WILLIAM H. ANGOFF MEMORIAL LECTURE SERIES

Research Report

Test-Based Accountability Systems:

The Importance of Paying Attention to Consequences

Suzanne Lane

ETS®

William H. Angoff
(1919–1993)

William H. Angoff was a distinguished research scientist at Educational Testing Service (ETS) for more than 40 years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text, "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible. The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. These lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory. The William H. Angoff Lecture Series reports are published by the Center for Research on Human Capital and Education, ETS Research and Development.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Test-Based Accountability Systems: The Importance of Paying Attention to Consequences

Suzanne Lane

University of Pittsburgh, Pittsburgh, PA

The impetus for test-based accountability systems is to improve the educational opportunities afforded to all students so as to improve their learning; therefore, integral to the validity argument of these systems is the appraisal of test-based inferences and decisions in terms of their consequences. Both positive and negative consequences of test-based decisions have different effects on different groups of students and in different schools, and these differential effects need to be examined as part of the validity argument and in addressing fairness issues. This paper addresses intended and potentially unintended consequences of test-based accountability systems in the validity argument. Legislation for test-based accountability systems, as well as studies evaluating their consequences, is discussed. A conceptual framework that provides a principled approach for evaluating both intended and unintended consequences of assessment and accountability systems, including those that arise due to using tests as policy levers for educational change, is provided.

Keywords validity; consequential evidence; test-based accountability; assessment; accountability systems

doi:10.1002/ets2.12283

Preface

The 17th William H. Angoff Memorial Lecture, *Test-Based Accountability Systems: The Importance of Paying Attention to Consequences*, was presented by Dr. Suzanne Lane, Professor in the Research Methodology Program at the University of Pittsburgh, at the National Press Club in Washington, DC, on October 16, 2019. Dr. Lane's research and professional interests focus on educational measurement and testing with an emphasis on design and validity issues in large-scale assessment programs as well as the effectiveness of education and accountability programs. Her work has appeared in educational measurement and testing journals.

In this paper based on her lecture, Dr. Lane addresses the positive and negative unintended consequences that can arise when assessments are used as part of accountability systems. She discusses the importance of validating test score inferences and uses for both instruction and accountability purposes, noting that the negative consequences typically affect students who are traditionally underserved. At stake, she says, are higher drop-out rates, lower enrollment in post-high school institutions of learning, a narrowing of topics taught, and an unbalanced ratio of teaching time versus time spent learning to take the test.

Dr. Lane compares and contrasts the tenets of the No Child Left Behind Act and the Every Student Succeeds Act. She also addresses the use of standardized tests for college entry as assessments of competency upon graduation from high school. Dr. Lane points out that one stakeholder's perceived value of a test-based decision often differs from and sometime conflicts with the perceived values of other stakeholders. In her conclusion, she stresses the need for a comprehensive evaluation of all possible consequences — negative and positive — of tests used in accountability systems.

ETS's Angoff Memorial Lecture Series was established in 1994 to honor the life and work of William H. Angoff, who died in 1993. For more than 50 years, Dr. Angoff made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Dr. Angoff's interests, this lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Ida Lawrence
Senior Vice President
ETS Research & Development

Corresponding author: S. Lane, E-mail: sl@pitt.edu

The intent of test-based accountability systems is to improve the educational opportunities afforded to all students by holding schools, educators, and, sometimes, students accountable. Tests serve as powerful tools for forwarding educational reforms while at the same time exposing societal and educational inequities; therefore, integral to validation efforts of tests used in these systems is the appraisal of test-based decisions and uses in terms of their consequences. The use of tests as policy levers bears on fairness and equity issues. Differential test outcomes for students living in economically disadvantaged communities, African American students, Hispanic students, and English learners have led to undesirable consequences, including narrowing the curriculum and instruction and increased high school drop-out rates, for the very same students that the tests were intended to support. An evaluation of the use of tests in accountability systems should address whether the intended positive consequences are achieved and potential unintended negative consequences are minimized.

When a test is used as an instrument of policy, its consequences must to be evaluated (Cronbach, 1982). The values inherent in the testing program need to be made explicit, and the consequences of the decisions and uses based on tests are an integral aspect of the validity argument (M. T. Kane, 1992). There is an inherent conflict in using tests for instruction and accountability purposes as test design and development decisions differentially affect the validity of using a test for both instruction and accountability purposes. The consequences of tests used in accountability systems typically have different impacts on different groups of students and in different schools and districts, and these differential impacts need to be examined as part of the validity argument (Lane & Stone, 2002).

The focus of this paper is on the use of tests and their consequences in test-based accountability systems. It includes three main sections: consequences for tests used for accountability purposes, legislation for test-based accountability systems and their consequences, and a conceptual framework for evaluating consequences.

Consequences for Tests Used for Accountability Purposes

Importance of an Interpretation/Use Argument and Validity Argument

Validation entails constructing and evaluating coherent arguments for and against proposed test interpretations and uses (Cronbach, 1988; M. T. Kane, 1992; Messick, 1989). A clear statement of the interpretations and uses of tests and resulting scores is critical in the validation of tests used for both instruction and accountability purposes. The argument-based approach to validity entails both an interpretation/use (IU) argument and a validity argument (M. T. Kane, 2006, 2013). An IU argument "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading to the observed performances to the conclusions and decisions based on the performances" (M. T. Kane, 2006, p. 7). It is important for the IU argument to extend beyond those conclusions and decisions based on student performance to decisions made when delineating the purposes and uses of tests as agents of educational change, such as using tests to serve as a tool to focus curriculum and instruction. IU arguments for tests used in accountability systems require the articulation of the claims about student performance and progress; curriculum and instruction focus; and student, educator, and school accountability. The articulation of both positive consequences and potentially unintended negative consequences is incorporated in the IU argument. It is important to recognize that some unintended consequences may be positive, but it is essential to delineate any potential negative consequences prior to implementing a test-based accountability system. It is these unintended negative consequences that typically adversely affect students who have been traditionally underserved, including students of color, students living in economically disadvantaged communities, English language learners, and students with cognitive disabilities.

The validity argument entails obtaining both logical and empirical evidence to evaluate the soundness of the claims and assumptions in the IU argument. It involves an overall evaluation of the plausibility of the proposed interpretations and uses of tests by providing a coherent analysis of the evidence for and against proposed interpretation and uses (American Educational Research Association [AERA] et al., 2014; Cronbach, 1988; M. T. Kane, 1992; Messick, 1989). The specification of the validity argument allows for the accumulation of evidence for and against score interpretations and uses. The validity of test use depends on the synthesis of the evidence for the evaluation of the IU argument (Haertel, 1999).

Theories of Action

IU arguments encompass a theory of action that specifies how the test decisions and uses are expected to lead to intended outcomes and any potential unintended consequences. Theories of action include the context in which a program is being

implemented, a description of the components of the program, what the program components intend to achieve and how they interact, and short-term and long-term outcomes. A theory of action for a test-based accountability system provides a framework for bringing together score meaning and impact claims, with an emphasis on claims about the intended consequences or outcomes of the test on both individuals and schools (Bennett, 2010). Evidence needs to be obtained and synthesized to support the theory of action, and the validity argument and any contradictions to the specified claims need to be identified and evaluated. Bennett (2010) proposed the following features to be included in a theory of action for an assessment system:

- the intended effects of the assessment system,
- the components of the assessment system and a logical and coherent rationale for each component, including backing for that rationale in research and theory,
- the interpretive claims that will be made from assessment results,
- the action mechanisms designated to cause the intended effects, and
- potential unintended negative effects and what will be done to mitigate them (Bennett, 2010, p. 71).

States are required to delineate a theory of action for their assessment and accountability programs, and an evaluation of whether the intended outcomes are achieved, while minimizing any negative consequences, is necessary to support the use of the assessments as agents of educational change and the use of scores for accountability and instructional purposes.

