.

These state policy decisions—some of them seemingly modest—had profound effects on the number and type of schools identified as failing to meet NCLB proficiency targets. For instance:

- Porter et al. (2005) found early in NCLB that Kentucky's design decisions (using a large student group size of 60 and confidence intervals) resulted in 75% fewer schools failing to meet their accountability targets than had the state made more stringent decisions.
- Polikoff and Wrabel (2013) found that two thirds of California schools used Safe Harbor to meet at least one AYP AMO in 2011.
- Davidson et al. (2015) documented the complex and interacting effects of state accountability policy decisions on failure rates, suggesting that school failure under NCLB was as much about minor differences in the ways state policy makers interpreted and implemented the law as it was about actual school performance.

As the NCLB era drew toward its conclusion, there was widespread dissatisfaction with NCLB accountability systems, with specific focus on (a) the unreasonableness of 100% proficiency targets (though again, in practice there were many ways to make AYP without hitting 100% proficiency), (b) the use of status measures of performance without considering growth, (c) the narrowing effects of the law resulting from its near-exclusive focus on English language arts (ELA) and mathematics test scores, and (d) the perceived ineffectiveness of NCLB's mandated interventions.

The ESEA Waivers

It was in this context that in 2011 (already 4 years after the original intended ESEA reauthorization), the U.S. Department of Education began offering states flexibility waivers from some provisions of NCLB's accountability rules. The waiver accountability policies were broadly aligned with NCLB policies and the NCLB theory of change, but offered some technical fixes aimed to address the shortcomings of NCLB accountability systems.

In place of NCLB's AYP rules, the waivers had several requirements for state accountability systems. First, state AMO targets had to be revised, with the waivers offering states two options: AMOs that increase in equal annual increments to 100% proficiency by 2019–2020 or reduce by half the percentage of *Below Proficient* students in the "all students" group and in each student group within 6 years. These tweaks were offered to reduce the pressure of the impending 100% proficiency deadline, which was then just a couple years off. Second, states were required to include a measure of student growth

in their accountability rules to address the widely known concerns about student proficiency rates as a measure of school performance. Third, states were mandated to identify at least 15% of schools for intervention, with 10% of schools identified as "focus" schools (intended to be schools contributing to state achievement gaps) and 5% of schools identified as "priority" schools (intended to be the state's lowest-performing schools). This provision offered states some relief in terms of the number of schools and districts they would have to identify and work with, but it also introduced a norm-referenced approach into what was originally intended to be a criterion-referenced policy system based on schools meeting targets. Though the waiver accountability systems were in place for only a few years, there are generalizable lessons in the revisions about how to set realistic targets and establish consistent definitions to avoid perverse incentives.

Waiver accountability systems represented important improvements over NCLB systems in some ways, but less so in others. Polikoff et al. (2014) reviewed these improvements and missed opportunities, which we summarize here. First, waiver accountability systems offered flexibility to include measures beyond test scores, responding to earlier critiques of NCLB as narrowly focusing on ELA and mathematics proficiency at the expense of all other measures of school effectiveness. States used this flexibility to include a variety of measures like graduation rates, attendance, school climate, and opportunity-to-learn measures. Only five states used just test scores in identification of their priority schools, and most states used at least one nontest measure in their identification of focus schools. Waiver plans also often included additional tested subjects in accountability (e.g., science, which was required to be tested but was not used for accountability under NCLB), further broadening the measures against which schools were evaluated.

Second, when waiver accountability systems did rely on test scores, they included measures that experts argued were superior to proficiency rates in terms of measuring and incentivizing school effectiveness (e.g., Ho, 2008; Linn, 2003; Polikoff & Wrabel, 2013; Ryan, 2004). About half of states used a measure of individual student achievement growth, and about a quarter used different status-based measures of performance, such as performance indices where students scoring *Basic* earned half a point and those scoring *Proficient* earned a full point (Polikoff et al., 2014). Third, waiver plans made more stable classifications—identifying schools and working with them for at least a few years before moving on to a new set of schools. This has two potential benefits: It reduces the extent to which ephemeral fluctuations in performance drive school classifications and it ensures more sustained intervention and support in struggling schools.

However, researchers also raised concerns about the design of waiver accountability systems. These concerns were along a number of dimensions. For example, in some states the state would use one approach to rate schools on, say, an A–F composite index, but then use a different approach altogether to identify schools for intervention. This created a lack of consistency that might cause unnecessary confusion. As another example, while nontest measures were included in many states' accountability systems,

proficiency rates based on state tests still dominated accountability classification rules in nearly all states, raising all of the same issues as discussed in the NCLB era. And finally, some were concerned about the use of so-called super-subgroups, which combined the performance of multiple groups of students, rather than fully disaggregating results as under NCLB (e.g., R. A. Hernández, 2013). There was concern that this approach to aggregating student groups could mask the performance of the disaggregated groups and therefore diminish attention to supporting these student groups.

Overall, the waiver period represented a controversial but important transition from NCLB-style accountability to ESSA accountability. The waivers clearly broadened the measures against which schools were evaluated, a transition from the nearly exclusive test-based focus of NCLB. The waivers narrowed attention to a small proportion of the lowest-performing schools, moving away from NCLB's extraordinarily high school failure rates and toward a more norm-referenced approach to school accountability. The waivers encouraged innovation in the use of student achievement test scores, including a move away from proficiency-based status measures and toward true student growth measures. These innovations carried over when ESEA was finally reauthorized as the ESSA in 2015.

The Every Student Succeeds Act

The ESSA passed in 2015, finally reauthorizing ESEA and updating requirements and expectations with regard to state accountability systems. The act's accountability provisions continued and expanded on some of the main innovations of the waiver era and continued the rollback of NCLB-era sanctions. The proportion of schools subject to accountability sanctions was reduced to just the bottom 5% (plus high schools with low graduation rates), and the sanctions were made much less prescriptive (with a focus instead on district creation and implementation of evidence-based plans for school improvement).

The statute describes five types of indicators that must be included in school accountability systems (Marion, 2016): first, an indicator of academic achievement; this can be as simple as percent *Proficient*, but ESSA also allows for states to use other status measures of performance (e.g., average scale scores, index-based approaches); second, another "valid and reliable academic measure,"[1] often a measure of either growth or achievement gaps; third, graduation rates, including extended (i.e., 5- or 6-year) graduation rates; fourth, progress toward English language proficiency for English learners; and fifth, another indicator of school quality or success that meaningfully differentiates and is valid, reliable, and comparable (in the text of the law, they list as options measures of student engagement, educator engagement, student access to and completion of advanced coursework, postsecondary readiness, and school climate and safety). In terms of actual accountability classifications, the statute also says that states must measure school performance against state-determined status and improvement goals in at least the areas of academic achievement, graduation rate, and student groups that are behind (Marion, 2016). Drawing on the waivers, the law requires that states produce

at least every 3 years a list of schools for comprehensive support and improvement, including (a) the lowest 5% of Title I schools, (b) high schools with graduation rates below 67%, and (c) schools with low-performing student groups (thus including both norm- and criterion-referenced approaches to school identification).

How did states implement these requirements? The approved state ESSA plans varied along every dimension, so characterizing them concisely is difficult. Indeed, if there is one way to characterize state ESSA accountability systems, it is that they are incredibly variable in form and content. But a few trends emerged in the many external reviews that were produced.

With regard to student achievement measures, many (but not a majority of) states use status measures of student performance that move beyond percent *Proficient* (e.g., average scale scores or "indices" that use a fuller range of student scores rather than a binary *Proficient*/*Nonproficient* distinction) and many also include achievement in subjects other than mathematics and ELA as another indicator of school quality or student success (English, 2017). Virtually all states (47 + Washington, DC; Education Commission of the States, 2018) include measures of student achievement growth in their ESSA systems, a sharp departure from waiver policies (American Institutes for Research, 2017; English, 2017). States also include test participation in their achievement measures, most commonly by assigning a score of zero to nonparticipants on state tests (English, 2017). The story for achievement measures at the high school level is even more variable from state to state; see Petrilli et al. (2016a, 2016b).

State goals and approaches also vary considerably with regard to graduation rate and other high school measures. State 4-year graduation rate goals range from a high of 100% (South Dakota) to a low of 84% (Nevada, Virginia), with the most common levels being 90% or 95% (21 and 10 states, respectively; Hackmann et al., 2019). States also include a wide variety of "college- and career-readiness" (CCR) measures, mostly at the high school level, but sometimes at lower levels. The most common of these are Advanced Placement/ACT/SAT/International Baccalaureate exam scores, industry-recognized credentials, Career and Technical Education (CTE) course sequence completion, advanced diploma completion, and other measures of course taking (for a full list, see Hackmann et al., 2019). In most but not all states, these measures are reported overall and separately for student groups. In many states, these diverse measures are aggregated into a single index (e.g., in Arizona, students can earn fractions of a CCR point for various accomplishments, with those student scores averaged across the state). In most states, the aggregated CCR measure is then combined with other measures to create an overall summative rating.

Another important trend is with regard to the "other academic indicator." While 33 states include the CCR measure as part of their other indicators (Education Commission of the States, 2018), many alternative accountability measures are also included, representing a continued growth in the number and variety of nontest measures that began under the waivers. For instance, the majority of states include chronic absenteeism as a measure (English, 2017), which is important because absenteeism is highly predictive of other long-term student outcomes (Ansari & Pianta, 2019; Liu et al.,

2021). Other measures in this area include conditions of learning or school climate, some measure of the well roundedness of students' education, and credit accumulation.

Student groups continue to play an important role in accountability systems, but perhaps less so than under NCLB. Although ESSA requires states to continue to report student group performance, it does not require a particular approach to the use of student group performance for accountability purposes. Some states under ESSA are using a variant on super-subgroups that began under the waivers, most often simply a group defined by the lowest-performing students (English, 2017). But other states combine low-performing student groups into a demographically defined student subgroup. Furthermore, state minimum subgroup size rules continue to decrease under ESSA, with a median and modal value of 20 (Alliance for Excellent Education, 2018).

Finally, a large majority of states combine multiple measures in some way to produce a summative rating of some kind. Thirty-six states use an A–F grade, a descriptive rating system (e.g., *Needs Improvement*, *Average*, *Good*, *Great*, *Excellent*), or a numerical index system (e.g., 1–10, 1–100; Education Commission of the States, 2018). These indices vary along every conceivable dimension in terms of how they aggregate and weight their various component measures.

While most reviews of ESSA accountability plans view those plans positively, there are some concerns that are commonly raised. The most common critiques have to do with the treatment of student groups. While NCLB was critiqued by some for its overly harsh attention to student groups (its conjunctive approach meant that a single group failing a single AMO would lead to the whole school failing), ESSA may have swung the pendulum far in the opposite direction. Independent reviews suggest that most state plans give almost no attention to the performance of student groups, with school ratings being determined on whole-school performance (Bellwether Education Partners, 2017). Advocates for students with disabilities have expressed similar concerns, worrying that these students will fall through the cracks of the 33 state accountability systems that do not explicitly take student group performance into account in rating schools (National Center for Learning Disabilities, 2018). Another critique is that states' composite indices may appear transparent on the surface, but actually are overly complicated in ways that make their results confusing to educators and the general public (Bellwether Education Partners, 2017). There are also continued concerns about high school accountability, which is heavily reliant on graduation rates (a status-based measure of performance) and lacks the growth-based approach used when student-level longitudinal test score data are available. It is also not fully clear in many state plans how states actually plan to intervene in and improve their lowest-performing schools (though the literature does not exactly offer a road map for school improvement at scale).

The overall takeaway from ESSA plans can be summarized as follows. These plans expand on some of the trends that began under the waivers. In most cases, they dramatically increase the types and number of measures to be used in accountability, but proficiency rates in mathematics and ELA still play a major role and may drive educators' actions. They more appropriately measure student achievement status and

almost universally include measures of student achievement growth as well. However, they take some of the pressure off for school improvement (which is by design and a reaction to NCLB), substantially decreasing the number of schools that are likely to feel meaningful accountability pressure. The changes to student group reporting and accountability requirements may reduce awareness about the average achievement and progress of historically underserved students when these populations are small.

There are two other ESSA-era trends that are worth mentioning here. The first concerns changes in the high school summative assessment. As of 2019, half of the states were including either or both the SAT and the ACT as a high school summative measure (Olson, 2019). This change was allowed under the law's provisions for "a nationally recognized college entrance exam," and states were enthusiastic about the opportunity to adopt assessments that could serve multiple purposes. There are continued concerns about the alignment of the tests to state standards (e.g., National Council on Measurement in Education, 2019), but the Department of Education approved the use of the tests for statewide summative purposes regardless.

Second, as ESSA accountability systems were being implemented, the COVID-19 pandemic gripped America's schools in the 2019–2020 academic year. There were important implications of the pandemic on assessment and accountability. For instance, the Department of Education authorized states to extend the testing window, shorten assessments, and give assessments remotely (U.S. Department of Education, 2021). The Department of Education also allowed states to waive the 95% participation rate rules and to apply for waivers to the requirements for school identification. The Department of Education has been returning to ESSA accountability rules as the pandemic abates.

NAEP AND A THEORY OF PUBLIC ACCOUNTABILITY

Public accountability includes policies and commitments for reporting test data without predetermined stakes attached to results (Hutt & Polikoff, 2020). As Figure 16.1 indicates, proponents of public accountability assert or assume that publishing test score data can signal worthy goals, improve commitment to those goals, spur competition toward those goals, encourage actors to learn from organizations that are successful at achieving those goals, and align the system toward those goals. This logically requires appropriate measures that communicate effectively to intended audiences. The reasoning also requires that actors are motivated by public transparency without any automatic sanctions or rewards attached to results.

This section reviews the rise and species of "report cards" that represent the linchpin of test-based public accountability. We begin with the assessment known as the Nation's Report Card, NAEP (https://www.nationsreportcard.gov/), which was introduced briefly in the previous section. NAEP remains the foremost national and cross-state example of test-based public accountability. This section continues with a discussion of audiences and effects of public reporting.

The Rise of NAEP as a Mechanism for Public Accountability

NAEP had its origins in the federal Office of Education under the leadership of Francis Keppel in 1969. The earliest NAEP assessments that year tested citizenship, science, and writing skills of 9- 13-, and 17-year-olds, debuting technical features like matrix sampling. Over subsequent years, new innovations were included, such as comprehensive content frameworks, state comparisons, and, later, large urban district comparisons. In 1986, Secretary Bennett formed a study group chaired by Tennessee governor Lamar Alexander and Spencer Foundation president H. Thomas James to explore state comparisons using NAEP (Alexander et al., 1987). The resulting report compared NAEP to a weather map that only provided weather for the nation as a whole. The report described state-to-state comparisons as "the single most important change recommended by the Study Group" (p. 11). This recommendation was fiercely debated and ultimately included in the 1988 amendments of ESEA as a trial state assessment program beginning in 1990 that would be voluntary for states (Vinovskis, 2001). Later, NCLB required states to participate in NAEP biennially as a condition of receiving Title I funds.

NAEP also illustrated the value of compelling reporting metrics for public accountability by adopting the criterion-referenced score metric, "percent *Proficient*." The 1987 Alexander–James report included a published commentary by a National Academy of Education review committee chaired by Robert Glaser, the same scholar who coined the term *criterion-referenced testing* 24 years prior. Although the Alexander–James report did not mention criterion-referenced score reporting, the commentary does:

> We recommend that, to the maximal extent technically feasible, NAEP use descriptive classifications as its principal reporting scheme in future assessments. For each content area NAEP should articulate clear descriptions of performance levels, descriptions that might be analogous to such craft rankings as novice, journeyman, highly competent, and expert. Descriptions of this kind would be extremely useful to educators, parents, legislators, and an informed public. (Alexander et al., 1987, p. 58)

The subsequent 1988 legislation created the independent NAGB to oversee NAEP policy and assigned to NAGB the responsibility of "identifying appropriate achievement goals for each age and grade in each subject area to be tested" (PL 100-297, Part C, Section 3403(6)A, 1988). Shortly thereafter, in 1989, President George H. W. Bush convened all 50 governors in an education summit to achieve consensus on broad educational goals for the year 2000. In 1989, the National Governors Association established Goal 3: "American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter" (National Governors Association, 1990). A year later, in May 1990, the newly formed NAGB voted to establish three achievement levels for each grade and subject, with a central level of *Proficient* including the same descriptor: "Students reaching this level have demonstrated competency over challenging subject matter." Although the NAGB process for setting performance

levels for "proficiency" faced considerable scrutiny and controversy through the 1990s, (Shepard et al., 1993; Vinovskis, 1998), a 2017 evaluation concluded, "Through 25 years of use, the NAEP achievement levels have acquired a 'use validity' or reasonableness by virtue of familiarity" (NASEM, 2017, p. 11). The apparent interpretability of the 0- to 100-scaled percent *Proficient* statistic would help it to become a central feature of test-based public accountability policies in the decades to come, despite its technical flaws (Ho, 2008; Polikoff, 2016).

Although official reporting of NAEP scores includes states and participating large urban districts (from 6 in 2002 to 27 in 2019), NAEP data form the basis of a number of aggregate mapping efforts that represent a form of public accountability. The National Center for Education Statistics has regularly published a mapping of state performance standards to the NAEP scale (e.g., Bandeira de Mello et al., 2009; Ji et al., 2021), revealing wide variation in the rigor of proficiency standards across states under NCLB. These mappings show the rise and convergence of proficiency standards through the ESSA era. NAEP scores also form the basis of linking efforts that attempt to map school and district test scores to a common national scale for aggregate research and public engagement (Reardon et al., 2021). These secondary research efforts extend the reach of test-based public accountability even as they lag years behind the time of testing.

