This report is part of the RAND Corporation research report series. RAND reports present research findings and objective analysis that address the challenges facing the public and private sectors. All RAND reports undergo rigorous peer review to ensure high standards for research quality and objectivity.

Research Report

New Assessments, Better Instruction?

Designing Assessment Systems to Promote Instructional Improvement

Susannah Faxon-Mills, Laura S. Hamilton, Mollie Rudnick, Brian M. Stecher

Authors are listed in alphabetical order

Sponsored by the William and Flora Hewlett Foundation

The research described in this report was sponsored by the William and Flora Hewlett Foundation, and was produced within RAND Education, a unit of the RAND Corporation.

RAND OFFICES

SANTA MONICA, CA • WASHINGTON, DC

PITTSBURGH, PA • NEW ORLEANS, LA • JACKSON, MS • BOSTON, MA

DOHA, QA • CAMBRIDGE, UK • BRUSSELS, BE

Preface

In 2010, the William and Flora Hewlett Foundation initiated a new strategic initiative that focuses on students' mastery of core academic content and their development of *deeper learning skills* (i.e., critical-thinking, problem-solving, collaboration, communication, and learn-how-to-learn skills). In its efforts to encourage schools to promote deeper learning, the Foundation is looking for leverage points to influence change in schools. Assessment can be a lever for change, and the Foundation asked the RAND Corporation to explore various aspects of assessment related to deeper learning. In an earlier project, RAND Education tracked the extent to which U.S. students are currently assessed in a way that emphasizes deeper learning skills. This project explores the extent to which changing assessment is likely to lead to changes in educational practice by reviewing the research literature on the impact of assessment on instruction and the factors that may mediate that relationship.

This report should be of interest to assessment developers who want to understand how their tests might influence instruction and to education policymakers and practitioners who are seeking to understand how to incorporate assessments into efforts to improve education.

This research was conducted by RAND Education, a unit of the RAND Corporation. Funding to support the research was provided by the William and Flora Hewlett Foundation.

Contents

Figure

Tables

Summary

The Hewlett Foundation is committed to promoting "deeper learning" for America's students—learning that prepares students to master core academic content, think critically and solve complex problems, work collaboratively, communicate effectively, develop positive habits of mind, and learn how to learn—and it recognizes that assessment can play a key role in this effort. The Foundation's view of deeper learning is consistent with many other organizations' calls to expand and enrich the expectations we hold for students, including calls for teaching 21st century skills and international competence (Saavedra and Opfer, 2012) and the recent adoption by more than 40 states of the Common Core State Standards (CCSS). The Foundation has expressed interest in exploring factors that might promote high-quality instruction in response to the new standards, and in particular the role that new, CCSS-aligned assessments might play in improving instructional quality. The Foundation commissioned RAND to review research about the effects of assessment and to summarize what is known about assessment as a lever for reform. One of the primary goals of this work is to understand how new assessments that are aligned with the CCSS and that attempt to measure higher-order skills and processes might be expected to influence instruction. This exploration had two central questions:

- What does research tell us about the influence of testing on instructional practice, and what are the implications of this research for predicting the likely impact of new, CCSS-aligned assessments?
- What conditions could be put in place to promote a positive impact of assessments on instruction and, ultimately, deeper learning?

To explore the likely influence of new CCSS-aligned assessments on teaching practice and the conditions that moderate that relationship, we conducted a series of literature reviews that focused on the following topics: (1) high-stakes testing in U.S. public education, (2) performance assessment in U.S. public education, (3) large-scale educational assessment in international settings, (4) formative assessment and teachers' use of test results, (5) military and occupational testing, and (6) professional certification and licensure testing. In these reviews, we paid particular attention to assessment's role in promoting instructional change as well as the external conditions that might hinder or enable such change. We did not limit the search to research on assessments that measure higher-order skills or processes (although these skills are often the focus in the literature on performance assessment) but included assessments of any academic, occupational, or professional achievement.

We found considerable research on the effects of testing in U.S. schools, including studies of high-stakes testing, performance assessment, and formative assessment. Studies of international assessment, military and occupational testing, and professional certification and licensure yielded fewer relevant findings. The studies suggest a wide variety of effects that testing might have on

teachers' and students' activities in the classroom, including changes in curriculum content and emphasis, allocation of time and resources across different pedagogical activities, and teacher-student interactions. At the same time, extensive variability in how educators responded to tests, which we observed both across different studies as well as within individual studies, suggests that responses depend on the characteristics of teachers as well as on the contexts in which they work.