Consequences in the Argument-Based Approach to Validity

The consideration of consequences in the evaluation of the validity of test score interpretations and uses is not new, although there has been a lively debate by scholars regarding whether consequences should be embraced as an integral aspect of validity. Scholars such as Cronbach, M. T. Kane, Linn, Messick, and Shepard argued that consequences are integral to the validity argument. In fact, nearly five decades ago, Cronbach (1971) considered evaluation of decision and actions based on tests as part of validity evaluation. Linn (1997) argued that "the evaluation of consequences rightly belongs in the domain of validity" (p. 16) and the "best way of encouraging adequate consideration of major intended positive effects and plausible unintended negative effects of test use is to recognize the evaluation of such effects as a central aspect of test validation" (p. 16). Similarly, Shepard (1997) argued that the inclusion of the soundness of decisions based on test scores warrants the need to examine consequences that are a "logical part of the evaluation of test use, which has been an accepted focus of validity for several decades" (p. 5). An examination of consequences in the validity argument involves a critical evaluation of both the intended positive and potential unintended negative consequences of test use (M. T. Kane, 1992, 2006; Lane et al., 1998; Linn, 1997; Shepard, 1993, 1997; Sireci, 2016). It is essential that we attend closely to potential negative consequences, particularly for students who have been historically underserved. As articulated by Campbell (1979) and commonly referred to as "Campbell's Law," 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. X).

The validity argument should address differential impact given that the consequences of tests used in accountability systems, both positive and negative, are likely to have different impacts for different groups of students and in different schools (Lane, 2014; Lane & Stone, 2002). Judgments about the fairness of tests bear on their uses and consequences. A well-documented, unintended negative consequence for state tests is the narrowing of instruction for some students, typically students living in economically disadvantaged areas, to those topics measured by the tests (Stecher, 2002) and a focus on students slightly below the cut point for proficient (Booher-Jennings, 2005).

The evaluation of the congruency between the intended interpretation and uses of test scores and the actual or enacted interpretations and uses of test scores is integral to the validity argument. The validity of both social and personal consequences is jeopardized if there is a lack of congruency between intended and enacted interpretations and uses. Enacted interpretations and uses by educators and administrators are more complex and varied than intended interpretations and uses (Coburn & Turner, 2011; Moss, 2016). Tests are used locally in different contexts by educators and administrators to address a variety of their own needs (e.g., grade promotion), and these local uses need to be understood and evaluated in terms of their appropriateness and potential negative consequences for students.

Social Consequences and Value Judgments

Tests are used as mechanisms of social action, resulting in consequences from direct and indirect actions. As indicated by Cronbach (1988), the validity argument "must link concepts, evidence, social and personal consequences, and values" (p. 4). Although consequences of test use have been a part of the evaluation of test decision procedures, social consequence, or adverse impact, was not addressed until the 1960s. In a Messick, 1964 conference paper, Messick reflected on the ethical aspects of whether a test should be used for a particular purpose that requires a justification of the proposed use of the test in terms of social values. Messick's (1989, 1994) perspective on adverse social consequences as having a role in test validation was limited, however, to them being traced back to construct-irrelevant components in scores and to construct underrepresentation: "If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced … then the validity of the test use is not overturned" (Messick, 1989, p. 88). M. T. Kane (2006) argued that Messick's perspective gave consequences a secondary role in validation and supported Cronbach's (1971, 1988) position that consequences have a more integral role in the validity argument and that negative consequences could invalidate test use even if they cannot be traced to a test flaw: "Tests that impinge on the rights and life changes of individuals are inherently disputable" (Cronbach, 1988, p. 6). Those responsible for collecting evidence to support the validity argument have an obligation to evaluate whether a use has appropriate consequences for both individuals and institutions and, more importantly, "to argue against adverse consequences" (Cronbach, 1988, p. 6). Consequences of testing are an integral part of validity and extend beyond consequences due to "invalidity of tests." As indicated by M. T. Kane (2013), consequences that have "potential for substantial impact in the population of interest, particularly adverse impact and systemic consequences" (p. 61) should be a central aspect of the validity argument. The consideration of social consequences of testing programs is an inherent aspect of the IU and validity argument.

Decisions made in test design, development, and use are grounded in value judgments (Messick, 1989; M. T. Kane, 2006). Value judgments are involved in the articulation of the intended purposes and uses of a test, the specification of the construct and content standards to be assessed, the design and development of items and scoring rules to measure those content standards, the development of performance level descriptions, the development of performance standards for determining student proficiency, and the design of score reports and accompanying interpretive guides. The value judgments inherent in each of these test design decisions will have an impact on the validity of test score interpretations and the consequences of test use, such as what knowledge and skills are emphasized in curriculum and instruction and what knowledge and skills are needed to be considered proficient. Value judgments are inherent in the specification of the intended positive consequences and the potentially unintended, negative consequences, such as using a test in consideration of high school graduation and its differential effect on different schools and different groups of students. Decision procedures always rest on value judgments; therefore, values need to be made explicit when articulating the consequences of decision rules that need to be evaluated (M. T. Kane, 2006).

Different stakeholders, including policymakers, advocacy groups, administrators, educators, parents, students, business leaders, and the community, have different, and sometimes conflicting, perspectives on the purposes and uses of tests in instruction and accountability systems. What is considered of value with respect to test-based decisions, uses, and consequences depends on the stakeholder. As suggested by Bachman (2002), different IU arguments could account for the conflicting values of different stakeholders. The specification of different IU arguments for the various stakeholders may be impractical. Nevertheless, the articulation of the competing values from various stakeholders is warranted so as to better understand the potential consequences, both positive and negative, of test use.

Test-Based Accountability Systems and Consequences

Policies are implemented in educational reforms with the intent that such policies will result in certain consequences, including improving student achievement and learning, enhancing curriculum and instruction, narrowing the achievement gaps, and preparing students for college and careers. Tests used in accountability systems provide policy makers with tools to hold schools, administrators, educators, and students accountable for student achievement and learning. Tests appeal to policy makers as agents of educational change because they are relatively inexpensive, can be mandated externally and implemented relatively quickly, and have visible results (Linn, 2000). Advocates view test-based accountability systems as a way to raise academic standards for students and to focus instruction on content aligned to rigorous

content standards. The intent of test-based accountability systems is for all students to be exposed to rigorous content standards and to achieve success on acquiring the knowledge and skills reflected in those standards.

There is an inherent tension, however, between using tests to achieve the goals of instruction and using them to measure accountability due to practical constraints in the design of assessment and accountability systems (Lane & DePascale, 2016). Design decisions are made that will affect the different goals of the tests and systems, and these design decisions should be informed by an explicit delineation of the multiple purposes and uses of the test (e.g., student, educator, and school accountability; instructional support) and the intended consequences, including social consequences and potential unintended negative consequences, of these conflicting uses. When a test is used as an instrument of policy, its consequences need to be evaluated (Cronbach, 1982).

In evaluating the effectiveness of tests used for accountability purposes, attention needs to be paid to how well the tests are achieving their goals in terms of the motivation and practices of school administrators and educators and in terms of improved student performance on the state content standards (Lane & Stone, 2002). Those who mandate and use tests are obliged to make a case for the appropriateness of the decision rules in the context in which they are being used, and any potential unintended negative consequences need to be weighed against intended positive consequences before using a test as an instrument for policy. The articulation and documentation of intended and potential unintended consequences should occur as the test purposes and claims are articulated and documented. A systematic evaluation of how the test achieves its goals and the cost of attempting to obtain its goals is needed. The relevant stakeholders should be involved in deciding on the questions that should be investigated when evaluating consequences of test use because of their different perspectives on the value of the positive and potentially negative consequences.

A principled approach to the evaluation of consequences is needed so we do not continue the practice of "begging the question;" that is, accepting part of the argument without serious evaluation (M. T. Kane, 2006). The claim that tests will lead to better instruction and student achievement is typically taken at face value. It is important that those who mandate tests, develop tests, and use tests adhere to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), and as Standard 1.6 states, "When a test use is recommended on the grounds that testing or the testing program itself will result in some indirect benefit, in addition to the utility of information from interpretation of the test scores themselves … logical or theoretical arguments and empirical evidence for the indirect benefit should be provided" (p. 24). The comment following Standard 1.6 further acknowledges that "certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices" (AERA et al., 2014, p. 24); therefore, the evaluation of whether such testing programs have the intended positive effects on instruction, or potentially negative unintended effects on instruction, is integral to the IU argument and validity argument.

Standards for Educational and Psychological Testing (AERA et al., 2014) also indicates that the extent to which "individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test has several implications for test scores for their intended uses" (p. 56). Validity and fairness of test score interpretations and uses as well as the use of tests as vehicles for educational change may be compromised due to inequities in education and, therefore, can lead to negative consequences. As an example, the use of tests for high school graduation has resulted in negative consequences for students who have not had the opportunity to learn the content and skills measured by the test.