NCLB and ESSA required states to participate in only two NAEP subjects, reading and mathematics. Additional NAEP subjects have included science, economics, arts, foreign language, civics, geography, and technology and engineering literacy, all typically reported at the national level only. NAGB interest and budgetary constraints have led to less frequent testing in these subjects in recent years. This is consistent with a theory of public accountability: Absent required participation, competition, and alignment with state goals, tests at other levels of the system will be less effective at inspiring progress.

Hutt and Polikoff (2020) distinguished between a range of variables for public accountability, not only test-based variables but also curriculum materials, budget, and salary data. They classified NAEP as relatively inactionable compared to other variables because of NAEP's high level of aggregation and delayed reporting schedule. In the remaining subsections, we review other audiences for public accountability that might use report cards like NAEP but at different levels of aggregation and for different purposes.

Forms of Report Card Accountability

There is a long history of states implementing report card accountability, where they publish publicly available report cards of school performance indicators. These report cards are intended to provide information to parents and stakeholders to both inform parent choices and encourage external pressure for schools to improve and meet benchmarks. These "public accountability" systems began to be implemented in the mid-1990s (Hanushek & Raymond, 2005), such that by the onset of NCLB in 2002, about

20 states had report card–style accountability systems without attaching consequences. (Of note, early research on report card–style accountability systems found they did not have the same positive effects that consequential systems did; Hanushek & Raymond, 2005). Of course, states with consequential accountability systems might also put out report cards, and indeed, this became required under NCLB.

Under ESSA, all states have accountability report cards that summarize school and district performance and compare them to state benchmarks. State (and Washington, DC) report cards and dashboards vary tremendously in the ways in which they present accountability data. Of the 51 report cards produced, 26 have some form of summative rating. Of these, 12 give schools a letter grade, 8 a numerical score, 7 a categorical label (e.g., "*Commendable*"), and 5 a star rating (some states give schools multiple summative ratings). The Data Quality Campaign (2019) offers a high-level overview of progress on state accountability report cards each year. In their report, they highlighted strengths and weaknesses of report cards across the nation. Strengths include that the majority of state report cards are easily found through Internet searches, are in formats that are useful to parents (PDF and mobile friendly), and include downloadable data. Weaknesses include that most state report cards do not offer translation to any language other than English, include text at too high a reading level, lack disaggregated data for at least one federally required student group, and lack reporting on important measures like college enrollment and teacher demographics and effectiveness. In addition, the Education Commission of the States (2014) offered a more detailed summary of the similarities and differences in report cards as of that time.

Another form of reporting educational data is through public-facing "dashboards," which display school and district performance along multiple dimensions. California adopted a school dashboard, which they have modeled on dashboards from other jurisdictions (Darling-Hammond et al., 2014). In California's dashboard (Polikoff et al., 2018), schools are given ratings for "status" and "change" on each of several dimensions (including ELA and math achievement, chronic absenteeism, suspension rates, and English learner progress) on a five-point color-coding rating scheme. These public-facing dashboards are intended to drive parent decisions and involvement in states' (nonconsequential) public accountability system, the Local Control and Accountability Plans, but evidence suggests that few access these data (Polikoff et al., 2018) or participate in these processes (Marsh & Koppich, 2018).

How Parents Use Accountability Data

One of the main targets of accountability data is parents. Providing accountability data to parents could affect performance in several ways. For an individual parent, receiving accountability data and using it to select more effective schools for their child could result in improvements in that child's performance. Accountability data use by parents could also affect the broader system. Or if enough parents made school choices based on accountability data, they could drive systemic improvements in school

performance by leading to the growth of more effective schooling options or the closure of less effective ones (for more on this topic, see the literature on "exit" and "voice," for instance, Hirschman, 1970; for a recent critique of the theory by which public accountability of this form can result in educational improvement, see Hutt & Polikoff, 2020). And, of course, individual parents might also use their child's state test results to press for educational improvement for their individual child (e.g., needed instructional supports, acceleration opportunities).

School accountability report cards themselves have not been the subject of much empirical study, but there are a few exceptions. A study by the Institute of Education Science's National Center for Education Evaluation and Regional Assistance (2019) tested the impact of five different features (e.g., the balance of numbers and graphs, the ordering of schools by distance or performance, and whether district averages were included) on school report cards on parents' knowledge, satisfaction, and school choices in an experimental simulation study. They found evidence that these features matter (though to a fairly modest extent), with each feature mattering in different ways. For instance, they found that presenting accountability results as numbers only increased parents' ability to accurately answer factual questions about school performance indicated on the report card, but presenting numbers and graphs increased parent satisfaction. They also found that including parent satisfaction in accountability report cards boosted parents' satisfaction, but slightly diminished their knowledge. Finally, they found that adding more information to the presented results boosted parents' satisfaction but may have led them to select fewer effective schools. The Education Commission of the States (2014) also analyzed parents' preferences with regard to school report cards and accountability data, finding that ease of interpretation, thoroughness of data (including nontest measures), and ability to compare schools were especially important in parents' evaluations of report card quality. Finally, Jacobsen and colleagues (2014) tested certain features of accountability report cards more formally, finding that the form by which results are reported (e.g., A–F, proficiency rate) substantially affects voters' satisfaction with high- and low-performing schools (for instance, letter grades widened the perceived quality gap between high- and low-performing schools as compared to proficiency rates).

There is also more general literature on how parents evaluate schools. This literature is not necessarily about accountability reporting per se, but their results speak to the kinds of measures that might be valued by parents on accountability report cards. Haderlein (2022) divided these literatures into stated preferences and revealed preferences. Stated preferences research relies on parent self-reports about the factors they think are most important in evaluating schools. Revealed preferences research estimates parent preferences without explicitly asking them (i.e., by analyzing their behaviors).

Survey studies of parents suggest that they value academic quality measures the most when evaluating schools. The particular academic quality variables range by study, but include teacher quality (Schneider et al., 1998), curriculum quality (Van Dunk

et al., 1998), and academic outcomes such as test scores (Armor & Peiser, 1998). Some studies suggest that there are heterogeneous preferences for these factors (with low-income and less-educated parents expressing even stronger emphasis on academic measures; Schneider et al., 1998). There is also some evidence that school demographics may play a role in parents' decisions, especially for White parents (Holme, 2002; Roda & Wells, 2013).

Revealed preferences research generally relies on housing data, school choice preference data, or experimental data (Haderlein, 2022). Housing studies confirm that parents are willing to pay more for access to schools that have higher average academic achievement (Bayer et al., 2007; Black, 1999; Figlio & Lucas, 2004). Further, Figlio and Lucas (2004) found that accountability report card results affect housing values, with parents willing to spend more money to live on the boundary of an "A" school versus a "B" school. In the school choice context, studies using parent choice data reveal that parents value a variety of factors, including average student achievement, growth in student achievement, proximity to home, racial demographics (preferring schools with demographics closer to their own), and other nonacademic factors like extracurricular activities (Glazerman & Dotter, 2017; Harris & Larsen, 2017; Hastings et al., 2009). Recent evidence suggests that parents' perceptions of school effectiveness may actually be proxied by their perception of students' peer quality (Abdulkadiroglu et al., 2017). Finally, experimental survey studies show that parents value demographics when making school decisions, but that academic achievement levels and growth are the two most important factors (Billingham & Hunt, 2016; Haderlein, 2022). A recent and innovative study by Haderlein (2022) tested parents' choices in three ways, finding that the ways they evaluate school quality do not always line up with how they make choices among schools. In particular, parents value academic achievement growth the most when choosing among schools even though the factors contribute equally to their rating of school quality. Nontest measures also matter in parents' decisions, though more so in their rating of schools than in their selection of schools.

Taken together, the existing literature clearly demonstrates that school accountability data are important for affecting parents' evaluations of school quality. Parents clearly seem to value student performance data, including both achievement status and growth. They also seem to value nontest measures to the extent they are available, especially information on teacher quality and instructional programs. In addition, parents seem to respond to student demographics by choosing schools that are more racially similar to their own race/ethnicity. While there is a very small literature on the ways in which accountability data are presented and the impacts of that presentation on preferences and decisions, that literature does suggest that key features of data presentation matter for parents' judgments, perhaps especially at the tails of the performance distribution. Unfortunately, research also suggests that more advantaged parents are more likely to access and use accountability data, implying that more needs to be done if the data are to be useful to diverse audiences.

How Teachers Use Accountability Data

Accountability data could also be used by teachers to improve instructional quality and, through it, student performance. For example, receiving data on student performance could help teachers identify gaps in performance that need to be addressed (e.g., topics or skills where students are performing poorly or well). These data could also be used to group students (heterogeneously or homogeneously) or to target interventions. There is a growing body of literature on "data-driven decision-making"—the ways that educators use data for instructional improvement (see Marsh et al., 2006, for an early review or Marsh & Farrell, 2015, for a more recent summary). This section does not review this entire body of literature, but rather summarizes some of the key findings as they pertain to the role of state accountability performance data in teachers' decision-making.

There is broad agreement that state-level achievement data often arrive too late to be instructionally useful (see Marsh et al., 2006). This has led to a proliferation of district-level assessment systems that are intended to provide more actionable evidence about student performance on a timeline that is more usable by educators; these are very widely adopted now, especially in urban districts (Burch, 2010). Indeed, the creators of summative state tests have recognized the importance of (and market for) aligned interim assessments. This has led to, for example, the Smarter Balanced Assessment Consortium creating and offering interim assessments that are more seamlessly integrated with summative state tests.

More generally, researchers have identified the key factors driving the use of educational data by educators. As just mentioned, one dimension is timeliness—if the data do not arrive sufficiently quickly, they cannot inform instructional decisions. Other key dimensions include the following (Marsh et al., 2006):

- accessibility of data, including technological capacity;
- quality of data, including both the reality and the perception of data quality (i.e., whether results accurately reflect performance);
- motivation to use data, including both internal and external forms of motivation;
- educator capacity to understand and use data, including the degree to which educators have been sufficiently supported on data use;
- educators' other beliefs and values, including their beliefs about the abilities of their students (Bertrand & Marsh, 2015);
- organizational culture and leadership, including school norms, presence of collaborative opportunities, and leadership modeling effective data use;
- organizational barriers, including pressures from the curriculum and adequate time to interpret and use data; and
- local and state context, including the history of data use and accountability.

State summative accountability results may be less useful for educators than more locally developed performance data because of their timeliness and their perceived distance from instruction.

The use of performance and accountability data by teachers can also have unintended consequences (these are discussed below in the section on accountability effects).

Whether in the context of state summative assessment data or interim/benchmark data, the results of these assessments can sometimes be used to "game" accountability systems by targeting students who are close to accountability thresholds. These so-called bubble kids are the ones whose performance is most likely to drive accountability ratings, so it is a rational (if potentially pernicious) response for educators to attend to them. There is some evidence that these responses are widespread (Booher-Jennings, 2005; Marsh et al., 2006), though the actual impact on student performance of these kinds of approaches is unclear. We discuss other unintended instructional responses in the section "The Impact of Accountability Policies".

How Researchers and Policy Makers Use Assessment Data From Accountability Systems

Assessment data from state accountability systems offer comparable performance metrics across school systems and over time. This comparability enables researchers to link relative improvements in test scores to resource allocations and policies. Phillips et al. (2018) laid out a typology of different uses of assessment data for research and policy. One type of use is to inform service provision, including to help coordinate services across government agencies (e.g., schools, social work). Another type of use is to generate descriptive evidence to support continuous improvement efforts (for example, see the University of Chicago Consortium on Schools research that has informed Chicago Public Schools' policies and practices for several decades, e.g., Allensworth & Easton, 2005; Roderick et al., 2014). A third type of use is to identify schools or districts that are performing particularly well or poorly and use them to inform improvement efforts more broadly (e.g., Cannata et al., 2017). Fourth and finally, state longitudinal data systems have been used to conduct large amounts of quantitative research evaluating specific policies and programs, ranging from teacher evaluation to accountability and curriculum reforms.

Several states have been national leaders in the creation and dissemination of educational data for research, and as a result these states are overrepresented in the literature using statewide longitudinal data systems to evaluate policy. Phillips et al. (2018) summarized these states, which include Florida, North Carolina, Texas, and Washington. State creation and maintenance of longitudinal data systems grew sharply under the Obama administration, which used Race to the Top and substantial federal dollars to incentivize states to beef up their data systems (Howell, 2015).

Another potential use of accountability data by policy makers is for continuous improvement; this is a focus of policies like California's Local Control and Accountability Plans. Hough et al. (2018) identified three key issues in the use of educational data for improvement processes. First, data use for improvement is different from data use for accountability purposes. In particular, educational data must be a part of a change-oriented pathway, which involves school and district leaders evaluating changes, running predictive analytics, and establishing priorities (Yeager et al., 2013). Second, data use for improvement must be embedded in an aligned improvement process. This includes

attention to establishing cycles of inquiry; building norms, culture, and mindset; and fostering organizational capacity (Marsh et al., 2006). Third, the specific data needed for continuous improvement vary by user and phase of the improvement process. State accountability data are not nearly enough to drive improvement in this context—more and different kinds of data are needed, and they must be provided in usable forms (The Victorian Quality Council, 2008).

In general, for data to be more useful for research and policy, Phillips et al. (2018) suggested that several factors are important. First, data are more useful to the extent that they are well linked over time and across agencies (e.g., across agencies at a given age level and longitudinally from early childhood through higher education and beyond). Second, data are more useful to the extent that they are clean and straightforward to use. Third, data are more useful to the extent that political barriers do not get in the way of their use. For instance, laws that prohibit the connection of student data with individual teacher identifiers (as in California) will limit the ability to conduct analyses related to teacher policies. Fourth, data are more useful if there are not too many administrative and bureaucratic requirements that limit the ability of researchers to publish their results (e.g., onerous data use agreements).

STUDENT AND TEACHER ACCOUNTABILITY IN K–12 EDUCATION

At lower levels of the system in Figure 16.1 are student and teacher accountability systems. These systems contrast with school accountability policies and public accountability strategies by using student test scores to inform decisions with direct impact on individual students and teachers. The intended direct mechanisms include matching students and teachers with curricular or professional experiences appropriate for their current skills (e.g., sorting students or teachers, assigning them to programs, awarding them credentials). Their intended indirect mechanisms include directing student and teacher effort toward desired goals, as well as indicating who holds responsibility, management, and control over key outcomes (Airasian, 1987).

The Main Types of Student Accountability Policies

Broadly, there are two main types of high-stakes student accountability assessments in use across the United States in the early 21st century: high school exit exams (HSEEs), which are sometimes delivered as comprehensive exams covering multiple subjects and sometimes delivered as end-of course (EOC) exams, and grade promotion exams.

High School Exit Exams

HSEEs are a policy by which students must pass one or more standardized tests to receive a high school diploma. Exit exams have evolved over time from the minimum competency testing (MCT) movement of the 1970s (Chudowsky et al., 2002). Like those tests, cut scores were set at a low level of difficulty and were more often used

to identify and remediate low-performing students (Dee, 2003; Jacob, 2001). As the standards movement grew in the 1990s and into the 2000s, more states began implementing HSEEs, and those with MCTs began increasing the rigor of these assessments. By the resurgence of the movement around 2010, the majority of states had implemented high-stakes HSEEs (Center on Education Policy, 2012). These exams were especially common in states serving larger proportions of non-White students—in 2011–2012, for instance, 85% of Hispanic students attended school in states requiring HSEEs, as compared to 69% of students overall. Since that time, the number of states requiring an exam to graduate has declined—at last count for 2018–2019, just 13 states still had a traditional exit exam in place (Gewertz, 2020).

Even as cut scores have risen above historical levels, they remain at relatively low levels of difficulty, often representing middle school or early high school content (Achieve, 2004). Furthermore, the vast majority of exit exam policies have come with numerous escape valves or alternative methods to pass the assessments. All states that had exit exams offered students the opportunity to retake the exams if they failed, with states allowing anywhere from 2 to 12 retakes (Center on Education Policy, 2009, 2012). More than 80% of states also offered alternative pathways to meet the standards—for instance, many states allowed test scores from other tests such as SAT or ACT to count, permitted students to use portfolios or projects to substitute for test scores, or offered waivers or appeals for students who met other requirements. Finally, nearly all states specifically offered alternative pathways for students with disabilities (Center on Education Policy, 2009, 2012). Given these alternative pathways, the actual impact of the test itself on students' graduation may have been moderated somewhat (Harris et al., 2020).

Beginning late in the NCLB era and continuing to the present, states began to move toward EOC exams to replace more traditional exit exams, recognizing that student pathways are increasingly differentiated in high school. EOC exams differ from traditional exit exams (which are sometimes referred to as "comprehensive" exit exams) in two key ways (Center on Education Policy, 2008; see also Domaleski, 2011, for a comprehensive discussion of EOC exams). First, EOC exams are intended to assess student mastery over specific course content, as opposed to providing a more general assessment of student knowledge across subjects. Second, EOC exams are given to students when they complete particular courses, rather than at a fixed point (i.e., most exit exams are given starting in 10th grade).