Much of the research on testing has occurred in the context of accountability, where there are important consequences associated with test results and hence the tests have have high stakes. Much of the research shows that educators respond to high-stakes assessments differently than to lower-stakes assessments (Firestone, Mayrowetz, and Fairman, 1998; Pedulla et al., 2003). Therefore, the stakes that are attached to test performance clearly represent an important mediating factor of a test's effects on instructional practice. The literature also identifies a number of other conditions that affect the impact that assessment may have on practice. These include:

- attributes of the tests, such as their purposes, technical quality, and format
- background, beliefs and knowledge of teachers and administrators, including their domain knowledge, familiarity with the assessment, beliefs about teaching and learning, and endorsement of the assessment
- characteristics of the school and students, such as grade configuration and demographics
- district/school policies including those related to professional development, teacher collaboration, and curriculum.

Specifically, tests of deeper learning are likely to promote desirable changes in practice under the following circumstances:

- Test content and format should mirror high-quality instruction.
- Tests should be used only for purposes for which they were designed and validated.
- Score reporting should be optimized to foster instructional improvement.
- Teachers should receive training and support to interpret and use test scores effectively.
- The test scores should "matter," but important consequences should not follow directly from test scores alone.
- If there are externally mandated, high-stakes tests, they should be part of an integrated assessment system that includes formative and summative components.
- Accountability metrics should value growth in achievement, not just status, and should be sensitive to change at all levels of student performance, not just a single cut point.
- Assessment should be one component of a broader systemic reform effort.

By themselves, tests of deeper learning are likely to have some impact on classroom instruction. However, research suggests that the benefits of tests will be enhanced by policies ensuring that the tests have features to make them helpful for instructional improvement, are accompanied by specific supports to help teachers increase their relevant knowledge and skills and modify their practices, and are part of a larger, systemic change effort.

Acknowledgments

Marc Chun at the Hewlett Foundation first approached us about reviewing the literature on the impact of assessment, and he was very helpful in framing this investigation. Helpful suggestions were provided by RAND Quality Assurance Manager Cathy Stasz and by our reviewers, Margaret Goertz and Jennifer McCombs. We appreciate the assistance that Donna White provided in preparing the document.

Abbreviations

AfL	Assessment for Learning
CCSS	Common Core State Standards
MSPAP	Maryland School Performance Assessment Program
NCLB	No Child Left Behind
PARCC	Partnership for Assessment of Readiness for College and Careers
Smarter Balanced	Smarter Balanced Assessment Consortium
TEPA	Technology-Enhanced Formative Assessment
WSS	Work Sampling System

1. Introduction

The Hewlett Foundation is committed to promoting "deeper learning" for America's students—learning that prepares students to master core academic content, think critically and solve complex problems, work collaboratively, communicate effectively, develop positive habits of mind, and learn how to learn (William and Flora Hewlett Foundation, 2013)—and it recognizes that standards and assessment can play a key role in this effort. The Foundation's view of deeper learning is consistent with many other organizations' calls to expand and enrich the expectations we hold for students, including calls for teaching 21st century skills and international competence (Saavedra and Opfer, 2012). All these efforts are designed to prepare students to meet the demands of a changing workplace and a more tightly interconnected world. Because a number of states have recently adopted the Common Core State Standards (CCSS), which emphasize some deeper learning skills to a greater extent than earlier state standards did, the Foundation has expressed interest in identifying factors that might promote high-quality instruction in response to the new standards. The CCSS-aligned assessments that are being developed by two consortia might play a particularly important role in promoting improved instructional quality in response to the CCSS, especially if they address deeper learning to a greater extent than existing state tests, which address these skills to only a minimal degree (Yuan and Le, 2012).

The Foundation also understands that the relationship among standards, assessments, instruction, and learning is complex; the common notion that "what you test is what you get" (Resnick and Resnick, 1992) is too simplistic to use as a guide for program development or implementation. While there is ample evidence that what you test (i.e., the topics that are covered and the ways they are measured) influences what you get (i.e., the subject areas teachers emphasize and the knowledge and skills students learn), the effects of testing on learning depend on many other factors, including the alignment of tests with standards, curriculum, and other features of the education system and the consequences associated with test results. In particular, there is still much to learn about how changes in testing might influence the education system and how tests of deeper content and more complex skills and processes could best be used to promote the Foundation's goals for deeper learning.

Given the gaps in evidence regarding the link between testing and student outcomes, the Foundation commissioned RAND to review research about the effects of assessment in education and in other fields, and to summarize what is known about assessment as a lever for reform. One of the primary goals of this work is to understand how new assessments that are aligned with the CCSS and that attempt to measure higher-order skills and processes might be expected to influence instruction. This exploration had two central questions:

- What does research tell us about the influence of testing on instructional practice, and what are the implications of this research for predicting the likely impact of new, CCSS-aligned assessments?
- What conditions could be put in place to promote a positive impact of assessments on instruction and, ultimately, deeper learning?