Research on the Consequences of State Performance Assessments

Two types of test-based accountability systems that should have a substantial impact on instruction and learning are performance assessments and alternate assessments for students with the most significant cognitive disabilities. The renewed interest in performance assessments in the early 1990s was, in part, because performance assessments were considered to be valuable tools for educational reform and they were more direct measures of rigorous content standards. Performance assessments help shape instructional practice by providing an indicator to educators of what is important to teach and to students of what is important to learn. Performance assessments have instructional value, are aligned to the learning goals and content standards in academic disciplines, and serve as powerful professional development tools, particularly if teachers are involved in the design of the assessment and scoring of student performances (Lane & DePascale, 2016; Lane & Stone, 2006).

Linn (1993) argued that consequential evidence in support of the validity argument is "especially compelling for performance-based assessments … because particular intended consequences are an explicit part of the assessment

system's rationale" (p. 6). Research studies have demonstrated that the implementation of large-scale performance assessments in the 1990s was related to positive changes in instruction and student learning, with a greater emphasis on problem-solving, critical thinking, and reasoning. My colleagues and I examined validity evidence for the consequences of the Maryland School Performance Assessment Program (MSPAP), which was composed entirely of performance tasks that were integrated across content domains and required collaborative problem-solving for some of the tasks. Using growth model analyses we found a positive impact of the performance assessment and accountability system on both student learning and classroom instruction (Lane et al., 2002; Parke & Lane, 2008; Parke et al., 2006; Stone & Lane, 2003). Teacher reported use of performance-based instruction accounted for differences in school performance on MSPAP in reading, writing, mathematics, and science. In particular, schools with a focus on performance-based instruction, such as the engagement of students in critical thinking and reasoning, had higher MSPAP scores than schools in which their instruction was less performance-based (Lane et al., 2002; Stone & Lane, 2003). The use of more reported reform-oriented problem types in instruction had a significant impact on rates of change in MSPAP school performance in reading and writing; the more reported impact MSPAP had on instruction, including a focus on higher level thinking and reasoning skills and rigorous content, the greater rates of change in MSPAP school performance in mathematics and science were over a 5-year period (Lane et al., 2002; Stone & Lane, 2003). A systematic evaluation of mathematics instructional artifacts obtained from the classrooms revealed that they were more aligned with MSPAP and the state content standards for higher achieving schools and higher gain schools as compared to lower achieving schools and lower gain schools. It is important to note however that classroom assessments had a relatively weak alignment with the nature and content of MSPAP and the state content standards (Parke & Lane, 2008). MSPAP had a greater effect on performance-based instruction as opposed to performance-based classroom assessment practices, suggesting the need for better support of teachers in classroom assessment practices.

The MSPAP results pertaining to mathematics were supported by a study by Linn, Baker, and Betebenner (2002). They found that trends in mathematics student gains for state National Assessment of Educational Progress (NAEP) and MSPAP mathematics assessments were similar, suggesting that increased student performance on MSPAP was most likely a result of actual gains in student achievement in mathematics across the school years. The student performance gains on MSPAP were not limited to the format and content of the performance assessment. Gains on MSPAP could be extrapolated to other assessments that measure somewhat different content using different item formats. Such positive results may have resulted, in part, from schools using MSPAP data along with other information to guide instructional planning (Michaels & Ferrara, 1999).

When using test scores to make inferences regarding the quality of education, contextual information is needed to inform the inferences and resulting actions (Haertel, 1999). In the MSPAP studies, a school contextual variable, socioeconomic status (SES), which was measured by the percent of students receiving free or reduced-price lunch, was significantly related to school level performance on MSPAP in mathematics, reading, writing, science, and social studies (Lane et al., 2002; Stone & Lane, 2003). SES however was not significantly related to school performance gains on MSPAP in mathematics, writing, science, and social studies. It could be argued that MSPAP did not have an adverse impact for students living in economically disadvantaged areas because of comparable growth on MSPAP school scores in these subjects regardless of the school SES status. One could argue however that greater MSPAP gains for schools located in economically disadvantaged areas should have occurred given that the initial MSPAP performance was lower for these schools and a greater emphasis was needed in ensuring these students were provided with enriched instruction.

Other state assessment programs that included performance tasks in the early 1990s provided some additional evidence of the impact of performance assessments on instruction and student learning. Teachers in Vermont reallocated instruction time to reflect the goals of the Vermont Portfolio Assessment in mathematics and writing, including allocating more time to problem-solving and communication in mathematics and providing students opportunity to engage in extended writing projects (Koretz, Barron, Mitchell, & Stecher, 1996; Stecher & Mitchell, 1995). Teachers in Washington reported that both short-response items and extended-response items on the Washington Assessment of Student Learning were influential in improving instruction and student learning (Stecher, Barron, Chun, & Ross, 2000), and teachers used mathematics and writing scoring rubrics in instruction in a way that reinforced meaningful learning (Borko et al., 2001). In examining the relationship between improved instruction and gains in Kentucky school-level scores, Stecher, Barron, Kaganoff, and Goodwin (1998) found inconsistent findings across disciplines and grades.

A positive relationship however was found between standards-based instructional practices in writing and the Kentucky direct writing assessment. For example, more seventh-grade writing teachers in high-gain schools versus low-gain schools reported integrating writing with other subjects and increasing instructional emphasis on the process of writing.

Research on the Consequences of Alternate Assessments

The motivating force behind alternate assessments is to improve the educational opportunities, including increased access to general educational curriculum and inclusive instructional classrooms, for students with the most severe cognitive disabilities. Marion and Perie (2009) argued that consequential evidence is of particular importance for evaluating the validity of score interpretation and use for alternate assessments and the use of these assessments as educational change agents because these assessments provide a mechanism to promote grade-level academic instruction for students who are typically underserved. In examining the consequences of alternate assessments in the 2000s, Kleinert and Kearns (2001) reported that teachers from schools with high student performance on the alternate assessment indicated that the assessments had a positive impact on instruction and the inclusion of students with disabilities in classrooms with students without disabilities, whereas teachers from schools with low performance indicated that the alternate assessment had little impact. Towles-Reeves et al. (2006) evaluated the impact of the use of an alternate assessment on the development of instructional educational plans (IEPs). They found that the impact of the alternate assessment on IEP development was significantly less than its impact on instruction, which may have been due to the lack of an explicit connection between the alternate assessment and IEP development. The teachers who reported that the alternate assessment had no or little impact on their instruction also indicated that the assessment was not important to them, they lacked support for the implementation of the assessment, and they did not know how to use the assessment results to inform IEP development.

Roach, Elliott, and Berndt (2007) examined consequential evidence for the validity of the Wisconsin alternative assessment. It was expected that performance on the assessment and teacher's indication of student access to the general curriculum would account for the majority of the variance on teachers' perception of the alternate assessment. The only predictor, however, that accounted for the variance on the teachers' perception of the assessment was student grade level. Teachers reported less positive perceptions about the assessment as students progressed through the grades. More specifically, 10th-grade teachers were more skeptical about the meaningfulness of the alternate assessment scores and their use for instructional planning. An implication of these results is that teachers would benefit from more professional development on how to provide an inclusive instructional environment and standards-based instruction to students with the most severe cognitive disabilities. The assessment consortium that was established under Race to the Top (RTTT; U.S. Department of Education, 2009) recognized this need and made it central in their alternate assessment and instruction systems.

Legislation for Test-Based Accountability Systems and Consequences

Tests used at the national and state level have undergone substantial changes since the 1970s when the minimum competency testing (MCT) reform gained traction to restore confidence in the high school diploma. Although the goal of MCT reform was admirable — all students attain certain standards of minimum competency — it led to a number of undesirable consequences, including shifts in curriculum and instruction to focus on lower level skills and an increase in high school dropout rates, especially for those students living in economically disadvantaged communities. During the period of MCT and leading into the early 1990s, states were using tests for accountability purposes, with high-stakes for both educators and administrators. The 1990s led to standards-based accountability systems, and some states, such as Maryland, focused on the use of performance assessments that allowed for measuring cognitively demanding content and skills. This section describes several test-based accountability policies since the 2000s and the consequences of their implementation.

No Child Left Behind Act

The use of performance assessments declined in the early 2000s with the increased demand on testing, the need to provide individual-level scores, and limited resources under No Child Left Behind (NCLB, 2002). The intent of the NCLB Act, which was in effect from 2002 to 2015, was to create incentives for educators and students to focus on rigorous content in mathematics, English language arts (ELA), and science, and to have all students reach proficiency by

Table 1 Proficiency Under No Child Left Behind for Grade 8

State	Percent of students at proficient AYP starting point on state test	Percent of students at proficient or above on NAEP
Arizona	7	21
North Carolina	75	30

Note. AYP = adequate yearly progress; NAEP = National Assessment of Educational Progress.
Source: Linn (2003).

setting challenging performance standards. State tests were used as the primary measure of student outcomes, providing individual student scores and aggregate scores to hold schools, educators, and students accountable. Although its intent to focus instruction on rigorous content was admirable, when enacted, the emphasis tended to be on school, teacher, and student accountability.