An analysis by Tyner and Larsen (2019) documented the rise of EOC exams across the United States. In 1996, just 2 states had EOC exams. Over the next 12 years, the count rose an average of 1 state per year to 14 states by 2008. The next 9 years saw more rapid spread, with 30 states having adopted one or more EOC exam by 2017. Since then, the count has ticked down to 26 states (not counting any COVID-related pauses). Importantly, EOC exams are widely used for both student and school accountability, and their school accountability uses (i.e., as part of high school accountability metrics) outnumber their student accountability uses.

Early advocates for EOC exams tended to emphasize that they were more appropriate than HSEEs as tools for accountability because they were better aligned with the high school curriculum (Center on Education Policy, 2008). Furthermore, they offered the opportunity to provide feedback to teachers and students about student performance to drive targeted supports and remediation activities to students. Indeed, in some states EOC exam scores are factored into students' course grades instead of or in addition to students being required to pass the exam to receive credit (Domaleski, 2011). Still, these exams are broadly based on the same theory of action that more traditional comprehensive exit exams are based on.

Early evaluations of HSEEs describe the motivations behind the policies (Chudowsky et al., 2002). One overriding motivation is to provide some assurance that the high school diploma "means something"—that it signifies that the diploma recipient has indeed demonstrated sufficient attainment of some body of knowledge and skills needed to succeed beyond high school. By establishing clear expectations and ensuring that students must meet those expectations to receive the diploma, the HSEE should ensure at least a minimum standard has been obtained. If the standard is sufficiently rigorous, another motivation for HSEEs is that they could boost the rigor or quality of the education students are provided (i.e., setting a baseline beneath which high school teaching and learning will not fall). Third, HSEEs are seen as motivating individual students to work harder, and indeed there is evidence that high school students in the United States have not felt sufficiently challenged (Boser & Rosenthal, 2012). Fourth, HSEEs that are well aligned to state standards can provide additional weight behind standards-based reform policy, driving high schools to better align instruction to standards. Fifth, these exams could have formative uses if their results are used to identify students falling below the target and provide them support or remediation to succeed. And last, HSEEs are seen as possibly strengthening college applicant pools or providing a new source of assessment data that is better aligned with the high school curriculum than are the SAT and ACT. Advocates also note that many competitor countries in Europe and Asia have high-stakes student-level exit exams.

Grade Promotion Exams

A second main type of assessment for student accountability is the grade promotion exam. Grade promotion exams are used by states and districts to hold individual students accountable for meeting performance thresholds to be promoted to the next grade. The most common type of grade promotion exam is the third-grade reading exam, which was used by 17 states and the District of Columbia as of the 2018–2019 school year (National Conference of State Legislatures, 2019). Third-grade reading is seen as especially consequential for students' future success, with evidence that the large majority of students who eventually drop out were struggling readers by the time they were in third grade (D. J. Hernandez, 2012). Eight other states allowed the use of state tests for third-grade retention purposes but did not require it. And there are a number of large urban districts that now use or have in the past used third-grade reading

assessments for this purpose (e.g., Chicago, New York). Importantly, there are very few states or localities that do not offer numerous exemptions to the test-based promotion rules. These so-called good-cause exemptions include students with limited English proficiency (typically 3 or fewer years in an English language acquisition program); special education students; participating in an intervention; parent, principal or teacher recommendations; previous retention; demonstrating proficiency through a portfolio (student work demonstrating mastery of academic standards in reading); or passing an approved alternative reading assessment. While well intentioned and aligned with best practices in test use, recent research found that these exemptions are inequitably applied, with children of more educated mothers substantially more likely to obtain exemptions (LiCalsi et al., 2019).

There are also some test-based promotion policies at other grade levels, though it is harder to get an accurate count on the prevalence of these kinds of policies. For instance, Chicago's test-based promotion policy that began in the 1990s included promotion tests in third, sixth, and eighth grades (Jacob & Lefgren, 2009; Roderick & Nagaoka, 2005), and New York City's included Grades 3 through 8 (Mariano & Martorell, 2013).

The theory of action behind grade promotion exams is typically twofold. First, grade promotion exams are generally tied to specific remediation and intervention programs, such as summer school, assignment to a high-quality teacher, or extended reading instruction (Ozek, 2015; Winters & Greene, 2012). Through this path, the exam identifies students who need additional support, and the policy responds by providing that support. If the support is effective, student abilities and skills will increase, producing long-lasting effects. Second, like exit exams, promotion exams are thought to have motivational effects on students. Indeed, a common complaint from teachers about most state tests under standards-based accountability is that students do not have much incentive to try hard because there are few consequences for individual students; promotion exams address this concern and may also cause students to work harder in their learning. We discuss the literature on the impact of these student accountability systems in the section "The Impact of Accountability Policies".

TEACHER ACCOUNTABILITY POLICY SYSTEMS

A relatively recent target of test-based accountability is individual teachers. The Obama administration's Race to the Top and ESEA flexibility waiver initiatives dramatically changed the way that teachers around the nation were evaluated. These changes were in response to two streams of research. First, a variety of studies provided compelling evidence that teachers have large effects on students' achievement (e.g., Rivkin et al., 2005; Rockoff, 2004). Subsequent research has deepened this literature and also extended it to teacher impacts on nontest measures (e.g., Blazar & Kraft, 2017; Chetty et al., 2014a; Jackson, 2018). Second, academic research (Almy, 2011; Donaldson, 2009; Sartain et al., 2011; Sinnema & Robinson, 2007; Stronge & Tucker, 2003; Toch & Rothman, 2008; Tucker, 1997) and one high-profile report (Weisberg et al., 2009) suggested that

teacher evaluation systems largely failed to differentiate teachers or to provide useful feedback to improve their teaching.

As a result of Race to the Top and the waivers, states substantially redesigned their teacher evaluation systems. The main features of these revised teacher evaluation plans included (a) the use of multiple measures of performance to evaluate teachers, (b) the inclusion of student outcomes (measured most often by state test scores or student learning objectives) in evaluation, (c) the use of three or more levels of ratings (i.e., not just "*Pass/Fail*"), and (d) the attachment of consequences to results of evaluations (Steinberg & Donaldson, 2016).

In terms of the component measures of new teacher evaluation systems, by far the most common measure—appearing in every state and district plan analyzed in a 2016 survey of 50 states and 25 large districts—was classroom observations (Steinberg & Donaldson, 2016). Observation-based ratings of effectiveness also comprised the largest portion of teachers' overall evaluation scores, on average representing a bit more than half of the total rating. About one fifth of states and half of large districts also included a separate, but very lightly weighted (about 2%–5%, on average) measure of teachers' professional conduct. Measures of student learning were also an essential part of teacher evaluation systems postwaiver/Race to the Top (RTTT). Test-based measures—value-added models or student growth percentiles, which we discuss later—were part of 80% of states' and large districts' evaluation systems. In other states and districts, or in nontested grades, the most common measure of student performance was student learning objectives. Finally, some districts and states included other measures in teacher evaluation systems. For instance, 26%–30% of states and districts included schoolwide achievement data in individual teachers' evaluations, and about one in six states and districts included results from student surveys.

There were a number of important technical challenges in the design and implementation of these systems. One of the major challenges was in including measures of student learning that could be applied to all teachers. The large majority of teachers do not teach in subjects and grades for which student growth models can be calculated from state tests (i.e., all teachers below Grade 4 and above Grade 8, teachers of any subject other than mathematics or ELA). For these teachers, alternative measures of student learning could sometimes be computed. But these could sometimes result in seemingly nonsensical and indefensible measures being used, such as the art teacher who is evaluated in part by the growth in students' mathematics achievement (Jacobs, 2015). Another common approach for these teachers was individualized "student learning objectives," which were widely used but lacked standardization and adherence to psychometric standards (Buckley, 2015). They also ultimately do not seem to have resulted in meaningful instructional improvement (e.g., Lachlan-Haché, 2015).

Research has begun to shed light on the implementation and effects of these teacher evaluation policies. First, just as before these systems were implemented, very few teachers seem to be rated as needing improvement—a study of 24 state evaluation systems found fewer than 5% of teachers rated as needing improvement (Kraft & Gilmour,

2017). Second, the low percentage of teachers rated as needing improvement does not comport with district leaders' subjective ratings of teacher performance (i.e., when asked directly, leaders rate far more than 5% of teachers as *Below Proficient*; Kraft & Gilmour, 2017). Third, teachers' overall ratings in evaluation systems are highly sensitive to these systems' design decisions (Steinberg & Kraft, 2017). For instance, the greater the weight on student achievement and other norm-based measures of performance, the larger the proportion of teachers deemed *Not Proficient*. Finally, where there is a large literature on achievement-based measures of teacher performance, researchers have argued that relatively less is known about the properties of nontest measures of performance, such as teacher observations (J. Cohen & Goldhaber, 2016).

What we do know about nontest measures suggests that they may have many of the same challenges as test-based measures. For instance, there is evidence that the attachment of stakes to evaluation scores can change the distribution of those scores (Liu et al., 2019) and that there are racial gaps in teacher effectiveness ratings that are driven by school-level demographic differences (Steinberg & Sartain, 2021). Teacher observation scores can have substantial measurement error unless they are averaged over multiple lessons and raters (Ho & Kane, 2013). However, findings from randomized studies also suggest that observation scores can be unbiased estimates of instructional quality (Bacher-Hicks et al., 2017). Summarizing the impact of the Obama-era push for teacher evaluation reform, a recent national study found precisely zero effect on student achievement nationwide (Bleiberg et al., 2024).

THE IMPACT OF ACCOUNTABILITY POLICIES

What have been the effects of accountability on key teacher and student outcomes? This section reviews the available literature on these impacts, focusing on the effects on (a) student achievement as measured by standardized tests; (b) longer term outcomes, including graduation, college, and outcomes beyond college; (c) teacher outcomes, including their instruction; and (d) unintended consequences. This section draws on several recent reviews of accountability and its effects (Figlio & Loeb, 2011; National Research Council, 2011; Polikoff & Korn, 2020). First, we discuss the evidence focused on school accountability systems and then teacher and student accountability.

Effects of School Accountability on Student Achievement

There are several main methods that have been used to estimate the impact of school accountability policies on student achievement on state tests. The most prominent studies have used state-level data to compare changes in achievement levels and trends between states that implement new accountability systems and those that already had existing systems in place (using difference-in-differences or comparative interrupted time series methodologies). Investigating the impact of NCLB on student achievement, two studies (i.e., Dee & Jacob, 2011; Wong et al., 2015) used analyses of this type on data from NAEP. Both studies found evidence of positive to null effects, depending

on grade and subject. Dee and Jacob found achievement impacts of around a quarter of a standard deviation in elementary mathematics. Wong and colleagues found similar impacts in elementary mathematics, as well as statistically significant impacts in middle school mathematics and marginally significant impacts in elementary reading. Interestingly, these positive effects come despite the fact that NCLB's consequences were largely ineffective—supplemental education services saw low participation and ineffective program design (Heinrich et al., 2010), and vanishingly small numbers of students participated in public school choice, often because there were no options available (Lauen, 2008). Earlier studies similarly leverage variation among states in either the timing or the strength of accountability in attempts to isolate a causal effect; these studies generally find positive achievement effects of similar magnitude (e.g., Carnoy & Loeb, 2002; Hanushek & Raymond, 2005).

Another common approach to estimate the impact of accountability is to leverage within-state variation in state or district policies. Figlio and Loeb (2011) grouped these studies into three categories: (a) studies that look within states or districts over time (e.g., Jacob, 2005; Klein et al., 2000); (b) studies that compare districts that implement accountability policies to those that do not (e.g., Ladd, 1999; S. S. Smith & Mickelson, 2000); and (c) studies that examine the differential responses of schools subject to and not subject to accountability pressure using regression discontinuity design (e.g., Rockoff & Turner, 2010). They concluded that there is evidence of a positive relationship between school accountability and student achievement, but that this relationship is not universal and varies in magnitude.

The designs of most state accountability systems are well suited to the use of regression discontinuity designs, with schools just above and below the threshold functionally similar except for their accountability classification. Studies in North Carolina (Ahn & Vigdor, 2014), Wisconsin (Chakrabarti, 2014), and Kentucky (Bonilla & Dee, 2020) found positive effects of accountability on student achievement. In contrast, Dee and Dizon-Ross (2017) found null effects in Louisiana, Hemelt and Jacob (2020) found essentially null effects in Michigan, and Atchison (2020) found null to negative effects in New York State. A study of New York City school grades by Winters and Cowen (2012) also used regression discontinuity (RD) designs at the C/D and D/F thresholds. That study found heterogeneous effects, with positive impacts at the D/F margin but negative impacts at the C/D margin.

Studies have also investigated the heterogeneous effects of accountability policies on different types of students, often with a focus on the impact of accountability on student achievement gaps. This is an important area of research, because NCLB and other accountability policies generally aim to close racial and socioeconomic achievement gaps. One study of achievement gaps in North Carolina found that NCLB accountability narrowed the Black–White test score gap by around 4%–10% of the baseline gap in both mathematics and reading (Gaddis & Lauen, 2014). Other studies also provided evidence of achievement gap effects by showing that accountability affected subgroup achievement differently, again generally leading to a modest narrowing of gaps (e.g., Dee

& Jacob, 2011; Figlio et al., 2009; Lauen & Gaddis, 2012). Pre-NCLB accountability studies found mixed effects, with Hanushek and Raymond (2005) finding a narrowing of the White–Hispanic gap and Harris and Herrington (2004) finding no impact of accountability on performance gaps.

None of the above-cited studies is generalizable unconditionally, but the results taken together demonstrate that school accountability policies often have a positive impact on average student achievement. These effects may be concentrated in mathematics and in the early grades. And the effects on achievement gaps are modest, at best. The impacts of accountability appear not to be uniform, however, because there is evidence that these effects are stronger for the lowest-performing schools (Figlio & Rouse, 2006; Jacob, 2005), in states with higher proficiency standards (Wong et al., 2015), and in states with greater degrees of local autonomy (Loeb & Strunk, 2007).

Our conclusion about the impact of accountability aligns with that of Figlio and Loeb (2011), who said, "Taken as a whole, the body of research on implemented programs suggests that school accountability improves average student performance in affected schools, at least in general" (p. 410). And our conclusion is more positive than that of the National Research Council (2011), which noted,

> Test-based incentive programs . . . have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests . . . the overall effects on achievement tend to be small and are effectively zero for a number of programs. Even when evaluated using the tests attached to the incentives, a number of programs show only small effects. (p. 4)

The differences between our conclusion and that of the National Research Council may have to do with at least two factors. First, the literature on this topic has grown dramatically in recent years, with many of the positive effects coming in more recent studies using more advanced methods. And second, the difference hinges on the definition of "small" as pertains to the magnitude of accountability effects. We agree with Kraft (2020) that effects of even just .2 standard deviations in performance are not small, even though they pale in comparison to the magnitude of performance gaps and are clearly insufficient to narrow gaps with higher performing nations. This is especially the case insofar as the effects were cumulative across grades and years—the resulting long-term impacts could indeed be quite large.

Despite the positive evidence overall, there appears to be very modest evidence of any significant gap-closing effects of accountability systems. This is notable insofar as student group accountability is intended to drive educators to work to narrow these gaps—that portion of the theory of change clearly appears to be failing. Furthermore, concerns about accountability policies are often driven by equity issues—that standards-based accountability undermines possibilities for culturally relevant pedagogy (Royal & Gibson, 2017), leads to proceduralized instruction in classrooms serving historically marginalized youth (Polikoff & Struthers, 2013), and overall undermines

educational opportunity (Rebell, 2008). At the very least, it is by now clear that performance gaps will need to be closed with other policy levers rather than through test-based accountability.

Effects of School Accountability on Longer Term Student Outcomes

Much less is known about the long-term effects of accountability or the effects on outcomes besides student academic achievement. In large part, this is due to data limitations that may be addressed as new state longitudinal data systems mature. There are two exceptions to this pattern. A study using longitudinal graduation rate data from before and after the NCLB era found that the introduction of accountability systems coincided with a significant bump in high school graduation rates equal to about one fifth of the descriptive increase over that time (Princiotta, 2019). A second study, by Deming et al. (2016), explored the effect of Texas high school accountability policies during the 1990s on college attainment and future earnings. They found that schools on the threshold of a *Low Performing* label increased student achievement on high-stakes exams, which led to higher rates of college enrollment and completion and higher salaries at age 25. In contrast, they found a negative effect of accountability in higher performing schools, which they attributed to the reclassification of low-performing students into categories exempt from taking state tests (e.g., special education). There is a great need for further research on the impact of accountability on longer term outcomes.

Effects of School Accountability on Instruction and Teacher Outcomes

The theory by which accountability policies boost student achievement operates through these policies' effects on teacher behaviors. The effect of accountability on teacher practice and other teacher outcomes is even more difficult to estimate. This is primarily because there is a lack of high-quality measures of teachers' instruction at the kind of scale needed to conduct causal analyses. However, there are many correlational and descriptive studies that explore teachers' instructional responses to school accountability, especially focusing on teachers' instructional alignment to standards. The results of this work generally suggest modest instructional change in the direction of the standards, as called for by the theory underlying standards-based accountability.

A large number of self-report studies using surveys or interviews indicated that teachers believe they practice aligned instruction and are making efforts to increase the alignment of their instruction with standards and assessments (see, for instance, Hamilton, 2012; Pedulla et al., 2003). However, when teachers' practice is assessed by external means, such as through observation, it often appears less aligned to the intent of standards than teachers think (see, for instance, D. K. Cohen, 1990; H. C. Hill, 2001; McDonnell, 2004; Spillane, 2004). Still, there is evidence that the quality of instruction

and school goals for student performance have improved over the early years of the accountability era (Hunter, 2019; Lee & Lee, 2020).