RAND took a broad approach to answering these two questions. We started with an examination of the research on high-stakes testing in public education and expanded into other areas, including formative assessment in K–12 education and assessments of occupational and professional knowledge and skill in other fields. We included studies of assessments of higher-order skills or processes but did not limit the review to these types of assessments. We conducted a scan of the research in each area and examined more closely those studies that were most relevant to the two guiding questions.

Answers to these questions will be of interest to educators and policymakers across the country. More than 40 states have adopted the Common Core State Standards (CCSS) in reading and mathematics, which entail more rigorous academic content and place greater emphasis on critical thinking, problem solving, and communication than most of the standards that states had previously adopted. In addition, two national consortia are developing "next generation" assessments aligned with the CCSS, and these assessments are being designed to measure many aspects of deeper learning. While the final designs for the assessments are still being developed, both the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC) have indicated that their new tests will address more rigorous content and do so in richer ways than previous large-scale assessments. Because most states have plans to adopt one or the other set of assessments, it seems appropriate to consider them as part of the context for this paper. Thus we describe them more completely here and return to them in the concluding section, where we consider the findings from our literature reviews.

As of July 2013, both Smarter Balanced and PARCC aim to include summative, interim, and formative components for both English language arts and mathematics. PARCC assessments will be delivered at grades 3–12, and Smarter Balanced assessments will be administered in grades 3–8 as well as grade 11. PARCC includes two required summative assessments—one performance-based assessment to be administered three-quarters of the way into the school year and one end-of-year assessment. According to PARCC's website, these summative assessments will be designed to "make 'college- and career-readiness' and 'on-track' determinations, measure the full range of standards and full performance continuum, and provide data for accountability uses, including measures of growth" (Partnership for Assessment of Readiness for College and Careers, no date). The summative assessments from Smarter Balanced are designed to achieve similar goals but will be administered in the last 12 weeks of the school year. Like PARCC, these tests will include performance tasks as well as computer-enhanced and constructed-response items.

The interim and formative components of both assessment systems are optional, though PARCC does require an assessment of students' speaking and listening skills, which is not included in the overall summative score. PARCC's other non-summative components are designed to inform instruction. They include diagnostic assessments that are intended to indicate student knowledge and skills and mid-year performance-based assessments that emphasize "hard-to-measure" standards. PARCC is also developing a range of assessment tools for grades K–2 that are aligned to both the CCSS and to the PARCC assessment system, to both prepare students for later grades and prepare teachers for incoming 3rd-grade students. Smarter Balanced interim assessments aim to be flexible, so that educators can locally select the item sets they want to measure and the timing of assessments so that it can be strategically placed within the instructional year. These assessments, as well as additional formative tools developed by the consortium, are intended to help educators gain a better understanding of where students are in their learning.

While both assessment systems will be computer-delivered, one significant difference between the two is that Smarter Balanced uses computer adaptive testing for both its summative and interim components. For computer adaptive assessments, the computer program will adjust the level of difficulty of questions throughout the course of the assessment, depending on whether a student has answered the previous question correctly. According to Smarter Balanced, "these assessments present an individually tailored set of questions to each student and can quickly identify which skills students have mastered" (Smarter Balanced Assessment Consortium, no date).

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) released a report in early 2013 (Herman and Linn, 2013) that utilizes Norman Webb's depth of knowledge taxonomy (Webb et al., 2005) as a framework for analyzing the representation of deeper learning in the two consortia's assessment models. Though the report acknowledges that there are several potential challenges—particularly budgetary—that may constrain the intentions of both consortia, its initial analysis indicates that both assessment systems "represent many goals for deeper learning, particularly those related to mastering and being able to apply core academic content and cognitive strategies related to complex thinking, communication, and problem solving" (p. 17).

PARCC and Smarter Balanced are not the only options for states that want to align their assessments with the CCSS. Commercial publishers are also developing new, CCSS-aligned assessment systems, and states continue to have the option of developing their own assessments. Whichever assessment state education officials decide to adopt, their expectations regarding how the assessments will influence teaching and learning should be informed by what is known about the relationship between testing systems and classroom practice.

The remainder of the report is organized in four chapters. The first describes the methods we used to review the relevant research literature and how we integrated findings that were drawn from research designs with different levels of rigor. This chapter also offers a framework for

summarizing the empirical evidence, and it presents a simple conceptual model for the influence of testing in the larger education system. The second chapter summarizes the literature on how educators respond to testing, organized in terms of the conceptual model. Chapter Three reviews the research literature on the factors that mediate responses to testing, i.e., the assessment features and contextual conditions that enhance the potential benefits of testing. Finally, in the fourth chapter we draw conclusions about ways to enhance the role of assessment in promoting the Foundation's goals for deeper learning.