A primary goal of NCLB was for schools to focus on and improve the education of low-achieving students and to reduce the achievement gaps between White students and African-American, Hispanic, and Native American students — students who were traditionally "left behind" or underserved in the educational system. In an effort to encourage schools to focus instruction on these groups, rewards, such as monies for after-school programs and teacher pay-for-performance programs, and sanctions, such as the loss of federal Title 1 funding and school takeovers, were attached to student performance and led to negative consequences for students, educators, and administrators.

States were required to annually report students' proficiency in ELA and mathematics in Grades 3 – 8 and once in high school using state summative tests. The tests were used to evaluate whether schools made adequate yearly progress (AYP) toward 100% of students being proficient or above in ELA and mathematics by 2013 – 14 and to provide rewards and sanctions based on each school's AYP status. These polices led to pressure to raise test scores, resulting in unintended negative consequences, including narrowing the curriculum and shifting resources away from those subjects that were not tested, focusing instruction in mathematics and ELA to content measured by the tests, replacing instruction with days of test preparation, and focusing on students just below the proficiency cut scores.

NCLB allowed states to provide their own definitions of proficient with the requirement that all students needed to achieve proficiency on the state tests by 2013 – 14. In addition to states having their own definition of proficient, they also had their own starting point for evaluating AYP. The starting point indicated the percentage of students who were proficient in 2001 – 02. As illustrated by Linn (2003), some states had the proficient performance standard set close to the 70th percentile, which is similar to NAEP's proficient achievement level and is considered challenging. As an example, for Grade 8 in mathematics, Arizona had only 7% of students at the proficient starting point in 2002 and North Carolina had 75%, whereas the NAEP results revealed a much smaller discrepancy between the states with 21% of the students in Arizona and 30% of the students in North Carolina at the proficient level or above (see Table 1). States who set lower starting points were at an advantage for meeting AYP than states who had a high percent of students proficient at the starting point; however, the trustworthiness of the state results is debatable given the state NAEP results.

These results reveal the inherent problem in using state-developed content standards, state-developed tests, and state-defined performance levels for accountability purposes and for comparing state performance. The use of state-defined performance standards for accountability and for state comparisons was summarized by Linn (2003): "The variability in the percentage of students who are labeled proficient or above due to the context in which the standards are set, the choice of judges, and the choice of method to set the standards is, in each instance, so large that the term proficient becomes meaningless" (p. 12). This unintended consequence, in part, led to a shift from the requirement of states to achieve AYP to the requirement of states to identify the lowest performing 5% of schools that would benefit from assistance in 2012.

NCLB Act and Consequences

In a study examining the effects of NCLB in three states, teachers reported that NCLB led to improvements in academic rigor and instruction as well as a focus on student learning; however, teachers also reported some negative features and consequences of NCLB, including content standards that were too difficult for some students but not challenging enough for other students and more instructional time spent on content standards that were tested and less instructional time

on content standards that were not tested (Stecher et al., 2008). Schools and districts reported improving the alignment of curriculum to the content standards and the state test, using data to inform instruction and focusing instruction on low-achieving students. Both teachers and administrators indicated that lack of time and resources had a negative impact on their efforts to improve instruction and student learning.

A review of research on the implications of test-based incentives for schools, teachers, and students indicated that although test-based incentive programs are associated to some extent with increased student achievement, they did not close the gap between student performance in the United States and the highest performing countries (National Research Council [NRC], 2011, p. 4.). In evaluating the results for tests attached to incentives, it was found that schools with NCLB-like school-level incentives had the largest estimates of student achievement effects (.08 *SD*), with the largest effects occurring in mathematics in elementary grades. The use of high school exit exams tended to decrease the rate of high school graduation, an undesirable negative consequence, although for programs that provided rewards for graduation, the use of exit exams tended to be related to an increase in the rate of high school completion (NRC, 2011, pp. 4–5).

A study conducted by Dee and Jacob (2011) examined whether NCLB had an impact on student achievement on state assessments. Using a comparative interpretive time series analysis (compared groups who have accountability polices in place and those that did not prior to NCLB), state assessment data, and NAEP state data, the researchers found increases in the average mathematics performance of students in Grade 4 (effect size = .23 by 2007) across the score scale, with slightly larger effects among Black and Hispanic students, and a smaller, but not statistically significant effect in Grade 8 mathematics performance, with a larger effect for Hispanic students. There was no evidence of increased reading performance in Grade 4 (Dee & Jacob, 2011). The positive results for mathematics would have been strengthened if there had been evidence of meaningful changes in the content of mathematics instruction for these students.

Race to the Top

The RTTT grant under the American Recovery and Reinvestment Act of 2009 called for tests to be grounded in academic standards that reflect 21st-century skills, including "cognitively challenging skills that are difficult to measure" and designed to create "conditions for education innovation and reform" (U.S. Department of Education, 2009, p. 2). RTTT required states to provide a theory of action for their test-based accountability systems, including the intended benefits and outcomes of the system (U.S. Department of Education, 2010). Specifically, the call required a "theory of action that describes in detail the causal relationships between specific actions or strategies … and its desired outcomes … including improvements in student achievement and college- and career-readiness" (U.S. Department of Education, 2010, p. 1,817). The assessments were to "play a critical role in educational systems; provide administrators, educators, parents, and students with the data and information needed to continuously improve teaching and learning" (U.S. Department of Education, 2009). To evaluate these desirable intentions, it was necessary for states to specify how they will occur and the mechanisms that will bring about the intended outcomes.

RTTT encouraged states to join one or more of the assessment consortia that developed assessments aligned to the Common Core State Standards (CCSS), and it supported measuring teacher effectiveness using the state test results. The goal of the two general assessment consortia, Partnership for Assessment of Readiness for College and Careers (PARCC, 2010) and Smarter Balanced (2015a, 2015b), was to develop assessments grounded in rigorous, rich content measuring high-level literacy, mathematical reasoning, and problem-solving using innovative item formats.

States adopted the CCSS or other rigorous content standards to focus instruction on challenging content and used one of the consortium tests, PARCC or Smarter Balanced, as their state assessment or developed their own test to signal whether students were college-and-career ready or on track. Table 2 provides an indication of the rigor of state content standards and performance standards during this time using the NAEP state results as a yardstick (Achieve, 2015, 2016, 2018). Overall, the consistency in the percentage of students proficient or above on both NAEP and state tests increased since the implementation of RTTT (2014–15), allowing for more confidence in using state test scores for instructional and accountability purpose and comparing the performance of states. As an example, for Grade 8 mathematics, the discrepancies between the percentage proficient or above on NAEP as compared to the state tests were considerable in 2013 prior to the implementation of RTTT, with 13 states' discrepancies ranging from −31 to −53, indicating that these 13 states had between 31% and 53% more students classified as proficient or above by the state test as compared to NAEP. Another 17 states had discrepancies ranging from −16 to −30, indicating that these 17 states had between 16% and 30%

Table 2 Grade 8 Mathematics State Test Performance and National Assessment of Educational Progress (NAEP) Performance

% Proficient or above discrepancy	2013–14 state data & 2013 NAEP data (Achieve, 2015)	2014–15 state data & 2015 NAEP data (Achieve, 2016)	2016–17 state data & 2017 NAEP data (Achieve, 2018)
−53 to −31 in 2013 −43 to −31 in 2015 −51 to -31in 2017	13	3	3
−30 to −16	17	4	5
−15 to −1	12	22	24
0 to +10 in 2013 0 to +22 in 2015 0 to +17 in 2017	2	16	19

more students classified as proficient or above by the state test as compared to NAEP, and for only two states were the discrepancies reversed, indicating that only two states had a larger percentage of students, between 0% and 10%, classified as proficient or above by NAEP as compared to the state test.

As a result of the adoption of the CCSS or other more rigorous content standards by states; the implementation of Smarter Balanced, PARCC, or other more cognitively demanding state tests; and the setting of more challenging performance standards, in 2015 fewer states exhibited such large discrepancies. Only three states' discrepancies ranged from −31 to −43, indicating that these three states had between 31% and 43% more students classified as proficient or above by the state test as compared to NAEP. Only four states' discrepancies ranged from −16 to −30, indicating that these four states had between 16% and 30% more students classified as proficient or above by the state test as compared to NAEP. In contrast to the results prior to the implementation of RTTT, 16 states had a higher percentage of students achieving proficiency or above on NAEP as compared to their state tests. In fact, one state had 22% more students achieving proficient or above on NAEP as compared to its state test. The 2017 discrepancy results followed a pattern similar to the 2015 results.