A different approach to studying alignment involves surveying teachers about the content of their instruction and then using alignment methods (discussed in more detail in the section "Methods of Measuring Alignment") to compare these survey responses to content analyses of standards and assessments (see Polikoff, 2012a; Porter, 2002). These approaches make it less likely that social desirability and other self-report issues could bias the results. Several studies that use these techniques find that: (a) the alignment of instruction to standards is low, much lower than reported on other kinds of surveys; (b) instructional alignment generally increases over time, especially in mathematics; and (c) some state policy features and characteristics of teachers predict stronger instructional alignment (Polikoff, 2012a, 2012b, 2013). Still, the findings taken together indicate significant gaps in the extent to which teachers are implementing standards in the classroom, even decades into the standards era and with considerable accountability pressure.

There is also some evidence that accountability affects various dimensions of teachers' working conditions, though these effects seem modest and mixed in direction. Two studies of NCLB, for instance, found contrasting effects. Grissom et al. (2014) found that the law had no impact on teachers' hours worked, job satisfaction, or commitment to teaching; had small negative effects on teachers' perceptions of cooperation; and had small positive effects on teachers' perceptions of classroom control and administrator support. Dee et al. (2013) found that the law increased teachers' salaries and teacher-reported student engagement. Two other studies have found that accountability decreased teachers' job security (Reback et al., 2014), increased the frequency of involuntary school transfers, and decreased the frequency of involuntary attrition (Sun et al., 2017). And a more recent study on just the impact of grade-level accountability testing requirements found no impact on teacher attrition (Fuchsman et al., 2020). In sum, the effect of accountability on teacher working conditions appears mixed and modest.

Unintended Consequences of School Accountability

As indicated in Figure 16.1, there are also a number of well-documented unintended consequences of accountability policies, which arise when teachers and schools strategically seek to maximize their scores and avoid accountability consequences. One type of gaming behavior is through the reallocation of instructional time. For example, schools have been shown to allocate instructional time away from nontested subjects and toward tested ones (Dee et al., 2013; Judson, 2013). Another variant on this is the allocation of instructional time to specific test-taking strategies (for instance, using class time to take practice tests, teach strategies for completing specific types of test items, or narrowly focus on content that is expected to be assessed; Jennings & Bearak, 2014; Jennings & Sohn, 2014). There is ample evidence that these behaviors have taken place under recent accountability regimes. Thus, positive effects that are measured on state

tests should generally be cross-validated with lower stakes exams like NAEP (Koretz, 2008).

Another type of strategic manipulation is when educators narrowly target resources and instructional efforts on the students who matter the most for accountability ratings. Under NCLB, for example, school performance was determined by the percentage of students scoring above a proficiency cut point on the state test. This creates incentives to target efforts on students just below the proficiency cut point (often referred to as bubble kids), since improving those students' scores would be more likely to boost a school's performance rating. There is some evidence that teachers diverted attention to bubble kids (Booher-Jennings, 2005; Jennings & Sohn, 2014), including evidence that positive effects of accountability were concentrated on students close to the proficiency cut score (Neal & Schanzenbach, 2010). However, other studies have not replicated the bubble kids finding using achievement data (e.g., Cronin et al., 2005).

The most damaging and pernicious unintended consequence of accountability policies is perhaps their effect on teacher cheating, which can be enabled or even encouraged by school and district leaders. High-profile cheating scandals have emerged in some places, and these scandals have often been blamed on the intense pressure associated with accountability policies. For example, an Atlanta cheating scandal ended in 11 teachers being convicted on racketeering charges (Blinder, 2015). A study in Chicago used unusual test score patterns and identified that approximately 5% of teachers had likely cheated on the state exams (Jacob & Levitt, 2003). Cheating-type behaviors have also been reported in nontest outcomes, such as graduation rates (Edwards & Mindrila, 2019).

Effects of Student Accountability

Exit Exams

The literature on the impact of HSEEs on student achievement, high school graduation, and other desired outcomes consistently indicates that these exams do not lead to the outcomes desired by their proponents (Holme et al., 2010). At the same time, the worst fears of HSEE opponents—that they would lead to large-scale reductions in opportunity, especially for disadvantaged groups—appear not to have materialized given that graduation rates have trended upward for decades and, prepandemic, stood at all-time highs overall and for all racial-ethnic groups (Harris et al., 2020).

A number of studies using quasi-experimental designs have examined the impact of HSEEs on student achievement (Grodsky et al., 2009; Jacob, 2001; Reardon et al., 2010), graduation/dropout rates (O. Baker & Lang, 2013; Dee & Jacob, 2006; Hemelt & Marcotte, 2013; Ou, 2009; Papay et al., 2010; Polson, 2018; Reardon et al., 2010), and labor market outcomes (Baker & Lang, 2013; Warren et al., 2008). In general, these studies have been focused on comprehensive exit exams rather than EOC exams. Some studies estimate the causal impact of just passing the exam compared to the counterfactual of just failing. Other studies estimate the causal impact of a policy year that includes an exit exam compared to the counterfactual of a policy year that does not.

Broadly speaking, this literature produces several clear conclusions (for a comprehensive review of evidence on exit exams, see Holme et al., 2010). First, exit exams seem to have no effect on student achievement as measured by standardized tests (Grodsky et al., 2009; Jacob, 2001; Reardon et al., 2010). Second, exit exams do modestly decrease graduation rates and increase dropout rates, and these effects are concentrated among Black students and very-low-performing students and in states that have more rigorous exams (Baker & Lang, 2013; Hemelt & Marcotte, 2013; Jacob, 2001). And third, these exams have no appreciable effect on labor market outcomes (Baker & Lang, 2013; Warren et al., 2008). In response to these disappointing findings, there was mounting pushback against these exams starting in the mid-2000s. This pushback followed many researchers who opposed these exams from their onset; for a discussion see, for example, Warren & Grodsky (2009).

Promotion Exams

There is a large literature on the impact of test-based promotion policies on a range of short- and long-term outcomes for students, including student achievement in the tested grade (Winters, 2018); subsequent achievement and promotion (Greene & Winters, 2007; Mariano & Martorell, 2013; Mariano et al., 2018; Schwerdt et al., 2017); attendance, graduation, and other attainment measures (Eren et al., 2017; Jacob & Lefgren, 2009; Mariano et al., 2018; Schwerdt et al., 2017); and other outcomes like behavior, incarceration, and college enrollment (Eren et al., 2018; Ozek, 2015; Schwerdt et al., 2017). In general, this literature uses regression discontinuity methods to provide convincing causal estimates of the impact of these policies. However, there are methodological challenges in this work, especially in terms of choosing an appropriate comparison group, tracking impacts longitudinally, and separating out the effects of the retention policy from other portions of the intervention, like summer school prior to the retention decision (see Roderick & Nagaoka, 2005, for a discussion of these and other challenges).

In terms of short-term effects, studies found mixed evidence that the threat of retention induced by the policy boosts student achievement, with some evidence that it does, especially for the lowest-performing students (Winters, 2018). Subsequent to the retention decision, research suggests there are positive and substantial effects of retention on achievement, but that these effects depend on who retained students are compared against (same-grade vs. same-age peers; see Mariano & Martorell, 2013, for a good discussion of this issue; Greene & Winters, 2007; Schwerdt et al., 2017; Winters & Greene, 2012). There are some studies showing smaller positive or null effects (Jacob & Lefgren, 2004; Roderick & Nagaoka, 2005) and there is also evidence that test score effects may diminish over time (Schwerdt et al., 2017; Winters & Greene, 2012). Test-based retention policies seem to have null or perhaps modest negative effects on long-term outcomes like high school completion (Eren et al., 2017; Jacob & Lefgren, 2009; Mariano et al., 2018; Schwerdt et al., 2017), with evidence that retention policies that occur later in students' schooling may be especially harmful for the completion of the lowest-performing students (Jacob & Lefgren, 2009; Mariano et al., 2018). Research

on other outcomes found that test-based promotion policies can lead to increases in adult crime and student misbehavior (Eren et al., 2017, 2018; Ozek, 2015).

In short, the literature suggests that whatever benefits there may be from test-based retention and support in the early years, the long-term impacts of these policies appear to be close to zero and may in some cases be negative. These modest findings are notable given the large cost of student retention and its disproportionate effects on historically marginalized student groups (West, 2012). Importantly, the effects of test-based retention policies undoubtedly vary as a function of the rigor of the assessment and cut score, the nature of the interventions and supports that are paired with test failure, and the extent to which there are nontest alternatives available for students.

Special Considerations for High-Stakes Student-Level Accountability

In their review of high-stakes testing for student placement and promotion, Heubert and Hauser's National Research Council report (1999) identified three criteria for fair test use for students: psychometric adequacy, opportunity to learn, and educational benefit. The *Standards* (AERA et al., 2014) also have specific criteria listed for these tests in an illustrative example:

> When tests are used for promotion and graduation, the fairness of individual score interpretations can be enhanced by a) providing students with multiple opportunities to demonstrate their capabilities through repeated testing with alternative forms or other construct-equivalent means; b) providing students with adequate notice of the skills and content to be tested, along with appropriate test preparation materials; c) providing students with curriculum and instruction that afford them the opportunity to learn the content and skills to be tested; d) providing students with equal access to disclosed test content and responses as well as any specific guidance for test taking (e.g., test-taking strategies); e) providing students with appropriate testing accommodations to address particular access needs; and f) in appropriate cases, taking into account multiple criteria rather than just a single test score. (p. 186)

Unlike Heuber and Hauser (1999), the *Standards* (AERA et al., 2014) do not count educational benefit to the marginal or average student as a criterion. As evidence continues to mount that these student-level accountability exams have no long-term positive effect for average or marginal students and are sometimes detrimental to historically marginalized groups (Penfield, 2010), the burden of proof should be considerable for current and new programs to argue that existing or proposed remediation strategies for failing students have a positive causal effect over social promotion. As we reviewed in an earlier section, tests serve political purposes as well as educational purposes, including signaling worthy goals and managing and controlling systems. It would be best if these purposes could be served with minimal detriment to students, including by following the guidance from the *Standards* and opening up multiple pathways to demonstrating readiness.

DERIVED ACCOUNTABILITY SCORES WITH A FOCUS ON GROWTH MODELS

As Table 16.1 indicates, the type of score creates opportunities to both achieve intended aims and inspire unintended negative consequences in accountability systems. The primary NCLB reporting metric was the percentage of proficient students, an aggregate status measure that was an obviously poor proxy for school quality because of its conflation with prior learning and other out-of-school opportunities to learn (Linn et al., 2002; Ryan, 2004). Proficiency percentages also exhibit deeply problematic properties when they become the basis of score trends and gap trends, where any group with proficiency percentages near 50% is expected to manifest greater trend magnitudes because of the density of the distribution local to the proficiency cut score (Ho, 2008; Holland, 2002). As we reviewed in previous sections, the statistic also enables gaming behavior related to triaging low-scoring students who may require greater effort to teach to proficiency. And it results in a considerable loss of information about differences and changes in student performance at other locations of the score scale. In these ways, the selection and properties of the derived score for any accountability purpose deserve scrutiny. In this section, we illustrate this by discussing the use of growth metrics for accountability purposes.

A debate about "status versus growth" motivated flexibility in the form of the GMPP (2005). In the GMPP announcement, Secretary of Education Margaret Spellings never defined what "growth" meant (U.S. Department of Education, 2005). The announcement insisted only that models adhere to seven "bright line principles," such as "Ensure that all students are proficient by 2014," and that the model "must track student progress." Subsequent guidelines from the peer review panel similarly left latitude for growth definition and model specification (U.S. Department of Education, 2006).

The guidelines at the announcement of the Race to the Top competition seemed to be more explicit and defined growth as "the change in achievement data for an individual

Table 16.1 An Example of a Categorical Model From Delaware's 2009–2010 School Year

	Year 2 Level				
Year 1 Level	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Proficient	0	0	0	0	300

student between two or more points in time" (U.S. Department of Education, 2009, p. 59742). However, it then continued, "A State may also include other measures that are rigorous and comparable across classrooms," along with its motivations "to allow States the flexibility to develop data and assessment systems" (p. 59742). This left space for GMPP models to continue and effectively took no position on the definition of growth, a degree of flexibility that continues under ESSA.

As the final report of the GMPP described (Hoffer et al., 2011), states took a variety of approaches to operationalizing growth. Models continue to proliferate through the ESSA era, including renewed focus on "through-course" or "through-year" models that track growth from a fall test through a spring test. In this section, we review prototypical models and demonstrate that these models operationalize growth using related but fundamentally distinguishable approaches, and differently at the student level than at the aggregate level.

The proliferation of growth models in a policy space constructed deliberately to allow for flexibility has led to confusion among terms and definitions. We follow the general framework and nomenclature provided by Castellano and Ho's *A Practitioner's Guide to Growth Models* (2013a). In their guide, Castellano and Ho attempted to be explicit about each growth model and included its aliases and statistical foundations. In addition, they articulated the primary interpretations that models support in terms of three answerable questions: how much growth, growth to where, and what caused growth?

The gain-based model is an intuitive and largely straightforward model that requires a vertical scale, a challenging constraint with stringent requirements reviewed in Moses (this volume). The categorical model is a flexible framework that considers student status in a small number of categories (usually four to nine) and operationalizes growth in terms of transitions between categories (R. Hill et al., 2006). And conditional status models, including the student growth percentile model (SGP; Betebenner, 2009), express growth in terms of status beyond expectations given past scores. Value-added models are related to conditional status models and associate aggregate conditional status with the value that teachers and schools add to test scores.

Gain-Based Models

The first and arguably most intuitive growth model is a trajectory that each student has over time that can be described by a slope or a gain. The statistical foundation is the simple difference between two scores. Extensions are straightforward and can include estimating trajectories over more than two points in time or allowing for nonlinear trajectories, in the tradition of longitudinal data analysis (e.g., Singer & Willett, 2003). Although gain scores are intuitive, they become problematic in that they rely on vertical scaling decisions, whereby expected gains may differ in magnitude across grades. Although average gains may be an accurate representation of the amount of learning in each grade on an absolute scale, this may be more attributable to the typical developmental trajectories of children than to schools. Straight comparisons of gains

on developmental vertical scales are thus inappropriate for comparing the amount of growth for which teachers and schools may be responsible.

Categorical Models

The categorical model, also known as the value table or the transition matrix model, divides each within-grade score scale into a smaller number of ordered categories (R. Hill et al., 2006). Table 16.1 shows Delaware's value table for the 2009–2010 academic year (Delaware Department of Education, 2010). A student who scores in Level 1A in Year 1 but Level 2A in Year 2 receives a growth score of 225, as shown, and the average across students in the school represents the school-level score. The model relies more than others on the selection of cut scores, where transitions between categories function as student growth data. Logically, the cut scores between Level 1B and Level 2A must have some basis for equivalence.

The categorical model is flexible in the sense that values for particular transitions between categories can be adjusted to user specifications. As R. Hill et al. (2006) demonstrated, careful selection of values for particular transitions can result in a pre-growth-era status model, where only proficiency is counted, or something that seems more gain based, where the gain is quantified as the number of levels that are gained or lost. The cost of this flexibility is the loss of information that comes with categorization, where the model cannot distinguish between the very highest and the very lowest scores in any given category. Although this may seem inappropriate for comparing growth for individual students, at the aggregate level, the errors due to coarse categorization are diminished, particularly as the number of categories increases. However, at a certain number of categories, judgments that support differing values become more difficult to distinguish and justify, and the model becomes likely to reduce to something similar to a gain-based model, with a number of categories approaching the number of score points.

The categorical model technically provides growth descriptions. However, the values that are selected for the categorical model may be motivated by inferences about whether a particular transition between categories is sufficient to warrant an "on-track" designation for students making that transition. To the extent that these inferences inform the choice and interpretation of values, the function of a categorical model is one of growth prediction, as well as growth description. In the case of some growth models, like Iowa's model under its GMPP (Hoffer et al., 2011), this took the form of values like those in Table 16.1, except any nonzero value was simply a 1. This was based in part on the argument that a gain in categories established students as being on track to *Proficient*. In this way, the categorical model can support both growth descriptions and growth prediction.

Conditional Status Models and the SGP Model

Betebenner (2009) introduced the SGP metric as a normative approach to describing student growth. The SGP metric uses nonlinear quantile regression to support

conditional status interpretations, where the current status of a student is referenced to expected percentiles given the score history of students. Although the name of the metric seems to indicate the percentile rank of a student's gain score, the statistical foundation is regression based, where a student's current status is considered in light of expectations given past scores.

Castellano and Ho (2013b) reviewed the SGP estimation procedure in detail. Fitting the statistical model can be time-consuming for large data sets and uses an open-source R library (Betebenner et al., 2023). The SGP is calculated by first estimating 100 nonlinear quantile regression manifolds, for quantiles from .005 to .995, where the outcome variable is the "current" score and the predictor variables are all prior-year scores. Castellano and Ho (2013b) demonstrated that this is practically similar to a straightforward linear regression model of current-year scores on past-year scores, where the SGP corollary is the percentile rank of residuals. The school-level SGP metric used most often in practice is the median SGP, which Betebenner (2008) argued for on the basis of the ordinal nature of percentile ranks. Castellano and Ho (2015) showed that means are superior, where medians require around three times the sample size for the same precision.

Value-Added Models for Teachers and Schools

As the focus in the 2000s shifted from proficiency to growth, value-added models (VAMs) rose as a candidate approach that unabashedly attempted to estimate the causal effect of a teacher or school on test scores, compared to a counterfactual average teacher or school. One set of questions concerned whether and how to estimate these causal effects, and another set of questions concerned whether and how to use them to hold teachers and schools accountable to student learning. High-profile consensus reports argued against (E. L. Baker et al., 2010) and for (Glazerman et al., 2010) the use of VAMs for evaluating teachers. The American Statistical Association (2014) offered a more neutral set of cautions that was rebutted by economists (Chetty et al., 2014b). And the American Educational Research Association (2015) listed limitations and specified eight technical requirements. We briefly review some of the technical issues here and reserve policy models for teacher and school evaluation for the next section.

Koedel et al. (2015) reviewed the substantial evidence about VAMs that emerged in this period. Using experimental evidence where students are actually assigned randomly to teachers, Kane and Staiger (2008) and Kane et al. (2013) estimated experimental causal effects and showed that nonexperimental VAM estimates are statistically indistinguishable from experimental estimates. Bacher-Hicks et al. (2014) and Chetty et al. (2014a) used a quasi-experimental approach to estimate causal effects of teachers from successive cohorts that have undergone staffing changes. They also found that VAM approaches recover these quasi-experimental estimates well. The upshot is that nonexperimental methods are good approximations of quasi-experimental and experimental results in the few studies we have. Further evidence of bias when test scores are high stakes would be useful.

Koedel et al. (2015) also reviewed the substantial evidence about VAM imprecision and intercorrelation, where the general finding is that VAM is positively but weakly correlated across subjects and years (Corcoran et al., 2011; Goldhaber et al., 2013; Lefgren & Sims, 2012; Lockwood et al., 2007). VAM estimates are also positively but weakly correlated across high-stakes versus low-stakes tests (Corcoran et al., 2011; Papay, 2011) and positively but weakly correlated with other evaluation measures like principal, peer, and student ratings (Kane & Staiger, 2012).

Koedel et al. (2015) established sensible technical recommendations for VAMs, including the use of student and school covariates, free estimation of coefficients for prior-year scores (instead of fixing coefficients to 1 to fit gain-based models), and corrections for measurement error (Lockwood & McCaffrey, 2014). They also noted that fixed-effects models (one-step estimation) are similar to averaging residuals (two-step estimation), with the latter approach benefiting from computational feasibility and separate inclusion of classroom and school covariates.

Suitability of Growth Metrics for Accountability Purposes

In test-based accountability systems, there are trade-offs among different growth models along dimensions of transparency, accuracy, and the potential for negative consequences (Ho, 2014). Gain-based models will seem transparent to users for individual reporting but create inequities in accountability models as a result of dependencies between initial status and expected growth. Categorical models may seem transparent to designers of accountability systems but reduce precision as a result of coarsening and confuse point-based aggregate policy goals with score-based individual learning goals. For student-level reporting, SGPs may seem like a black box, whereas at the aggregate level there may be insufficient control of sociodemographic variables for high-stakes purposes. And VAMs may be similarly imprecise and create incentives to artificially deflate test scores from prior years. In these ways, growth models illustrate the necessity of attending to the full complexity of Figure 16.1 by specifying the level of action, score type, and intended purposes, while anticipating predictable unintended consequences.

EVALUATING THE ALIGNMENT OF ASSESSMENTS FOR ACCOUNTABILITY

We next turn to a series of important technical issues underlying test-based accountability policies. There is no measurement concept more central to standards-based education policy efforts than alignment. Even the earliest versions of standards-based education reform in the 1990s had alignment at their core. (For instance, the text of the ESSA uses a form of the word "align" 72 times, demanding alignment among an array of policy instruments, including state content standards, state summative assessments, alternate assessments, English language proficiency assessments, and college entrance requirements.) A central goal of the standards movement is to get teachers to align their instruction with state content standards. To accomplish this, state policies

are meant to align challenging assessments with those standards and to support them through aligned curriculum materials and professional learning opportunities.

But what is alignment, and how has it been conceptualized and measured? This section draws on recent alignment reviews (Martone & Sireci, 2009; Polikoff, 2022) and discusses the role of alignment in accountability policy and the state of existing accountability assessments in terms of their alignment to standards.

Alignment and Validity

When assessments are to be used for accountability purposes, as has been required under federal law for several decades, it is imperative that those assessments align with the content standards they are intended to assess (and at the performance levels required) if the tests are to be used to gauge schools' contributions to student learning. This is certainly true from a policy design standpoint—aligned assessments send educators the message that the standards are important and should be well implemented, and they are at the heart of standards-based reform theory (M. S. Smith & O'Day, 1991). But it is also true from a validity standpoint—all of the inferences about student, teacher, and school performance that are made on the basis of assessment results rely to some degree on the extent to which those assessments align with the content standards they claim alignment to.

Consider the most common type of accountability use of assessments since the NCLB era—the rating of school performance based on its students' scores on state summative assessments. Setting aside policy design decisions about the construction of accountability metrics (e.g., status vs. growth and the use of nontest measures like attendance and graduation rates), the intended goal of these assessments is to gauge how effectively a school is educating children as to the content in state standards. Thus, to make valid inferences about school effectiveness, the assessments must measure student knowledge and understanding of the content and skills in those standards. Alignment is how we gauge the extent to which the assessments measure student knowledge and understanding of the standards. The argument is the same for teacher-level and student-level accountability assessments: To make valid inferences about teacher effectiveness or student mastery, the assessments must measure student knowledge and understanding of the standards.

Alignment is closely related to other terms—opportunity to learn, instructional validity, and instructional sensitivity. Opportunity to learn is a more general term that refers to "inputs and processes within a school context necessary for producing student achievement of intended outcomes" (Elliott & Bartlett, 2016). Alignment of instruction with standards can therefore be thought of as one component of opportunity to learn. For students to have the opportunity to learn the standards, they must (among other things) receive instruction that is aligned with the standards. For students to demonstrate their opportunity to learn the standards, the assessments must (among other things) be aligned to the standards.

Instructional validity and *instructional sensitivity* are other terms that have been used almost interchangeably with opportunity to learn. An assessment is instructionally

valid or instructionally sensitive to the extent that results on the assessment accurately reflect the content and/or quality of instruction students received (whereas some definitions of instructional validity only emphasize content, most implicitly cover both content and quality of instruction; see, e.g., D'Agostino et al., 2007). Presumably, a test could be aligned to standards and not instructionally valid (if, for instance, there were substantial gaps in the standards-aligned instruction students received). Instructional validity or sensitivity is more relevant to some uses of accountability tests than others (for instance, for a student promotion exam it might not matter how a student learned particular content, though it would matter tremendously that the student had the opportunity to learn it). For more on the distinctions among these terms, see Elliott and Bartlett (2016) and Polikoff (2010).

Methods of Measuring Alignment[2]

There are three main approaches to measuring alignment, but two of them are by far more widely used (Martone & Sireci, 2009). The most widely used procedure to investigate the alignment of state tests with state standards is the Webb approach. This approach compares tests with standards using four main criteria: categorical concurrence (agreement in coverage of broad content areas), depth of knowledge (agreement in coverage of depth or cognitive complexity), range of knowledge (breadth of coverage of the test relative to the standards), and balance of representation (even representation of content on the test). The Webb model also considers the source of challenge for test items (i.e., whether the source of challenge is construct relevant or whether the item is merely "tricky"). Expert raters analyze the content and cognitive complexity of each item on a test and use that information to calculate indices for each criterion. These indices are then compared with specified benchmarks to judge alignment along each criterion (see Webb, 1999, for more detail).

The second widely used alignment approach is based on the Surveys of Enacted Curriculum (SEC; Porter, 2002). Instead of directly comparing standards with assessments, this approach maps standards and assessments onto neutral content languages. These content languages, which define content at the intersection of specific topics (e.g., multiplying fractions) and cognitive demands (e.g., perform procedures), are intended to be exhaustive as to the content that would normally be taught in a particular grade and subject. Once multiple raters have coded the topics and cognitive demands in the standards and assessments, these codes are compared to one another by way of an alignment index that indicates the proportional agreement, as well as heat maps that graphically display areas of alignment and misalignment (Porter, 2002). Multiple alignment indices have been used in the literature (Polikoff et al., 2011). The SEC approach has also been extended to teacher surveys (e.g., Polikoff, 2012a, 2012b; Porter et al., 2011) and to curriculum materials (Polikoff, 2015), and similar alignment calculations have been done in those contexts.

The third most widely used approach was pioneered by Achieve to rate alignment of state tests with standards (Rothman et al., 2002). The Achieve approach rates the

individual test items on several criteria: confirmation of the test blueprint, the match of the content to its intended objectives, the match of the cognitive demand to what is called for by the objectives, the source of the item's challenge (whether it is the target objective or something construct irrelevant), and the overall level of cognitive demand. Then, it takes these item ratings and calculates holistic ratings for (a) level of challenge, (b) balance, (c) range, and (d) rigor (all but range are qualitative ratings; see Rothman et al., 2002, for a description of each dimension). Achieve differs from the previous two approaches by using a qualitative approach to holistic ratings—that is, relying on narrative summaries of the results rather than turning them into quantitative indices (Martone & Sireci, 2009).

There are also other approaches, which have been used more sporadically. For example, one approach has been recently created specifically for Common Core standards. This approach was used in a recent study of four assessment systems: the Partnership for Assessment of Readiness for College and Careers (PARCC), Smarter Balanced, ACT Aspire, and the Massachusetts Comprehensive Assessment System (MCAS; Doorey & Polikoff, 2016). This approach asks expert raters to analyze assessments against a large set of criteria that address both the content and the depth of the standards (for example, whether ELA items require students to read closely and provide evidence from the text, whether they use a balance of high-quality literary and informational texts, and whether they require a range of cognitive demand). These ratings are then rolled up to the whole test to indicate whether the assessment has a weak match, limited/uneven match, good match, or excellent match to the standards. This methodology has only been used in one study as of this writing, but it may be appropriate for more studies if the Common Core State Standards (CCSS) or similar standards remain in place. Other alignment approaches are summarized by Forte (2017).

There are strengths and weaknesses to each of these alignment approaches, and they are described in detail by Forte (2017). Some of the relevant distinctions include the following:

1. All of the methods are dependent on expert raters and are therefore time and labor intensive. The alignment ratings and their consistency across studies have sometimes been problematic (Herman et al., 2007).
2. While the Webb and Achieve approaches measure alignment along similar dimensions, Achieve is much more qualitative and rejects the consensus approach of the Webb methodology.
3. The Achieve approach also considers the role of test blueprints in translating standards into assessments, whereas the Webb and SEC approaches ignore blueprints.
4. The SEC approach is the only approach that can be used to measure alignment of other documents (e.g., curriculum materials, survey reports of teachers' instruction) with standards and assessments.
5. The Webb approach offers thresholds for determining the adequacy of alignment on each of its dimensions, while the SEC does not, but instead offers techniques for hypothesis testing of alignment indices.

Forte (2017) concluded that the choice of alignment method "should depend most on what information the state would find most helpful in improving the quality of its systems of standards and assessments in support of instruction" (p. 11).

How Well Aligned Are State Tests With State Standards?[3]

The main role of alignment methods is to investigate the alignment of tests with standards. Under NCLB and ESSA, states must submit evidence for peer review that their assessments are meaningfully aligned with their state standards. States have typically used the Webb procedure or a variant to conduct these studies (Martone & Sireci, 2009). Evidently, peer reviewers have been satisfied with the alignment evidence that has been produced, because these assessments continued to be developed and administered for accountability. Early Webb alignment studies (e.g., Webb, 2002) identified some common alignment issues that assessments had: failing to cover the full breadth of the standards, failing to cover the objectives under a standard, and failing to reach the adequate depth of knowledge of the standards. In that study, only one of the eight tests studied met all four alignment criteria. More recent studies (e.g., Webb, 2005, 2010) have generally found more acceptable levels of alignment of state tests with state standards and weaker alignment of other tests (e.g., ACT/SAT) with state standards (e.g., NORC, 2015; Roeber et al., 2018).

Test-to-standards alignment studies have also been conducted using the SEC. The most comprehensive of these was a 2011 analysis (Polikoff et al., 2011) that compared 138 pairs of state standards and assessments in mathematics, science, and ELA. That analysis found that assessments (a) routinely overemphasized some content in the standards and underemphasized other content, (b) assessed content at lower levels of cognitive demand than was called for in the standards, and (c) contained sometimes 20%–25% of content that was not in the corresponding grade-level standards. Furthermore, certain content areas of state standards were routinely not tested year after year.

A more recent analysis also investigated the alignment of state assessments, this time in the Common Core era (Doorey & Polikoff, 2016). This study did not consider alignment the way the Webb or SEC approaches do. Rather, it examined alignment in terms of key instructional shifts called for by the Common Core standards. It found that the new generation of assessments—specifically PARCC and Smarter Balanced—did outperform a "best-in-class" state assessment (MCAS) in terms of many of these key shifts. For instance, the new assessments were better at emphasizing close reading, writing to texts, research and inquiry skills, and matching the cognitive complexity of the standards than were MCAS. However, the ACT Aspire—another test intended to align to Common Core standards—scored lower than PARCC and Smarter Balanced (and, indeed, the MCAS) on several dimensions. The PARCC and Smarter Balanced tests did fall short of aligning to standards in several ways. For instance, they failed to assess speaking and listening standards, and they did not match the complexity of the standards in mathematics. Separately, there is also the recent special case of states using college entrance examinations (ACT, SAT) as their summative assessments for high school. These assessments were not constructed for—and are plainly not aligned

to—state high school standards, though they may offer other benefits such as boosting college awareness and enrollment among groups that may be less likely to otherwise enroll in college (Lorié, 2020).

Together, these studies do suggest that state tests have sometimes struggled in certain key areas with regard to alignment to standards. In particular, they have struggled to assess the full cognitive complexity of the standards (falling short of capturing the most cognitively complex content). This is perhaps not surprising because the most cognitively complex content is often difficult to assess unless there is an extended period or some degree of student research and writing. State tests have also sometimes failed to assess key chunks of standards content (for instance, NCLB-era standards overemphasized reading skills and underemphasized writing skills relative to the standards), perhaps undermining the content messages of the standards somewhat. The differences in findings across studies also suggest that different alignment approaches (which are conducted with different groups of raters) may lead to different answers as to how well aligned assessments are with standards. There is no one "best" alignment method, nor is it even clear how one would compare results from different alignment methods to reach broader implications, which makes it difficult to make definitive statements about overall test-to-standards alignment.

KORETZIAN INFLATION THEORY AND THE VALIDATION OF TEST-BASED INFERENCES IN ACCOUNTABILITY SYSTEMS

Alignment also plays a role in the implications and evaluation of test-based inflation, where systematic misalignment creates opportunities for students, teachers, and systems to attend to construct-irrelevant features of tests. We use the term Koretzian inflation theory to credit the empirical and theoretical work of Daniel Koretz and his coauthors toward the validation of test-based inferences under high-stakes conditions (Holcombe et al., 2013; Koretz, 2008, 2015; Koretz & Barron, 1998; Koretz & Hamilton, 2006; Koretz et al., 2001, 2016). Koretzian inflation theory is a corollary of Campbell's law (Campbell, 1979) as it applies to high-stakes test-based educational inferences. Campbell's law states, "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p. 49). Koretz and Hamilton (2006) described inflated test-based inferences as follows: "Test scores can become inflated—that is, can be higher than proficiency in the measured domain warrants . . . particularly severe[ly] when the consequences for scores are substantial" (p. 542).

Empirical studies of Koretzian inflation theory ask a counterfactual question of any high-stakes inference drawn from test scores: How would the inference differ had we used a lower stakes test of the intended domain? The theory predicts that the inference drawn from the high-stakes (focal) test is inflated and biased over the inference drawn from a well-aligned low-stakes (audit) test. The canonical case is that of cohort-to-

cohort measures of average progress such as those from the NCLB-era New York State test, which from 2005 to 2009 showed an increase of over half a standard deviation in Grade 8 Mathematics. In contrast, the cohort-to-cohort NAEP progress in New York from 2005 to 2009 was less than 0.1 standard deviations (Koretz, 2017).

Without a clearer specification of the intended inference and measurement, discrepant trends are not necessarily evidence of inflation. Ho (2007) summarized state–NAEP discrepancies from 26 states from 2003 to 2005 and found that state test score trends from 2003 to 2005 were approximately four times the magnitude of NAEP score trends (0.12 to 0.03 standard deviation units over 2 years). Jacob (2007) looked at four states from 1992 to 2003 and found focal test score trends of 0.09 and 0.06 standard deviations per year in math and reading, respectively, compared to 0.05 and 0.02 standard deviations per year on NAEP.

How can we explain discrepant trends? Any differences in differences in sampling, administration conditions, or behaviors may result in a trend discrepancy. For example, this pattern of trend discrepancies might arise if lower proficiency students began to opt out of the high-stakes test but not NAEP. Student motivation may increase disproportionately for the state test over NAEP. Teachers may increase the alignment of the curriculum to tested standards for the state test over NAEP. Teachers may begin to drill tested standards for the state test over NAEP. Or students or teachers may begin to cheat on the state test but not on NAEP. In all of these cases, it is differences in differences, or the interaction effect, not just differences between tests, that account for trend discrepancies.