PARCC, Smarter Balanced, and other state tests administered since 2015 assess more challenging content standards and set more rigorous performance standards as compared to state tests administered prior to 2015. NAEP is used as a yardstick in state comparisons because it is grounded in assessment frameworks that reflect rigorous content. It can be argued, however, that differences between the state tests and NAEP results are partly due to differences in what is assessed by the different tests. A study that examined the alignment between the NAEP mathematics assessment framework and the CCSS found that there were no "wide areas" of content that were not aligned; however, there were differences in "specificity and conceptual understandings" (Hughes et al., 2013, p. 11). As an example, the researchers indicated that there is more rigorous content in Grade 8 algebra and geometry in the CCSS than in the NAEP mathematics assessment frameworks. Overall, however, the results in Table 2 provide support that states are raising their standards.

Race to the Top and Consequences

In a study examining the consequences of the CCSS and Smarter Balanced and PARCC on curriculum and instructional practices in five states, T. J. Kane et al. (2016) found that there was a relationship between student performance on PARCC and Smarter Balanced for several mathematics CCSS implementation strategies used by educators in Grades 4 through 8 and administrators. Test performance was related to a number of instructional features, including teachers obtaining explicit feedback on the implementation of the CCSS after their instruction was observed, number of days of teacher professional development on the CCSS, and the inclusion of student PARCC or Smarter Balanced scores in teacher performance evaluations; however, there was no relationship between student test performance and teacher reported shifts in instruction and materials.

An unintended consequence of these test-based accountability polices was the rise of the opt-out movement where parents do not allow their children to take the state test. The movement gained national recognition by 2015 when 20% of students in New York opted out of the state test. Concerns about evaluating teachers using student test performances and gains were prevalent. In 2015, the U.S. Department of Education encouraged states to review their testing programs to reduce the amount of testing time so that no more than 2% of instructional time was spent on testing. Shortening state tests, modifying academic standards, modifying graduation requirements, and modifying policies that use the tests to evaluate teachers have contributed to a decrease in the opt-out rate.

Alternate Assessments Under RTTT

Under RTTT, two assessment consortia for the students with the most severe cognitive disabilities were formed: National Center and State Collaborative (NCSC) and Dynamic Learning Map (DLM). The primary goal of alternate assessments for students with the most severe cognitive disabilities is to focus instruction on academic content and to improve academic learning for these students. The theory of action for the NCSC alternate assessments treated the summative assessment as a component of the overall system and was considered in light of the other system components (e.g., professional development for teachers, appropriate communication methods for the student and teacher, instruction aligned to grade-level content standards) when evaluating the system goals (Quenemoen et al., 2013). The long-term outcomes that were related to student achievement of academic content aligned to the CCSS included greater exposure to grade-level academic curriculum, which in turn contributed to students achieving increasingly higher academic outcomes and students leaving high school ready to participate in college, careers, and community. Within the theory of action for NCSC were the following assumptions related to curriculum, instruction, and professional development underlying the long-term outcomes:

- teachers are given resources for and training on instruction in academic knowledge and skills needed for college, career, and community readiness;
- teachers have the knowledge, skills, and orientation necessary to access the standards and provide academic instruction; and
- teachers have the resources, training, and supports necessary to develop symbolic language and build communicative competence with students (Quenemoen et al., 2013, p. 6).

The theory of action for DLM identified short-term outcomes, intermediate outcomes, and long-term outcomes. Short-term outcomes focused on instruction and use of scores: teachers have knowledge and skills to implement effective instruction, teachers use data to make instructional decisions, and district and states use data to make resource decisions. The focus of the long-term outcomes was on pedagogy and the incorporation of features specific to the DLM system, including the use of learning maps: teachers understand how to build breadth and depth of conceptual understanding and make useful decisions from diagnostic information, teachers think differently about how to educate students with severe cognitive disabilities (SWSCD) in the context of learning maps, and the educational experiences of SWSCDs are improved.

Initial studies have started to examine the impact of DLM on teacher instructional practices for students with the most severe cognitive disabilities. DLM tests provide learning profiles for students that include a map of the knowledge and skills students have most likely acquired and the skills students are most likely ready to work on. In a study using surveys, teachers reported that the learning profiles from the end of the year assessments provide some useful information to inform instruction; however, some teachers viewed a learning profile as a fixed map for instruction, which may lead to narrowing of the instruction on only the skills reported, lower teacher expectations of students, and limiting student progress (Karvonen et al., 2017). In a follow-up study using focus groups, the use of the DLM reports by teachers was evaluated. Teachers indicated that they use fine-grained information from the learning profiles generated from the interim assessments in the development of IEPs and their plans for instructional groupings, and they use information from the end of year assessment reports as a general framework to inform the IEP goals (Clark et al., 2018).

Every Student Succeeds Act

Under the Elementary and Secondary Education Act (ESEA, 1965), the Every Student Succeeds Act (ESSA, 2015) replaced NCLB in December 2015, requiring states to modify their test-based accountability systems for operational use in the 2017–18 academic year. ESSA extends to the 2020–21 academic year. The goals of ESSA, which are similar to those of NCLB, are to promote fair, equitable, and high-quality education for all students, to close educational achievement gaps, to hold all students accountable to high academic standards that will prepare them to succeed in college and careers, and to hold schools accountable to effect positive change, in particular, in the lowest performing schools. ESSA narrows the role of the federal government and allows states and districts to play a larger role in accountability while continuing the practice of using tests as tools for instructional change and accountability.

ESSA stipulates that states use other measures, in addition to tests, in their school accountability systems to help alleviate the pressure on raising test scores at the expense of meaningful actions to improve student learning. States must continue to report by all required subgroups specified under NCLB. States are required to include the following:

- a measure of student achievement in ELA and mathematics (in each of Grades 3 – 8, plus one grade in high school) and a measure of student achievement in science (once in each grade span: 3 – 5, 6 – 8, and high school)
- another "valid and reliable statewide academic indicator" for elementary and middle schools, which can be a measure of student growth
- the 4-year adjusted cohort graduation rate for high schools (states may add an extended adjusted cohort graduation rate if they choose)
- a measure of progress in English language proficiency for English language learners (in each of Grades 3 – 8, plus one grade in high school
- at least one other measure of school quality or student success that is valid, reliable, and comparable across the state, such as student engagement, educator engagement, student access to advanced coursework, postsecondary readiness, school climate and safety, or other measure

The accountability requirements under ESSA include the following:

- indicators of school performance/quality
- state defined annual targets for indicators
- identification of schools requiring support and improvement
- annual reporting by indicator and subgroup
- state provision of support to schools identified — evidence-based, equity-enhancing approaches to reduce the opportunity gaps

Use of College Admission Tests Under ESSA

ESSA provides additional flexibility compared to previous policies, making it easier for states to adopt a college entrance exam, the ACT or *SAT*®, as their high school accountability test. Several studies have revealed some gaps in the alignment between state content standards and college entrance exams that could result in changes in what is taught in mathematics and ELA for those states that adopt college entrance exams for their high school accountability test. This potential negative consequence needs to be weighed against the positive consequence of using a college entrance exam as the high school accountability test to further the educational opportunities of students who may not have considered higher education. I and many stakeholders would argue that using a college entrance exam as a lever of educational opportunity for underserved students outweighs any misalignment of the exam with high school content standards. Evidence is needed, however, to determine whether the use of college entrance exams increases the higher education opportunities for students.

Maine was one of the first states to adopt a college entrance exam as its state test in high school. In 2006, Maine began administering the SAT as its high school accountability test and was required to augment the SAT mathematics section with additional items to improve its alignment with the state content standards. A study investigating whether more students in Maine entered college as a result of the use of the SAT found that a 2% – 3% increase in college-going rate was attributable to the use of the SAT (Hurwitz et al., 2015). In Michigan, the ACT was adopted as that state's high school accountability test, and the use of it resulted in an increase of approximately 50% of students living in economically disadvantaged areas scoring at a college-ready level but only an increase of 6% of students living in economically disadvantaged areas enrolling at a 4-year institution (Hyman, 2017). Although the use of the ACT increased the number of students enrolling in college, there still remains a high percentage of college-ready students living in economically disadvantaged areas who are not enrolling at a 4-year institution. These studies suggest that more programs are needed to support and encourage student to enroll in and attend higher education institutions, especially students who have been traditionally underserved.