In the case where trends differ because of what is measured, which one is correct? Koretzian inflation theory (Koretz, 2008; Koretz & Hamilton, 2006; Koretz et al., 2001) is an applied validation framework that emphasizes a sampling principle of testing and, because it is most commonly applied to measures of progress, a distinct temporal dimension. We present a simplified version of the framework in Figure 16.2, where a test and a target of inference are intersecting circles of a Venn diagram full of measurable and inferable elements of performance. This is a common depiction where the tested (but not inferred) section represents construct-irrelevant variance and the inferred (but not tested) section represents construct underrepresentation. If what is tested is a random sample of the population of elements and the construct-irrelevant variance is random, then we appear to be assured of valid inferences.

Koretzian inflation theory adds a temporal dimension to this representation under high-stakes conditions. Applying Campbell's law and a sampling principle of testing, it asserts that even an initially random sample of tested and targeted elements can become the focus of disproportionate weight or outright corruption, if that sample becomes predictable over time. Students and teachers may notice an initially representative subsample and begin to weight it disproportionately over desired inference weights. This may lead to inflation if, for example, such elements of performance are efficient psychometrically and end up being reused or because such elements form the basis of an equating item set that ensures comparability over time.

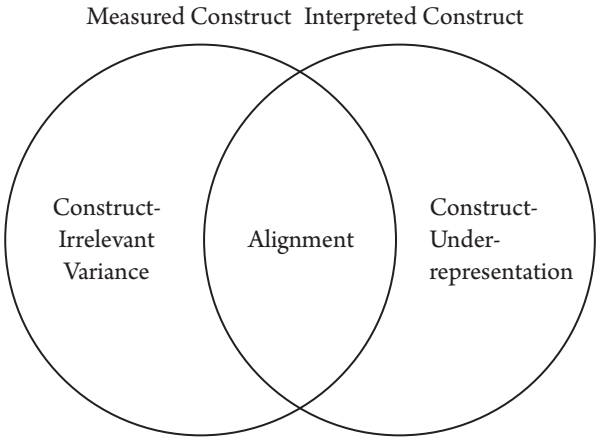

FIGURE 16.2
Illustrating Mechanisms for Inflation When Test Developers Sample Predictably From Content Domains

The theory defines *reallocation* as a process of increasing weights on tested content disproportionally over their inference weights over time. Reallocation also occurs when teachers and students decrease weights on untested elements that are nonetheless part of the target inference. This is commonly called *narrowing the curriculum* or *teaching to the test*, and we discussed evidence of its prevalence in previous sections. Koretzian inflation theory suggests that no test can simply be "worth teaching to." As long as a test is a predictable sample of desired elements, stakes will lead to invalid interpretations of progress. Reallocation degrades inferences at the right side of Figure 16.2, increasing emphasis on the intersection (what is measured) at the expense of important but underrepresented content.

Similarly, initially ignorable construct-irrelevant variance can also inflate progress over time under high-stakes conditions. The theory defines *coaching* as the disproportionate weighting of tested elements that are not part of the target inference. This may include elements like "process of elimination," where students are coached to eliminate obviously incorrect options in a multiple-choice item and, if necessary, guess among the remaining possibilities. Coaching works at the left side of Figure 16.2 by increasing emphasis on construct-irrelevant but predictable features. To the extent that eliminating possibly incorrect answers is not part of the target of inference for, say, a reading test score, Koretz (2017) argued that coaching is at best a waste of instructional time and at worst a source of bias and inflation in test scores.

Holcombe et al. (2013) added multiple levels to the simple heuristic figure sketched in Figure 16.2. The opportunities for score inflation are numerous and appear at every stage of the standard development of tests, including selecting the domain, selecting the elements of the domain, selecting standards for testing, and selecting item representations from tested standards. Each stage introduces possible construct underrepresentation and construct-irrelevant variance. The key insight is that this potential for inflation rises

with higher stakes on results, incentivizing educators to employ coaching and gaming and resulting in inflated indicators of progress.

Koretzian inflation theory also motivates a number of tools and solutions. If high stakes on individual measures are necessary, the theory clearly motivates the use of a low-stakes audit test. NAEP has served that de facto role for states and urban districts since the reporting of state and district scores in the 1990s. Although state and district leaders are held publicly accountable to NAEP progress via transparent reporting, the stakes are not as high or as direct as they are for school leaders and teachers. Numerous district tests may also serve lower stakes purposes and can serve as audit measures, although they may not be uniform and comparable at the state level.

Koretz and Beguin (2010) proposed self-monitoring assessments that embed two types of audit items into operational tests. A content audit item would assess a desired target of inference that may be underrepresented in operational items. A style audit item would measure existing content in an unpredictable way to protect against possible inflation from coaching. Koretz et al. (2016) concluded that the approach had theoretical promise but faced substantial practical constraints given the already limited space for items on operational tests. There are also potentially disproportionate costs when developing items from underrepresented content domains that may be underrepresented because they are costly to measure.

Neal (2013, 2018) proposed an alternative solution whereby two separate tests or test sections would achieve separate purposes, the first to measure progress and the second to estimate the causal effects of teachers and schools on test scores. The first test would be largely stable to measure change over time but have no direct stakes on teachers and schools. The second test could change substantially from year to year to sample broadly from the domain with no constraints on equating items, balance across content areas, or consistent item formats. Such a test would render reallocation and coaching difficult or impossible and theoretically encourage educators to focus on the curriculum standards rather than construct-irrelevant performance elements. Because models that estimate teacher and school effects regress current-year scores on past-year scores, they do not need to be equated to the same scale.

Koretz (2017) offered additional policy solutions, the most straightforward of which is to reduce stakes or dilute them through the use of a multiple-measure indicator system. Other solutions include setting reasonable and achievable targets and providing supports that are commensurate with stakes in a reciprocal accountability system (Elmore, 2004).

ACCOUNTABILITY TESTING FOR SPECIAL POPULATIONS

The logic of test-based accountability suggests principles of inclusion and comparability for all subpopulations. Following Figure 16.1, if students from special populations

are excluded or not disaggregated, it sends a signal that they are not a priority. If student scores are not comparable or counted, then the incentives to teach them become uncertain or diminished. In this section, we briefly discuss the inclusion of students with disabilities (SWDs) and English learners (ELs, sometimes also called limited English proficient, English language learners, or emergent bilinguals). Our review is necessarily brief and focuses primarily on issues directly related to test-based accountability for students in these groups. For additional consideration of these groups, see Zwick and Rodriguez and Thurlow in this volume.

NCLB's accountability provisions brought substantial focus on the measurement of all SWDs and ELs. From the earliest days of NCLB, scholars called attention to the salient accountability issues for these groups. Issues related to SWDs were prominent under NCLB, with NCLB policy specifically identifying two groups: (a) up to 1% of the tested population—those with the most severe cognitive disabilities—to be measured against alternate achievement standards; and (b) up to 2% of students with less severe disabilities but who would be unlikely to meet grade-level standards to be measured against modified achievement standards (Elledge et al., 2009). And NCLB is credited with bringing about substantial change and standardization in the processes used to assess EL students within and across states.

McLaughlin and Thurlow (2003) described some of the major substantive and technical issues related to educational accountability for students with disabilities. The authors first noted the inherent tensions of standards-based accountability with key provisions of special education law, such as the guarantees of the Individuals with Disabilities Education Act. The heart of federal law for SWDs is the individualized education program, which is not generally standards based and focuses on each child's individualized performance, goals, services, and participation in the general education classroom. In contrast, NCLB's accountability provisions required participation in the state's standards-based assessment for all but a small percentage of students with the most severe disabilities. The very nature of content standards themselves seems to run counter to the individualized nature of special education services, and special education scholars have questioned the appropriateness of standards for some SWDs. For students given alternate assessments based on the nature of their disabilities, McLaughlin and Thurlow discussed the challenges of aggregating these results with the results of the general population, including setting appropriate and attainable goals for these students and issues of confidentiality based on small group sizes. They also discussed test modifications commonly offered to SWDs and the potential for these modifications to affect the validity of inferences made on the basis of student performance. And for SWDs who are excluded for some or all of the general education curriculum, there are concerns about students' opportunity to learn and the validity of the assessment results for inferences about the general education curriculum.

Similarly, Abedi and Dietel (2004) summarized some of the major substantive and technical issues related to educational accountability for ELs. First, EL students are typically among the lowest-performing groups on content-area assessments, so the uniform

accountability targets of NCLB were seen as especially—and perhaps unrealistically—ambitious for this group. Second, EL students are especially sensitive to the language demands of accountability tests, and Abedi and Dietel made clear that these tests measure both students' content knowledge and their language ability. Third, EL students transition out of the category over time as their English language proficiency improves. Because new EL students often have low scores and high-scoring EL students transition out of the category, the average achievement of the EL category can often be low despite academic growth of EL students. If accountability systems incentivize average proficiency for EL students without attention to accurate reclassification, this creates an incentive to retain English learners in the EL category even after they may benefit from these services.

Other issues raised by Abedi and Dietel (2004) include the variation in approaches and assessments used to identify EL students, the diversity among EL subgroups (e.g., Chinese-speaking ELs versus Spanish-speaking ELs), and the tautological relationship between English language proficiency and EL classification (i.e., if all students were proficient in English, as was NCLB's goal, there would be no EL students by definition).

In the prior edition of this volume, Koretz and Hamilton (2006) discussed these and other issues relating to SWDs and ELs in the context of assessment for accountability. For instance, they discussed the heterogeneity of SWD and EL populations and noted the implications for group size (e.g., many EL or SWD groups are too small to be numerically significant for reporting). And they considered specific disability categories and the implications of students' disability types for validity issues related to construct relevance (e.g., a dyslexic student's challenges in reading text in the context of a mathematics assessment). The primary focus of their treatment of SWD and EL issues is their more extensive discussion of accommodations and modifications and their implications for validity. Changes in accommodation and modification policies brought on by NCLB's inclusion mandates had important implications for the kinds of students included in state assessment and accountability systems. While these changes brought needed attention to the academic performance of students with disabilities and raised expectations for many of these students, scholars have questioned the appropriateness of grade-level standards for many students with disabilities.

There are several important innovations pertaining to accountability assessment for ELs and SWDs under ESSA policy. These innovations broadly attempt to deal with some of the unintended consequences and issues that emerged under NCLB-era regulations. For instance, ESSA regulations do away with the 2% group for modified standards, while maintaining the 1% group with the most severe disabilities who can be tested against alternate standards. This addresses concerns that too many students were being excluded from accountability against regular grade-level standards with the 2% cap. Addressing the tension mentioned above between individualized education programs and inclusion, ESSA regulations specifically require states to have guidelines for the inclusion or exclusion of students with disabilities in the alternate assessment

system. In principle, this could again improve the appropriateness of assessments offered to students with disabilities. ESSA law also emphasizes the importance of universal design for learning in the test development process, which it defines as follows:

> The term "universal design for learning" means a scientifically valid framework for guiding educational practice that—(A) provides flexibility in the ways information is presented, in the ways students respond or demonstrate knowledge and skills, and in the ways students are engaged; and (B) reduces barriers in instruction, provides appropriate accommodations, supports, and challenges, and maintains high achievement expectations for all students, including students with disabilities and students who are limited English proficient. Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

This is intended to reduce concerns about construct irrelevance, some of which were highlighted by Koretz and Hamilton (2006). For more on the implications of the universal design for learning in the context of accountability assessments, see Rose et al. (2018). For more discussion of alternate assessments and the English Learner population, also see Rodriguez and Thurlow in this volume.

PERSPECTIVES ON EQUITY IN ACCOUNTABILITY TESTING

Educational policies like ESSA have stated aims and mechanisms to provide low-income students with improved educational opportunities. This reflects one definition of equity, the disproportionate allocation of resources to those with greatest need. The intended role of educational tests in these policies is to measure students' academic needs so that educational systems can similarly provide low-scoring students with disproportionate support. As we have reviewed in earlier sections in this chapter, there are unintended consequences when these systems place unrealistic expectations on teachers and school leaders to raise test scores for low-scoring students. In these cases, attempts to improve equity for low-scoring students may be inequitable for the teachers and leaders who support these students.

Tests may also fail to improve equity when they are disproportionately burdensome to the lowest-scoring students who take these tests. This can occur in the act of testing, when low-scoring students encounter assessment tasks and prompts that are not as engaging or affirming to them as they are to higher scoring students, or when they receive test scores, when low-scoring students may disproportionately interpret their low test scores as evidence of their irredeemable deficits rather than areas for potential growth. This section reviews selected perspectives on equity in accountability testing.

Equity Monitoring Systems

In the 2019 NASEM report, *Monitoring Educational Equity*, the authors argued that relying primarily on test scores to measure school and student performance can

undermine educational equity by neglecting the inputs and context for learning. A more comprehensive system could reveal inequality earlier in the life span, set realistic expectations for student growth, and provide national benchmarks to spur investment and improvement. Such a system would consider factors like school funding levels, curriculum quality, teacher qualifications, class sizes, student health and well-being, extracurricular opportunities, and other resources that impact learning.

The report recommended that the indicators track inequality in opportunities at the school and district level, as well as for various student demographic subgroups. This would provide insights into how policies differentially impact groups like students of color, low-income students, ELs, and SWDs. Monitoring progress in closing equity gaps could then help target resources and support. Comparisons of progress across the country could inspire learning from proven practices.

The report emphasized that developing a national monitoring system will require considerable input to secure buy-in before full implementation. Using NAEP as an example, the authors proposed a board like NAGB that oversees and reports on a system of comparable equity indicators across states and over time, inspiring improvement through public accountability. They argued that indicators must be selected carefully and with appropriately developed and publicly reported validity evidence. Ongoing evaluation would also be needed to ensure the system meets its goals of revealing inequities and spurring corrective action over time.

Returning to Figure 16.1, adopting an "equity monitoring" approach as proposed by these authors holds promise to both reveal and reduce the unintended negative consequences listed on the right-hand side of the framework. For example, highlighting inputs rather than outputs can help to set realistic expectations for growth. To the extent that there are stakes through public accountability, these stakes would be more distributed across additional indicators beyond test scores, reducing the likelihood of inflation. Unintended consequences of public reporting of indicators, like segregation, would be monitored directly. And systematic attention to disaggregating indicator data by sociodemographic categories, particularly for indicators related to educational inputs, would help to avoid deficit frameworks by highlighting inequality in educational access and opportunities (Quinn & Desruisseaux, 2022).

Culturally Sustaining Assessment and Test-Based Accountability

If the NASEM report is an effort to "measure equity" by documenting disparities in a broader range of school- and student-related indicators, there is also increased effort to "measure equitably." This effort involves measuring outputs like achievement with particular investment in understanding the learning experiences of historically marginalized students and designing assessments that engage them. This builds on theories of "culturally sustaining pedagogies" (Ladson-Billings, 1995; Paris, 2012), which are educational practices that not only acknowledge but also build on the cultural

knowledge and experiences students bring to the classroom. In classroom contexts, culturally sustaining assessments are activities designed to value, engage, and empower students (Lyons et al., 2021; Randall, 2021). (In this volume, see Ercikan & Flores; Lane & Marion; Rodriguez & Thurlow; and Zwick for discussion of culturally responsive assessments.)

Figure 16.1 distinguishes between direct score-based actions and indirect non-score-based influencing actions. For culturally sustaining assessments to function well in accountability systems, they must have positive direct and indirect actions. Scores from culturally sustaining assessments would need to accurately identify both schools and students who need support. And the practice of culturally sustaining assessment would need to improve agency and engagement among students and educators. An unintended negative consequence would be if accountability pressures on educators led them to customize tests under the banner of "culturally sustaining assessments" that inflated scores for students who needed support, thus decreasing the likelihood that they would receive the support they need.

Dee and Penner (2017) provided evidence that an ethnic studies course designed with principles from culturally sustaining pedagogy has positive effects on student attendance and grade point average in other courses. Using a regression discontinuity approach, they estimated that students near the assignment cutoff (an eighth-grade grade point average of 2.0) would have a 1.4-point higher ninth-grade grade point average had they been assigned to take the course, compared to the counterfactual of not being assigned. Bonilla et al. (2021) showed that there are still detectable causal effects in terms of high school attendance, high school graduation, and postsecondary matriculation. However, the ethnic studies course was not designed with culturally sustaining assessment for accountability as a goal, nor were the properties of scores or grades from this course a target of analysis.

Theoretical foundations for culturally sustaining testing exist in sociocognitive models for learning and assessment (Bennett, 2023; Lyons et al., 2021; Mislevy, 2018; Moss et al., 2008; NASEM, 2018; Shepard, 2021). Sociocognitive models incorporate the past experiences of students and the current context of learning and assessment, including students' relationships with other students and the teacher in the classroom. These models typically result in inferences that are richer and context referenced and thus are more limited in claims about generalizability. The sociocognitive perspective on conventional large-scale tests is that these are also context referenced to isolated individual completion of a particular subset of decontextualized cognitive tasks. Scores on these tests are then similarly limited in generalizability to other performances on tasks that also require isolated individual completion of decontextualized cognitive tasks.