Pilot Program Under ESSA

In an attempt to minimize negative curriculum and instructional consequences and to promote rigorous instructional content, ESSA authorized a pilot program that allowed up to seven states to develop innovative test-based accountability

systems, incorporating new measures of student performance that better supported student learning. The intent of the program was to "promote high-quality instruction, mastery of challenging state academic standards, and improved student outcomes, including for each subgroup of students" (U.S. Department of Education, 2018, p. 26). The pilot program provides states with the opportunity to build balanced assessment systems that lessen the tension between the goals of instruction and the goals of accountability. The systems can incorporate competency-based assessments, instructionally embedded assessments, interim assessments, performance-based assessments, and other forms of assessment.

The assessment system is to be developed in collaboration with a broad range of stakeholders, including parents, educators, and civil rights organizations, and the results need to be valid, fair, and comparable for all students and subgroups. An important provision in the pilot program is that states are to report annually how they will ensure that all students receive the instructional support to meet standards. Thus, evidence is needed that documents the nature of the instruction students are receiving as a result of this program. New Hampshire and Louisiana were approved in 2018, followed by Georgia and North Carolina in 2019.

New Hampshire's Performance Assessment of Competency-Based Education (PACE) pilot program was approved by the U.S. Department of Education in March 2015 and by the ESSA pilot program in the fall of 2018 (State of New Hampshire Department of Education, 2018). It includes a combination of locally developed performance tasks and common performance tasks that are shared by participating districts with a smaller state assessment component. Figure 1 provides the program's theory of action that specifies the program's intended impact on classroom practices, including the design features and mediating effects required to achieve its end goal of students being college-and-career ready (Figure 2). The evaluation of this innovative assessment system should include consequential evidence for the mediating effects and their impact on changes in instruction. The theory of action also considers contextual factors such as district size and negative consequences such as doing "no harm" on the state summative assessment and to minimize sources of construct irrelevant variance (State of New Hampshire Department of Education, 2018, p. 54).

Underlying this theory of action are intermediate goals and claims required to achieve the end goal that students are college-and-career ready.

1. Goal: Stake holders committed to PACE. Claims: Local leadership is committed and participating districts collaborate with one another.
2. Goal: Assessments are based on sound test design principles. Claims: Teachers developing performance assessments are trained and knowledgeable of the *Standards* and performance assessments adhere to the *Standards*, including ensuring equity.
3. Goal: Performance assessments are successfully implemented. Claims: Teachers receive effective training and supports to administer the performance assessments with fidelity, implementing the performance assessments as intended enhances and extends desired instructional practices, and student engagement and student learning increases/deepens when performance assessments are implemented as intended.
4. Goal: Scores are accurate and reliable. Claims: Scorers are effectively trained and scorers attain successful rates of interrater agreement and reliability.

An initial evaluation of whether participation in the PACE program does "no harm" to student performance on the state summative tests was undertaken. In examining the effects of the PACE program on student performance on the state test, using multilevel modeling, small positive effects — ranging from 3% to 14% of a standard deviation — were found for students in Grades 8 and 11 in PACE schools in contrast to students not in PACE schools (Evans, 2019). PACE students tended to perform slightly better than predicted on the state test as compared to students not in PACE schools but with comparable school contextual characteristics.

The development of local performance tasks provides the PACE program an opportunity to build a test that is guided by sociocultural learning theories that reflect the social nature of learning and the cultural practices in the community that shape learning. Designing tasks that are informed by theories of motivation and identity development, as well as cognitive theories, have the potential to address fairness issues. The inclusion of locally designed performance tasks as part of the assessment system may help alleviate potential adverse consequences. In the design of performance assessments, we need to pay careful attention to Messick's (1994) advice that "it should not be taken for granted that richly contextualized assessment tasks are uniformly good for all students … [because] contextual features that engage and motivate one student and facilitate effective task functioning may alienate and confuse another student and bias or distort task functioning" (p. 18).

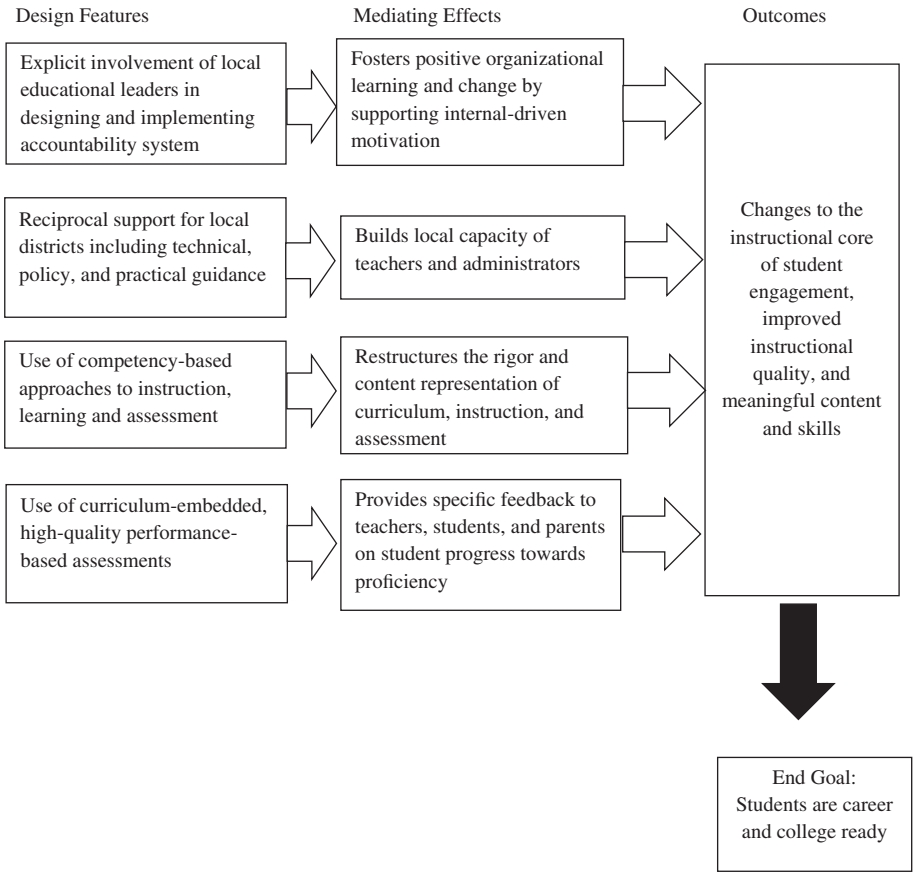

Figure 1 Theory of Action: PACE influence on classroom practices. Adapted from *New Hampshire Performance Assessment of Competency Education (PACE) Technical Manual (2016–2017)*, by S. F. Marion, Evans, & Lyons, 2017, p. 20. Copyright by the National Center for the Improvement of Educational Assessment.

Although it is challenging to design performance tasks situated within the culture of the students, the design of some tasks at the local level may result in more culturally responsive assessments.

Conceptual Framework for Evaluating Consequences

Figure 1 provides a conceptual framework for evaluating the consequences of the use of tests in state accountability systems that can be adjusted based on the theory of action and the IU argument underlying the particular test-based accountability system. The ultimate claim or decision to be made based on test score interpretations in accountability systems is whether students are college-and-career ready, or on track.

Components of the Conceptual Framework

The conceptual framework has a number of components that interact with each other, including the relevant stakeholders, the differing values held by the stakeholders, positive and negative individual and institutional level consequences related to using a test as a lever for educational change, positive and negative individual and institutional level consequences related to test score interpretations and uses, differential impact for subgroups or adverse impact, and contextual variables that provide information when interpreting the consequences of test use.

Stakeholders

Stakeholders have varying perspectives regarding the decisions made based on test scores and, consequently, the resulting consequences of these decisions. Different stakeholders, including policymakers, advocacy groups, administrators,

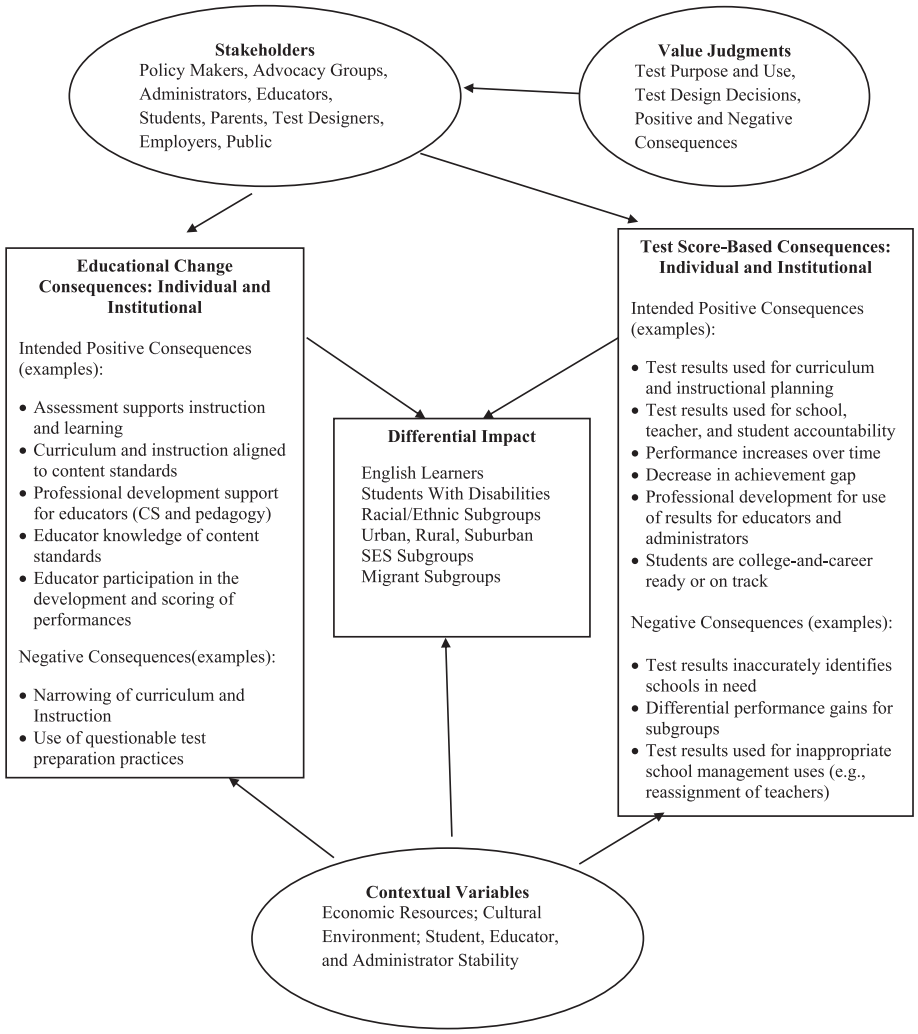

Figure 2 Conceptual framework for evaluating consequences for college-and-career ready claim

teachers, parents, students, business leaders, and the community, have different perspectives on the purposes, uses, and consequences of tests used in accountability systems.

Value Judgments

Value judgments are involved in the articulation of the consequences to be evaluated, and different stakeholders will vary in their determination of what consequences are most valued. As an example, policy makers may be less concerned than parents if test-based accountability systems lead to a narrowing of instruction and a curriculum focus on subject areas that are tested as compared to nontested subjects.

Value judgments that are involved in the initial delineation of the purposes and uses of the test and the construct to be measured, as well as in the test design and development decisions, will affect the testing outcomes and resulting consequences. Decisions made in test design, development, and use are grounded in value judgments (M. T. Kane, 2006). Design and psychometric considerations that bear on the consequences of test use include the specification of the content standards; the alignment of the test to the content standards; representativeness and relevance of the content being assessed; measurement invariance across subgroups; the comparability of the results across students, schools, and districts; generalizability of the assessment results; performance-level classification accuracy; and precision of scores across the score scale. These design considerations may bear on the resulting consequences of test use. As an example, the extent of alignment between the content standards and the test has implications for the usefulness of the test results to provide

instructionally relevant information. If the relationships among items or subdomains vary across subgroups or if the scores are less precise in the lower region of the score scale, the validity of the score interpretations and uses and their resulting consequences will be affected for subgroups and lower scoring students, respectively.

Consequences for Score-Based Decisions and for Tests as Levers of Educational Change

Consequences of decisions and uses based on test score interpretations as well as consequences of decisions and uses that are a result of the broader educational goals of the testing program need to be articulated. The consequences that occur based on the broader educational goals of the testing program and consequences that are a result of the test score-based decisions are multifaceted. Consequences occur at the individual and institutional level, where the institution can be the state, district, school, or classroom. Consequences can occur for individual students, educators, and administrators. There are intended positive consequences and potentially unintended negative consequences. The potential negative consequences that are most severe for individuals, and institutions should receive priority in the validity argument. Typically intended consequences of state assessment programs regarding instruction and learning include but are not limited to (a) student, educator, and administrator motivation and effort to make changes; (b) alignment of curriculum, instruction, and classroom assessment to content standards; (c) teacher professional development support; (d) improved learning for all students; and (e) student, educator, administrator, parent, and public knowledge about the assessment; criteria for judging performance; and how to use assessment results. Unintended, negative consequences include but are not limited to (a) narrowing of curriculum and instruction to focus only on the specific standards assessed and ignoring the broader construct reflected in the standards; (b) use of test preparation materials that are closely linked to the assessment without making meaningful changes to instruction; (c) inappropriate uses of test scores, such as questionable practices in reassignment of teachers or principals; and (d) decreased confidence and motivation to learn and perform well in the assessment because of past experiences with assessments (Lane & Stone, 2002).

Sources of evidence to bear on the evaluation of the consequences include student, educator, and administrator surveys as well as interviews and focus groups that allow for probing more deeply into survey results, instructional logs, and more direct measures such as instructional artifacts, including instructional tasks, classroom assessments, and classroom observations. Instructional artifacts and classroom observations complement the other sources, providing richer data with high fidelity. Studies evaluating whether changes in instruction as evidenced through classroom artifacts and survey data are related to changes in performance on assessments provide valuable information regarding the effects of tests (Lane & Stone, 2002).

The evaluation of consequences can occur at various points in time and on multiple occasions for the examination of trends. The obtainment of baseline data facilitates the interpretation of the consequential evidence. Decisions need to be made about how the consequences will be monitored and evaluated over time. How often and when evidence is collected and how much evidence is needed to make claims about the effect of the test decisions and uses need consideration.

Differential Impact

Differential impact, or adverse impact, bears on fairness issues and requires systematic evaluation. Differential impact should be examined for all relevant subgroups, including cultural/racial subgroups, students with disabilities, and English learners. Because of the history of the differential effects of testing for students who have been traditionally underserved, it is imperative that we learn from the past and minimize any unintended negative consequences for these students.

Contextual Variables

Contextual variables need to be considered in the evaluation of the consequences of an assessment and accountability program (Lane et al., 1998). The interpretation of consequences, particularly differential consequences or adverse social consequences, is informed by relevant contextual variables (Lane et al., 1998). The inclusion of school and community contextual variables, such as cultural environment; economic resources; student, educator, and administrator stability; and access to curriculum and instruction materials, informs the interpretation of the consequential evidence of tests used in accountability systems.

Concluding Thoughts

The reauthorization of ESSA can be informed by what we have learned about the use of assessment and accountability systems over the past 50 years as well as some of the features of the pilot programs supported by ESSA. The goals of these policies are worthy—high standards, quality instruction, and improved student learning—however, more attention to potential undesirable effects is needed so as to mitigate them. Some suggestions for the design and use of tests used for accountability under the reauthorization of ESSA include the following:

- Design coherent, integrated instruction and assessment systems.
- Design tests that model what is valued in instruction and that assess cognitively rich content using more direct measures (e.g., performance tasks).
- Develop locally some tasks and scoring of student performances by trained educators.
- Develop locally tasks that are guided by sociocultural learning theories that reflect the social nature of learning and the cultural practices in the community.
- Include comprehensive, systematic evaluation of consequences as part of the validity argument.
- Anticipate potential negative consequences and adverse consequences, and institute safeguards to alleviate them.

A comprehensive evaluation of all possible consequences of tests used in accountability systems, intended positive consequences and potentially unintended negative consequences, can be daunting. As already discussed, in prioritizing the consequences to be evaluated the following four components should be considered: (a) competing values and needs of stakeholders, (b) fundamental intended positive consequences, (c) severity of potential negative consequences, and (d) differential impact.

Acknowledgments

I would like to thank Ida Lawrence, Randy Bennett, Michael Kane, Irwin Kirsch, and Bob Mislevy for inviting me to present at the 17th William H. Angoff Memorial Lecture. My appreciation is also extended to Jim Carlson for his careful editing of the paper.