Thus, although there is optimism around culturally sustaining assessment that serves students and produces low-stakes scores for classroom inferences and scholarly research (Bennett, 2023; Mislevy, 2018; Shepard, 2021), there is skepticism about

the role that scores from culturally sustaining tests might play in accountability systems (Lyons et al., 2021). Stakes would have to remain low to avoid incentives to change assessment tasks to inflate scores. Traditional psychometric frameworks that require strict comparability of score interpretations across contexts would have to evolve. Costs for test development, deployment, evaluation, and scoring would have to be feasible. And building scholarly, political, and community consensus about content standards and test blueprints for culturally sustaining tests would take time and work.

As one example of such a process, the state of Hawaii has allowed Native Hawaiians to design a test and test-based accountability system for students in the Kaiapuni Educational Program, a Hawaiian language immersion program. The Kaiapuni Assessment of Educational Outcomes (KĀʻEO) follows ESSA guidelines, including the requirement for mandatory public reporting in terms of KĀʻEO standards in mathematics, Hawaiian language arts, and science. According to the KĀʻEO technical report (Hawaii State Department of Education, 2018), the development team used the achievement-level descriptors for the English language state test, the Smarter Balanced test, as a foundation for the development of Hawaiian language achievement-level descriptors. These, in turn, guided item writing and standard setting in Hawaiian. Shultz and Englert (2021) described this as "one example of how the test developers and community stakeholders successfully balanced the tension between maintaining the technical requirements of a state assessment and serving the needs of the community as defined by the community" (p. 5).

CONCLUSION

In a first draft of this chapter on test-based accountability, written before the onset of the COVID-19 pandemic, our opening sentence was, "Over the past several decades, the arc of K–12 educational assessments has bent steadily toward accountability." At the time, although ESSA had granted some flexibility from one-size-fits-all accountability, the statement generally seemed defensible. Years later, the sentence seems laughably dated, a reminder of the pendulum swings that are common between extremes of federal and local control and the imprecision of extrapolating from past trends.

Although predicting future trends in test-based accountability is not our goal, we believe the principles outlined in the beginning of this chapter and revisited throughout should guide future validation of test score use in educational accountability systems. Once educational test scores exist in complex systems, actors use them for a predictable range of purposes that extend well beyond simple descriptions of what individual students know and can do. Validation, and the corresponding anticipation of unintended negative consequences, requires specification of actors, score-based actions, score types, and indirect actions in social and political systems (Figure 16.1).

ACKNOWLEDGMENTS

The authors thank the coeditors, Linda Cook and Mary Pitoniak, the designated reviewers of this chapter, Laura Hamilton and Joan Herman, and the Editorial Advisory Board reviewer, Michael Rodriguez, for their constructive feedback and insights. We also thank Lily An and additional members of the Harvard Graduate School of Education's Educational Measurement Lab for their suggestions.

REFERENCES

Abdulkadiroglu, A., Pathak, P., Schellenberg, J., & Walters, C. (2017). *Do parents value school effectiveness?* (Working Paper No. 23912). National Bureau of Economic Research.

Abedi, J., & Dietel, R. (2004). Challenges in the No Child Left Behind Act for English-language learners. *Phi Delta Kappan, 85*(10), 782–785.

Achieve. (2004). *Do graduation tests measure up? A closer look at state high school exit exams*.

Ahn, T., & Vigdor, J. (2014). *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina* (NBER Working Paper No. 20511). National Bureau of Economic Research.

Airasian, P. W. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education, 95*(3), 393–412.

Alexander, L., James, H. T., & Glaser, R. (1987). *The Nation's Report Card: Improving the assessment of student achievement*. National Academy of Education.

Allensworth, E., & Easton, J. (2005). *The on-track indicator as a predictor of high school graduation*. The University of Chicago Consortium on School Research.

Alliance for Excellent Education. (2018). *N-size in ESSA state plans*.

Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Education Trust.

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448–452.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

American Institutes for Research. (2017). *Examining ESSA plans through the lens of research and practice: Reflections on state ESSA plans*.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*.

Ansari, A., & Pianta, R. C. (2019). School absenteeism in the first decade of education and outcomes in adolescence. *Journal of School Psychology, 76*, 48–61.

Armor, D. J., & Peiser, B. M. (1998). Interdistrict choice in Massachusetts. In P. E. Peterson & B. C. Hassel (Eds.), *Learning from school choice* (pp. 157–186). Brookings Institution Press.

Atchison, D. (2020). The impact of Priority School designation under ESEA flexibility in New York State. *Journal for Research on Educational Effectiveness, 13*(1), 121–146.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (NBER Working Paper No. 20657). National Bureau of Economic Research.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. w23478). National Bureau of Economic Research.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute.

Baker, O., & Lang, K. (2013). *The effect of high school exit exams on graduation, employment, wages, and incarceration* (NBER Working Paper 19182). National Bureau for Economic Research.

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007*. National Center for Education Statistics.

Bayer, P., Ferreira, F., & McMillan, R. (2007). *A unified framework for measuring preferences for schools and neighborhoods* (Working Paper No. 13236). National Bureau of Economic Research.

Bellwether Education Partners. (2017). *An independent review of ESSA state plans*.

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104.

Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal, 52*(5) 861–893.

Betebenner, D. W. (2008). *A primer on student growth percentiles*. National Center for the Improvement of Educational Assessment.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Betebenner, D. W., VanIwaarden, A., Domingue, B., & Shang, Y. (2023). SGP: Student growth percentiles & percentile growth trajectories (Version 2.1-0.0) [R package]. https://cran.r-project.org/web/packages/SGP/SGP.pdf

Billingham, C. M., & Hunt, M. (2016). School racial composition and parental choice new evidence on the preferences of White parents in the United States. *Sociology of Education, 89*(2), 99–117.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics, 114*(2), 577–599.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146–170.

Bleiberg, J. Brunner, E., Harbatkin, E., Kraft, M., & Springer, M. (2024). Taking teacher evaluation to scale: The effect of state reforms on achievement and attainment. *Journal of Political Economy: Microeconomics.* Advance online publication.

Blinder, A. (2015, April 1). Atlanta educators convicted in school cheating scandal. *The New York Times*. https://www.nytimes.com/2015/04/02/us/verdict-reached-in-atlanta-school-testing-trial.html

Bloom, B. S. (1968). Mastery learning. *Evaluation Comment, 1*(2), 1–11.

Bloom, B. S. (1973). *Cross-national study of educational attainment: Stage 1 of the I.E.A. investigation in six subject areas*. U.S. Department of Health, Education, & Welfare.

Bonilla, S., & Dee, T. (2020). The effects of school reform under NCLB waivers: Evidence from focus schools in Kentucky. *Education Finance and Policy, 15*(1), 75–103.

Bonilla, S., Dee, T. S., & Penner, E. K. (2021). Ethnic studies increases longer-run academic engagement and attainment. *Proceedings of the National Academy of Sciences, 118*(37), e2026386118. https://www.pnas.org/doi/10.1073/pnas.2026386118

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*, 231–268.

Boser, U., & Rosenthal, L. (2012). *Do schools challenge our students? What student surveys tell us about the state of education in the United States*. Center for American Progress.

Buckley, K. H. (2015). Separating the signal from the noise: An examination of student and teacher scores based on student learning objectives (SLOs) in one state. [Doctoral dissertation, Harvard University]. https://dash.harvard.edu/bitstream/handle/1/16461041/BUCKLEY-DISSERTATION-2015.pdf?sequence=1

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education, 85*(2), 147–162.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90.

Cannata, M. A., Smith, T. M., & Haynes, K. T. (2017). Integrating academic press and support by increasing student ownership and responsibility. *AERA Open, 3*(3), 1–13.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.

Castellano, K. E., & Ho, A. D. (2013a). *A practitioner's guide to growth models*. Council of Chief State School Officers.

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics, 38*(2), 190–215.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics, 40*(1), 35–68.

Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for high school graduation. *American Journal of Education, 98*(1), 1–34.

Center on Education Policy. (2008). *State high school exit exams: A move toward end-of-course exams*.

Center on Education Policy. (2009). *State high school exit exams: Trends in test programs, alternate pathways, and pass rates*.

Center on Education Policy. (2012). *State high school exit exams: A policy in transition*.

Chakrabarti, R. (2014). Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics, 110*, 124–146.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–2679.

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Discussion of the American Statistical Association's Statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy, 1*(1), 111–113.

Chudowsky, N., Kober, N., Gayler, K. S., & Hamilton, M. (2002). *State high school exit exams: A baseline report*. Center on Education Policy.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 311–329.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. U.S. Government Printing Office.

Corcoran, S., Jennings, J., & Beveridge, A. (2011, March 3–5). *Teacher effectiveness on high- and low-stakes tests* [Paper presentation]. Society for Research on Educational Effectiveness Spring Conference, Washington, DC, United States.

Council of Chief State School Officers. (1984). *Education evaluation and assessment in the United States*.

Cronbach, L. J. (2004). An interview with Lee J. Cronbach. In L. V. Jones & I. Olkin (Eds.), *The nation's report card: evolution and perspectives* (pp. 139–153). Phi Delta Kappa Educational Foundation.

Cronin, J., Kingsbury, G. G., McCall, M. S., & Bowe, B. (2005). *The impact of the No Child Left Behind Act on student achievement and growth*. Northwest Evaluation Association.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state standards-based assessment. *Educational Measurement, 12*(1), 1–22.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives, 22*(86), 1–34.

Data Quality Campaign. (2019). *Show me the data [Report]*. https://dataqualitycampaign.org/wp-content/uploads/2019/04/DQC-Show-Me-the-Data-04042019.pdf.

Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. *Educational Researcher, 44*(6), 347–358.

Dee, T. S. (2003). The "first wave" of accountability. In P. Peterson & M. West (Eds.), *No Child Left Behind? The politics and practice of accountability* (pp. 1–20). Brookings Institution.

Dee, T., & Dizon-Ross, E. (2017). *School performance, accountability and waiver reforms: Evidence from Louisiana* (NBER Working Paper No. 23463). National Bureau of Economic Research.

Dee, T. S., & Jacob, B. A. (2006). *Do high school exit exams influence educational attainment or labor market performance?* (NBER Working Paper 12199). National Bureau of Economic Research.

Dee, T. S., & Jacob, B. A. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*, 418–446.

Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis, 35*(2), 252–279.

Dee, T. S., & Penner, E. K. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal, 54*(1), 127–166.

Delaware Department of Education. (2010). *For the 2009–2010 school year: State accountability in Delaware*. http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics, 98*(5), 848–862.

Domaleski, C. (2011). *State end-of-course testing programs: A policy brief*. Council of Chief State School Officers.

Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Center for American Progress.

Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Thomas B. Fordham Institute.

Education Commission of the States. (2014). *Rating states, grading schools: What parents and experts say states should consider to make school accountability systems meaningful*.

Education Commission of the States. (2018). *50-state comparison: States' school accountability systems*.

Edwards, N. R., & Mindrila, D. L. (2019). Improving graduation rates: Legitimate strategies and gaming practices. *Education Policy Analysis Archives, 27*(41), 1–28.

Elledge, A., Le Floch, K. C., Taylor, J., & Anderson, L. (2009). *State and local implementation of the No Child Left Behind Act Volume V—Implementation of the 1 percent rule and 2 percent interim policy options*. U.S. Department of Education.

Elliott, S. N., & Bartlett, B. (2016). Opportunity to learn. In *Oxford handbooks online scholarly research reviews* (pp. 1–11). Oxford University Press.

Elmore, R. (2004). Conclusion: The problem of stakes in performance-based accountability systems. In S. Fuhrman & R. Elmore (Eds), *Redesigning accountability systems for education* (pp. 274 – 296). Teachers College Press.

English, D. (2017). *Proposed state accountability systems under the Every Student Succeeds Act: A summary of fall 2017 submissions*. American Institutes for Research.

Eren, O., Depew, N., & Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics, 153*, 9–31.

Eren, O., Lovenheim, M. F., & Mocan, N. H. (2018). *The effect of grade retention on adult grime: Evidence from a test-based promotion policy* (NBER Working Paper 25384). National Bureau for Economic Research.

Figlio, D. N., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. J. Machin, & L. Woessmann (Eds.), *Handbooks in economics: Economics of education* (Vol. 3, pp. 383–421). Elsevier.

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.

Figlio, D., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*(1–2), 239–255.

Figlio, D. N., Rouse, C. E., & Schlosser, A. (2009). *Leaving No Child Behind: Two paths to school accountability* [Working paper]. The Urban Institute.

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems*. edCount.

Fuchsman, D., Sass, T. R., & Zamarro, G. (2020). *Testing, teacher turnover and the distribution of teachers across grades and schools* (Annenberg Ed Working Paper No. 20–200). Brown University.

Gaddis, S. M., & Lauen, D. L. (2014). School accountability and the Black–White test score gap. *Social Science Research, 44*, 15–31.

Gewertz, C. (2020). Which states require an exam to graduate? *Education Week*. https://www.edweek.org/ew/section/multimedia/states-require-exam-to-graduate.html

Ginsburg, A. L., Noell, J., & Plisko, V. W. (1988). Lessons from the wall chart. *Educational Evaluation and Policy Analysis, 10*(1), 1–12.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*(8), 519–521.

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.

Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added*. Brown Center on Education Policy at Brookings.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review, 36*, 216–228.

Greene, J. P., & Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy, 2*(4), 319–340.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. National Education Goals Panel.

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis, 36*(4), 417–436.

Grodsky, E., Warren, J. R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971–2004. *Educational Policy, 23*(4), 589–614.

Hackmann, D. G., Malin, J. R., & Bragg, D. D. (2019). An analysis of college and career readiness emphasis in ESSA state accountability plans. *Education Policy Analysis Archives, 27*(160), 1–28.

Haderlein, S. A. K. (2022). How do parents evaluate and select schools? Evidence from a survey experiment. *American Educational Research Journal, 59*(2), 381–414.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary research and Perspectives, 11*(1–2), 1–18.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education, 104*(2), 1–34.

Haertel, E., & Ho, A. (2016). Fairness using derived scores. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 233–254). Routledge.

Hamilton, L. S. (2012). Measuring teaching quality using student achievement tests: Lessons from educators' responses to No Child Left Behind. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 49–75). Teachers College Press.

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.

Harris, D., & Herrington, C. (2004). *Accountability and the achievement gap: Evidence from NAEP* [Unpublished manuscript]. Department of Educational Leadership and Policy Studies, Florida State University.

Harris, D. N., & Larsen, M. F. (2017). *Demand, information, and the market for schooling: Evidence on revealed preferences from post-Katrina New Orleans*. Education Research Alliance for New Orleans.

Harris, D. N., Liu, L., Barrett, N., & Li, R. (2020). *Is the rise in high school graduation rates real? High-stakes school accountability and strategic behavior*. Brookings Institution.

Hastings, J. S., Kane, T. J., & Staiger, D. O. (2009). *Heterogeneous preferences and the efficacy of public school choice*. Yale University.

Hawaii State Department of Education. (2018). *KĀʻEO 2018 technical manual for grades 3-4 HLA, Math, and Science operational tests.* https://www.hawaiipublicschools.org/Reports/KAEO%20Technical%20Report_2018_Operational%20Tests.pdf

Heinrich, C. J., Meyer, R. H., & Whitten, G. (2010). Supplemental education services under No Child Left Behind: Who signs up, and what do they gain? *Educational Evaluation and Policy Analysis, 32*(2), 273–298.

Hemelt, S. W., & Jacob, B. (2020). How does an accountability program that targets achievement gaps affect student performance? *Education Finance and Policy, 15*(1), 45–74.

Hemelt, S. W., & Marcotte, D. E. (2013). High school exit exams and dropout in an era of increased accountability. *Journal of Policy Analysis and Management, 32*(3), 323–349.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education, 20*(1), 101–126.

Hernandez, D. J. (2012). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation*. Annie E. Casey Foundation.

Hernández, R. A. (2013). *Maintaining a focus on subgroups in an era of Elementary and Secondary Education Act waivers*. Campaign for High School Equity.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. National Academies Press.

Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal, 38*(2), 289–318.

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value-added models of student performance* (pp. 255–290). JAM Press.

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice, 26*(4), 11–20.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351–360.

Ho, A. D. (2014). Variety and drift in the functions and purposes of assessment in K–12 education. *Teachers College Record, 116*(11), 1–18.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.

Hoff, D. J. (2001). New standards' leaves legacy of unmet goals. *Education Week, 20*(43), 1.

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., & Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. U.S. Department of Education.

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation* (pp. 163–189). Information Age Publishing.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3–17.

Holme, J. J. (2002). Buying homes, buying schools: School choice and the social construction of school quality. *Harvard Educational Review, 72*(2), 177–206.

Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research, 80*(4), 476–526.

Hough, H., Byun, E., & Mulfinger, L. (2018). *Using data for improvement: Learning from the CORE Data Collaborative*. Policy Analysis for California Education.

Howell, W. G. (2015). Results of President Obama's Race to the Top. *Education Next, 15*(4), 58–67.

Hunter, S. B. (2019). New evidence concerning school accountability and mathematics instructional quality in the No Child Left Behind era. *Educational Assessment, Evaluation, and Accountability, 31*, 409–436.

Hutt, E., & Polikoff, M. S. (2020). Toward a framework for public accountability in education reform. *Educational Researcher, 49*(7), 503–511.

Institute of Education Sciences National Center for Education Evaluation and Regional Assistance. (2019). *Presenting school choice information to parents: An evidence-based guide*.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy, 126*(5), 2072–2107.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis, 23*(1), 99–121.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5–6), 761–796.