References

Achieve. (2015). *Proficient vs. prepared: Disparities between state tests and the 2013 National Assessment of Educational Progress (NAEP).* https://www.achieve.org/files/NAEPBriefFINAL051415_0.pdf

Achieve. (2016). *State test results are getting closer to student achievement on NAEP: Parents, students, and teachers benefit*. https://www.achieve.org/files/ProficiencyvsPrepared2.pdf

Achieve. (2018). *Proficient vs. prepared 2018: Disparities between state tests and the 2017 National Assessment of Educational Progress (NAEP)*. https://www.achieve.org/files/Proficient%20vs.%20Prepared%20May2018_1.pdf

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

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in education performance assessments. *Educational Measurement: Issues and Practice*, *21*(3), 5–18. https://doi.org/10.1111/j.1745-3992.2002.tb00095.x

Bennett, R. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*(2–3), 70–91. https://doi.org/10.1080/15366367.2010.508686

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Borko, H., Wolf, S. A., Simone, G., & Uchiyama, K. (2001, April). *Schools in transition: Reform efforts in exemplary schools of Washington* [Paper presentation]. American Educational Research Association Annual Meeting, Seattle, WA.

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

Clark, A. K., Karvonen, M., Romine, R. S., & Kingston, N. M. (2018, April). *Teacher use of score reports for instructional decision-making: Preliminary findings* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives*, *9*(4), 173–206. https://doi.org/10.1080/15366367.2011.626729

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3 – 17). Erlbaum.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418 – 446. https://doi.org/10.1002/pam.20586

Elementary and Secondary Education Act of 1965, 20 U.S.C. § 70 (1965).

Evans, C. M. (2019). Effects of New Hampshire's innovative assessment and accountability system on student achievement outcomes after three years. *Educational Policy Analysis Archives*, *27*(10), pp. 1 – 37. https://doi.org/10.14507/epaa.27.4014.

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015).

Haertel, E. (1999). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*(1 – 2), 19 – 23.

Hughes, G. B., Daro, P., Holtzman, D., & Middleton, K. (2013). *A study of the alignment between the NAEP mathematics framework and the Common Core State Standards for mathematics (CCSS-M)*. National Center for Education Statistics.

Hurwitz, M., Smith, J., Niu, S., & Howell, J. (2015). The Maine question: How is 4-year college enrollment affected by mandatory college entrance exams? *Educational Evaluation and Policy Analysis*, *37*(1), 138 – 159. https://doi.org/10.3102/0162373714521866

Hyman, J. (2017). ACT for all: The effect of mandatory college entrance exams on postsecondary attainment and choice. *Education Finance and Policy*, *12*(3), 281 – 311. https://doi.org/10.1162/EDFP_a_00206

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527 – 535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed)., *Educational measurement* (4th ed., pp. 17 – 64). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1 – 73. https://doi.org/10.1111/jedm.12000

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R. C., & Staiger, D. O. (2016). *Teaching higher: Educators perspectives on common core implementation*. Harvard University, Center for Education Policy Research.

Karvonen, M., Romine, R. W., Clark, A., Brussow, J., & Kingston, N. (2017, April). *Promoting accurate score report interpretation and use for instructional planning* [Paper presentation]. National Council on Measurement in Education annual meeting, San Antonio, TX.

Kleinert, H. L., & Kearns, J. F. (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities*. Brookes.

Koretz, D., Barron, S., Mitchell, M., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. Rand Corporation.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127 – 135. https://doi.org/10.7334/psicothema2013.258.

Lane, S., & DePascale, C. (2016). Psychometric considerations for performance-based assessments and student learning objectives. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 77 – 106). Routledge.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 24 – 28. https://doi.org/10.1111/j.1745-3992.1998.tb00830.x

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment*, *8*(4), 279 – 315. https://doi.org/10.1207/S15326977EA0804_1

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 23 – 30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x

Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement* (4th ed.). American Council on Education/Praeger.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, *15*(1), 1 – 6. https://doi.org/10.3102/01623737015001001

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, *21*(1), 14 – 16. https://doi.org/10.1111/j.1745-3992.1997.tb00587.x

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4 – 16. https://doi.org/10.3102/0013189X029002004

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Educational Policy Analysis Archives*, *11*(31), 1 – 37. https://doi.org/10.14507/epaa.v11n31.2003

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of the requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *20*(8), 15 – 21. https://doi.org/10.3102/0013189X020008015

Marion, S. F., Evans, C. M., & Lyons, S. (2017). *New Hampshire Performance Assessment of Competency Education (PACE) technical manual (2016–2017)*. National Center for the Improvement of Educational Assessment.

Marion, S. F., & Perie, M. (2009). An introduction to validity arguments for alternate assessments. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 321–334). Brookes.

Messick, S. (1964). Personality measurement and college performance. In A. G. Wesman (Ed.), *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110–129). Educational Testing Service.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Michaels, H., & Ferrara, S. (1999). Evolution of educational reform in Maryland: Using data to drive state and policy local reform. In G. J. Cizek (Ed.), *Handbook of educational policy*. Academic Press.

Moss, P. (2016). Shifting the focus of validity of test use. *Assessment in education: Principles, policy, & practice*, *23*(2), 236–251. https://doi.org/10.1080/0969594X.2015.1072085

National Research Council. (2011). *Incentives and test-based accountability in education*. National Academies Press.

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2002).

Parke, C. S., & Lane, S. (2008). Examining alignment between state performance assessment and mathematics classroom activities. *Journal of Educational Research*, *101*(3), 132–147. https://doi.org/10.3200/JOER.101.3.132-147

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, *12*(3), 239–269. https://doi.org/10.1080/13803610600696957

Partnership for Assessment of Readiness for College and Careers. (2010, June 23). Application for the Race to the Top comprehensive assessment systems, competition. http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Quenemoen, R., Flowers, C., & Forte, E. (2013, April). *Theory of action for the National Center and State Collaborative Alternate Assessments and its role in the overall assessment design* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA.

Roach, A. T., Elliott, S. N., & Berndt, S. (2007). Teacher perceptions and the consequential validity of an alternate assessment for students with significant cognitive disabilities. *Journal of Disability Policy Studies*, *18*(3), 168–175. https://doi.org/10.1177/10442073070180030501

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education*, *19*(1), 405–450. https://doi.org/10.3102/0091732X019001405

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *21*(1), 5–8, 13.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy, & Practice*, *23*(2), 226–235. https://doi.org/10.1080/0969594X.2015.1072084

Smarter Balanced Assessment Consortium. (2015a). *Smarter Balanced Assessment Consortium: End of grant report*. https://portal.smarterbalanced.org/library/en/v1.0/end-of-grant-report.pdf.

Smarter Balanced Assessment Consortium. (2015b). *Smarter Balanced Assessment Consortium: 2016–17 technical report*. https://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

State of New Hampshire Department of Education. (2018). *Application for the new authorities under the innovative assessment demonstration authority*. Author.

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability*. RAND.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000). *The effects of the Washington State education reform in schools and classrooms* (CSE Tech. Rep. No. 525). University of California, Center for Research on Evaluation, Standards and Student Testing.

Stecher, B., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996–97 RAND survey of Kentucky teachers of mathematics and writing* (CSE Tech. Rep. No. 482). University of California, Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., Russell, J., & Naftel, S. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006*. RAND.

Stecher, B., & Mitchell, K. J. (1995). *Portfolio driven reform: Vermont teachers' understanding of mathematical problem solving* (CSE Technical Report 400). University of California, Center for Research on Evaluation, Standards, and Student Testing.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education*, *16*(1), 1–26. https://doi.org/10.1207/S15324818AME1601_1

Towles-Reeves, E., Garrett, B., Burdette, P. J., & Burdge, M. (2006). Validation of large-scale alternate assessment systems and their influence on instruction—What are the consequences? *Assessment for Effective Intervention*, *31*(3), 45–57. https://doi.org/10.1177/073724770603100304

U.S. Department of Education. (2009). *Race to the top program: Executive summary*. https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2010). *Race to the top program: Purpose*. https://www2.ed.gov/programs/racetothetop-assessment/index.html

U.S. Department of Education. (2018). *Application of new authorities under the Innovative Assessment Demonstration Authority*. Author.

Suggested citation:

Lane, S. (2020). *Test-based accountability systems: The importance of paying attention to consequences* (Research Report No. RR-20-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12283

Action Editor: James Carlson