Jacob, B. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments*. National Bureau of Economic Research.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics, 86*(1), 226–244.

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics, 1*(3), 33–58.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843–877.

Jacobs, J. (2015). How is this fair? Art teacher is evaluated by students' math standardized test scores. *Washington Post Answer Sheet*. https://www.washingtonpost.com/news/answer-sheet/wp/2015/03/25/how-is-this-fair-art-teacher-is-evaluated-by-students-math-standardized-test-scores/

Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education, 121*(1), 1–27.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher, 43*(8), 381–389.

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.

Jerald, C. D. (2000). The state of the states 2000. *Education Week*. https://www.edweek.org/education/the-state-of-the-states-2000/2000/01

Ji, C. S., Yee, D. S. W., & Rahman, T. (2021). *Mapping state proficiency standards onto the NAEP scales: Results from the 2019 NAEP Reading and Mathematics assessments*. National Center for Education Statistics.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The Nation's Report Card: Evolution and perspectives*. Phi Delta Kappa International.

Judson, E. (2013). The relationship between time allocated for science in elementary schools and state accountability policies. *Science Education, 97*(4), 621–636.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Bill & Melinda Gates Foundation.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 9*(49), 1–22.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.

Koretz, D. (2015). Adapting the practice of measurement to the demands of test-based accountability. *Measurement: Interdisciplinary Research and Perspectives, 13*, 1–25.

Koretz, D. (2017). *The testing charade*. University of Chicago Press.

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. RAND.

Koretz, D., & Beguin, A. (2010). Self-monitoring assessments for educational accountability systems. *Measurement: Interdisciplinary Research and Perspectives, 8*(2–3), 92–109.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). American Council on Education/Praeger.

Koretz, D., Jennings, J. L., Ng, H. L., Yu, C., Braslow, D., & Langi, M. (2016). Auditing for score inflation using self-monitoring assessments: Findings from three pilot studies. *Educational Assessment, 21*(4), 231–247.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Center for the Study of Evaluation, University of California.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting "the widget effect": Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249.

Lachlan-Haché, L. (2015). *The art and science of student learning objectives: A research synthesis*. American Institutes for Research.

Ladd, H. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review, 18*, 1–16.

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice, 34*(3), 159–165.

Lauen, D. L. (2008). False promises: The school choice provisions in NCLB. In A. Sadovnik, J. O'Day, G. Bohrnstedt, & K. Borman (Eds.), *No Child Left Behind and the reduction of the achievement gap* (pp. 203–226). Routledge.

Lauen, D. L., & Gaddis, S. M. (2012). Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability pressure on student achievement. *Educational Evaluation and Policy Analysis, 34*(2), 185–208.

Lee, J., & Lee, M. (2020). Is "whole child" education obsolete? Public school principals' educational goal priorities in the era of accountability. *Educational Administration Quarterly, 56*(5), 856–884.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34*(1), 109–121.

LiCalsi, C., Ozek, U., & Figlio, D. (2019). The uneven implementation of universal school policies: Maternal education and Florida's mandatory grade retention policy. *Education Finance and Policy, 14*(3), 383–413.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*, 31–31.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind act of 2001. *Educational Researcher, 31*(6), 3–16.

Liu, J., Lee, M., & Gershenson, S. (2021). The short-and long-run impacts of secondary school absences. *Journal of Public Economics, 199*, 104441.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability, 31*(1), 61–95.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22–52.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Loeb, S., & Strunk, K. (2007). Accountability and local control: Response to incentives with and without authority over resource allocation and generation. *Education Finance and Policy, 2*(1), 10–39.

Lorié, W. (2020). *Addressing the alignment challenge associated with the use of college admissions tests under ESSA*. National Center for the Improvement of Educational Assessment.

Lyons, S., Johnson, M., & Hinds, B. F. (2021). *Confronting inequity in assessment*. Lyons Assessment Consulting.

Madaus, G. F., Stufflebeam, D., & Scriven, M. S. (1983). Program evaluation. In G. F. Madaus, D. Stufflebeam, & M. S. Scriven (Eds.), *Evaluation models* (pp. 3–22). Springer.

Mariano, L. T., & Martorell, P. (2013). The academic effects of summer instruction and retention in New York City. *Educational Evaluation and Policy Analysis, 35*(1), 96–117.

Mariano, L. T., Martorell, F., & Tsai, T. (2018). *The effects of grade retention on high school outcomes: Evidence from New York City schools*. RAND.

Marion, S. (2016). *Considerations for state leaders in the design of school accountability systems under the Every Student Succeeds Act*. National Center for the Improvement of Educational Assessment.

Marsh, J. A., & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. *Educational Management Administration and Leadership, 43*(2), 269–289.

Marsh, J. A., & Koppich, J. E. (2018). *Superintendents speak: Implementing the Local Control Funding Formula*. Policy Analysis for California Education.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. RAND.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*, 1332–1361.

McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Harvard University Press.

McDonnell, L. M. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education, 80*(2), 19–38.

McLaughlin, M. J., & Thurlow, M. (2003). Educational accountability and students with disabilities: Issues and challenges. *Educational Policy, 17*(4), 431–451.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge University Press.

National Academies of Sciences, Engineering, and Medicine. (2017). *Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2019). *Monitoring educational equity*. National Academies Press.

National Center for Learning Disabilities. (2018). *Assessing ESSA: Missed opportunities for students with disabilities*.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. U.S. Government Printing Office.

National Conference of State Legislatures. (2019). *Third grade reading legislation*. https://www.ncsl.org/research/education/third-grade-reading-legislation.aspx.

National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. U.S. Government Printing Office.

National Council on Measurement in Education. (2019). *National Council on Measurement in Education position statement on the use of college admissions test scores as academic indicators in state accountability systems*. Position statement. https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission_Statement_06-16-19.pdf

National Governors Association. (1990). *Educating America: State strategies for achieving the National Education Goals*.

National Research Council. (2011). *Incentives and test-based accountability in public education* (Committee on Incentives and Test-Based Accountability in Public Education, M. Hout & S. W. Elliott, Eds.). National Academies Press, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education.

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *The Journal of Economic Education, 44*, 339–352.

Neal, D. (2018). *Information, incentives, and education policy*. Harvard University Press.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics, 92*, 263–283.

NORC. (2015). *Alignment between the 2013 NAEP Grade 8 mathematics assessment and ACT EXPLORE mathematics assessment*.

Olson, L. (2019). *The new testing landscape: How state assessments are changing under the federal Every Student Succeeds Act*. FutureEd.

Ou, D. (2009). *To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam*. Centre for Economic Performance.

Özek, U. (2015). Hold back to move forward? Early grade retention and student misbehavior. *Education Finance and Policy, 10*(3), 350–377.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163–193.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis, 32*(1), 5–23.

Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational Researcher, 41*(3), 93–97.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. National Board on Educational Testing and Public Policy.

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher, 39*(2), 110–119.

Petrilli, M. J., Griffith, D., & Wright, B. L. (2016a). *High stakes for high schoolers: State accountability in the age of ESSA*. Thomas B. Fordham Institute.

Petrilli, M. J., Griffith, D., & Wright, B. L. (2016b). *High stakes for high schoolers: State accountability in the age of ESSA, Part II*. Thomas B. Fordham Institute. https://eric.ed.gov/?id=ED579520

Phillips, M., Reber, S., & Rothstein, J. (2018). *Making California data more useful for educational improvement*. Policy Analysis for California Education.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14.

Polikoff, M. S. (2012a). Instructional alignment under No Child Left Behind. *American Journal of Education, 118*(3), 341–368.

Polikoff, M. S. (2012b). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis, 34*(3), 278–294.

Polikoff, M. S. (2013). Teacher education, experience, and the practice of aligned instruction. *Journal of Teacher Education, 64*(3), 212–225.

Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal, 52*(6), 1185–1211.

Polikoff, M. S. (2016, July 12). A letter to the U.S. Department of Education. *On Education Research*. https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education/

Polikoff, M. (2022). Alignment. In S. Brookhart (Ed.), *Routledge resources online: Education*. https://doi.org/10.4324/9781138609877-REE4-1

Polikoff, M., & Korn, S. (2020). School accountability. In J. G. Dwyer (Ed.), *The Oxford handbook of children and the law* (pp. 521–550). Oxford University Press.

Polikoff, M. S., Korn, S., & McFall, R. (2018). *In need of improvement? Assessing the California Dashboard after one year*. Stanford University.

Polikoff, M. S., McEachin, A., Wrabel, S. L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher, 43*(1), 45–54.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal, 48*, 965–995.

Polikoff, M. S., & Struthers, K. S. (2013). Changes in the cognitive complexity of English instruction: The moderating effects of school and classroom characteristics. *Teachers College Record, 115*(8), 1–26.

Polikoff, M. S., & Wrabel, S. L. (2013). When is 100% not 100%? The use of safe harbor to make Adequate Yearly Progress. *Education Finance and Policy, 8*(2), 251–270.

Polson, C. (2018). TAKS-ing students? Texas exit exam effects on human capital formation. *Economics of Education Review, 62*, 129–150.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.

Porter, A. C., Linn, R. L., & Trimble, C. S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice, 24*(4), 32–39.

Porter, A. C., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher, 40*, 103–116.

Porter, A. C., & Polikoff, M. S. (2007, October). NCLB: State interpretations, early effects, and suggestions for reauthorization. *Social Policy Report, 21*(4), 1–15.

Princiotta, D. (2019). *Understanding the great U.S. high school graduation rate rise: 1998–2010* [Unpublished doctoral dissertation]. Johns Hopkins University

Quinn, D. M. (2020). Experimental effects of "achievement gap" news reporting on viewers' racial stereotypes, inequality explanations, and inequality prioritization. *Educational Researcher, 49*(7), 482–492.

Quinn, D. M., & Desruisseaux, T. M. (2022). Replicating and extending effects of "achievement gap" discourse. *Educational Researcher, 51*(7), 496–499.

Quinn, D. M., Desruisseaux, T. M., & Nkansah-Amankra, A. (2019). "Achievement gap" language affects teachers' issue prioritization. *Educational Researcher, 48*(7), 484–487.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82–90.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis, 32*(4), 498–520.

Reardon, S., Kalogrides, D., Ho, A., Shear, B., Fahle, E., Jang, H., & Chavez, B. (2021). Stanford Education Data Archive.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy, 6*(3), 207–241.

Rebell, M. A. (2008). *Moving every child ahead: From NCLB hype to meaningful educational opportunity*. Teachers College Press.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.) *Changing assessments: Alternative views of aptitude, achievement and instruction*(pp. 37–75). Springer.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

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

Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy, 2*(4), 119–147.

Roda, A., & Wells, A. S. (2013). School choice policies and racial segregation: Where White parents' good intentions, anxiety, and privilege collide. *American Journal of Education, 119*(2), 261–293.

Roderick, M., Kelley-Kemple, T., Johnson, D. W., & Beechum, N. O. (2014). *Preventable failure: Improvements in long-term outcomes when high schools focused on the ninth-grade year*. University of Chicago Consortium on School Research.

Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis, 27*(4), 309–340.

Roeber, E., Olson, J., Topol, B., Webb, N., Christopherson, S., Perie, M., Pace, J., Lazarus, S., & Thurlow, M. (2018). *Feasibility of the use of the ACT and SAT in lieu of Florida statewide assessments: Vol. 1. Final report*. Assessment Solutions Group.

Rose, D. H., Robinson, K. H., Hall, T. E., Coyne, P., Jackson, R. M., Stahl, W. M., & Wilcauskas, S. L. (2018). Accurate and informative for all: Universal Design for Learning (UDL) and the future of assessment. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices* (pp. 167–180). Springer.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. CSE-TR-566). National Center for Research on Evaluation, Standards, and Student Testing.

Royal, C., & Gibson, S. (2017). They schools: Culturally relevant pedagogy under siege. *Teachers College Record, 119*(1), 1–25.

Ryan, J. E. (2004). The perverse incentives of the No Child Left Behind Act. *NYU Law Review, 79*, 932–989.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation: Lessons learned from observations, principal-teacher conferences, and district implementation*. Consortium on Chicago School Research.

Schneider, M., Teske, P., Marshall, M., & Roch, C. (1998). Shopping for schools: In the land of the blind, the one-eyed parent may be enough. *American Journal of Political Science, 42*(3), 769.

Shultz, P. K., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education, 6*, 1–11. https://doi.org/10.3389/feduc.2021.701973

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics, 152*, 154–169.

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator, 45*(3), 28.

Shepard, L., Glaser, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels*. National Academy of Education.

Shepard, L. A., Marion, S. F., & Saldaña, C. M. (in press). Standards-based reform and school accountability. In L. Cohen-Vogel, P. Youngs, & J. Scott (Eds.), *AERA handbook of education policy research*. AERA.

Shultz, P. K., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education, 6,* 1–11. https://doi.org/10.3389/feduc.2021.701973

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Sinnema, C., & Robinson, V. (2007): The leadership of teaching and learning: Implications for teacher evaluation. *Leadership and Policy in Schools, 6*(4), 319–343.

Skinner, B. F. (1953). *Science and human behavior*. Macmillan.

Smith, E., & Tyler, R. (1942). *Appraising and recording student progress*. Harper & Row.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Taylor & Francis.

Smith, S. S., & Mickelson, R. A. (2000). All that glitters is not gold: School reform in Charlotte–Mecklenburg. *Educational Evaluation and Policy Analysis, 22*(2), 101–121.

Spillane, J. P. (2004). *Standards deviation: How schools misunderstand education policy*. Harvard University Press.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340–359.

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher, 46*(7), 378–396.

Steinberg, M. P., & Sartain, L. (2021). What explains the race gap in teacher performance ratings? Evidence from Chicago Public Schools. *Educational Evaluation and Policy Analysis, 43*(1), 60–82.

Stronge, J. H., & Tucker, P. D. (2003). *Teacher evaluation. Assessing and improving performance*. Eye on Education.

Sun, M., Saultz, A., & Ye, Y. (2017). Federal policy and the teacher labor market: Exploring the effects of NCLB school accountability on teacher turnover. *School Effectiveness and School Improvement, 28*(1), 102–122.

Toch, T. (1984). School chiefs endorse state-by-state comparisons. *Education Week, 9*, 12.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Education Sector.

Tucker, P. D. (1997). Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*, 103–126.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.

Tyner, A., & Larsen, M. (2019). *End-of-course exams and student outcomes*. Thomas B. Fordham Institute.

U.S. Department of Education. (2005, November 18). *Secretary Spellings announces growth model pilot, addresses chief state school officers' Annual Policy Forum in Richmond* [Press release].

U.S. Department of Education. (2006, January 25). *Peer review guidance for the NCLB Growth Model Pilot applications*. Washington, D.C.

U.S. Department of Education. (2009). *Race to the Top Fund. Final rule*. https://www.govinfo.gov/content/pkg/FR-2009-11-18/pdf/E9-27426.pdf

U.S. Department of Education. (2021). *U.S. Department of Education releases guidance to states on assessing student learning during the pandemic*.

Van Dunk, E., Meissner, D., & Browne, J. (1998). *Parental involvement and school choice: A look at private school choice in Cleveland and Milwaukee*. The Public Policy Forum.

The Victorian Quality Council. (2008). *A guide to using data for health care quality improvement*. Victorian Government Department of Human Services Rural and Regional Health and Aged Care Services Division.

Vinovskis, M. A. (2001). *Overseeing the Nation's Report Card: The creation and evolution of the National Assessment Governing Board (NAGB)*. National Assessment Governing Board.

Walter, M., Kukutai, T., Carroll, S. R., & Rodriguez-Lonebear, D. (2021). *Indigenous data sovereignty and policy*. Taylor & Francis.

Warren, J. R., & Grodsky, E. (2009). Exit exams harm students who fail them—and don't benefit students who pass them. *Phi Delta Kappan, 90*(9), 645–649.

Warren, J. R., Grodsky, E., & Lee, J. C. (2008). State high school exit examinations and postsecondary labor market outcomes. *Sociology of Education, 81*(1), 77–107.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Council of Chief State School Officers.

Webb, N. L. (2002, April 1–5). *An analysis of the alignment between mathematics standards and assessments for three states* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, United States of America.

Webb, N. L. (2005). *Alignment analysis of mathematics standards and assessments: Michigan Grades 3–8*. Michigan Department of Education.

Webb, N. L. (2010). *Alignment analysis of reading standards and assessments: South Dakota Grades 3–8 and 11, 2009 and 2010*. South Dakota Department of Education.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New Teacher Project.

West, M. R. (2012). *Is retaining students in the early grades self-defeating?* Center on Children and Families, Brookings Institution.

Winters, M. A. (2018). *The effect of Florida's test-based promotion policy on student performance prior to the retention decision*. Boston University.

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.

Winters, M. A., & Greene, J. P. (2012). The medium-run effects of Florida's test-based promotion policy. *Education Finance and Policy, 7*(3), 305–330.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness, 8*(2), 245–279.

Yeager, D., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. Carnegie Foundation for the Advancement of Teaching.

NOTES

1. Although the statute uses the terminology "valid and reliable academic measure," the current *Standards* (AERA et al., 2014) and Zwick (this volume) hold that validity is a property of uses of measures and not a property of measures themselves. A valid academic measure is thus one that has sufficient validity evidence to support its intended use in the accountability system.
2. This portion of the chapter is adapted from Polikoff (2022).
3. This portion of the chapter is adapted from Polikoff (2022).