2. Methods

To explore the likely influence of new CCSS-aligned assessments on teaching practice and the conditions that moderate that relationship, we conducted a series of literature reviews focusing on the relationship between testing and classroom practice within the following topic areas: (1) high-stakes testing in U.S. public education, (2) performance assessment in U.S. public education, (3) large-scale educational assessment in international settings, particularly in those countries in which schooling is organized in a similar manner as in the U.S., (4) formative assessment and teachers' use of test results, (5) military and occupational testing, and (6) professional certification and licensure testing. We selected these six topic areas because most of the published research on the effects of achievement testing on practice was conducted in these fields. In our literature reviews, we paid particular attention to the role of assessment in promoting instructional change as well as the external conditions that might hinder or enable such change. The first four areas had the greatest number of relevant sources and received the bulk of our attention.

The first step for each of these research areas was to identify relevant material from previous literature reviews on these topics, including those conducted by RAND researchers (e.g., Hamilton, Stecher, and Klein, 2002; Hamilton, 2003; Stecher, 2010) and by the National Research Council (e.g., Koenig, 2011). We also consulted with experts in these fields as a starting point for identifying relevant literature. Using these reviews and recommendations as our foundation, we conducted searches using Google Scholar, EBSCO, and ERIC to acquire any new or previously missed sources. While we did not restrict our searches to a specific time period, we paid particular attention to sources from the past ten years, since these studies were less likely to have been included in previous literature reviews. To cast the widest possible net, we used several search terms, including different combinations of the following terms: national tests, summative assessment, performance assessment, interim assessment, formative assessment, assessment data, high-stakes testing, portfolio assessment, performance-based assessment, standardized tests, classroom practice, instructional change, pedagogy, and teaching. We did not restrict the search based on study size or method, and thus our final collection of reports included large controlled studies and smaller qualitative case studies. We omitted from our analyses studies that did not include either findings or discussion about the influence of assessments on classroom practice or the factors that mediated that influence. As studies meeting this criterion were identified, we also reviewed their references and incorporated relevant sources into our review.

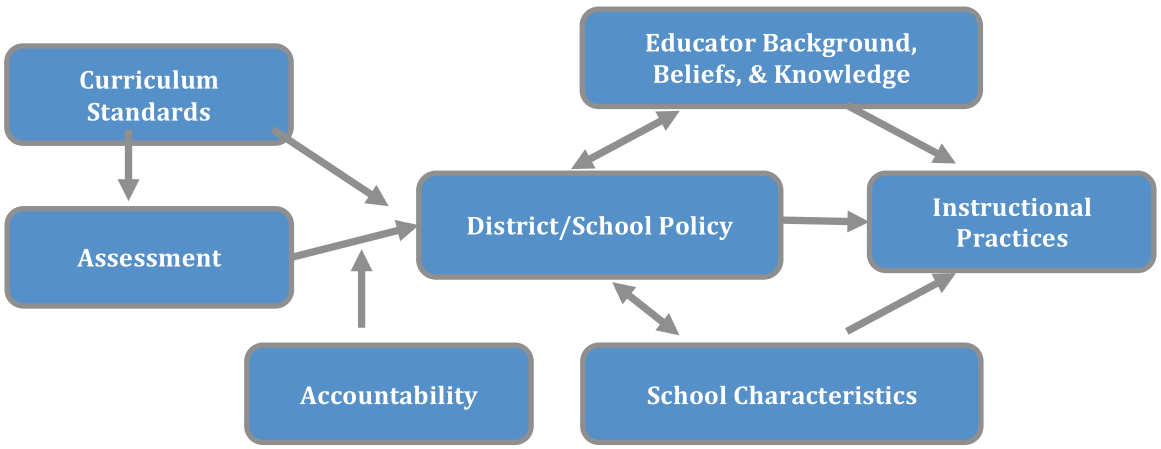
We identified more than 130 sources—including qualitative and quantitative studies, literature reviews, metasyntheses, and policy papers—for our analysis. For each of these studies, we wrote a summary that highlighted its purpose, methods, limitations, and pertinent findings.

We used the summaries to develop a set of codes for classifying the findings in terms of effects on practice and mediating factors (our two research questions). Initially, we created distinct codes for each of the changes in practice and mediating factors we identified in the literature; subsequently, we clustered similar codes together and created a two-level classification scheme with broad categories (e.g., changes in curriculum) subsuming more specific categories (e.g., changes in the sequence of topics). This process yielded nine specific changes in practice across three broader categories and 16 specific mediating factors across five broader categories, which we describe in Chapter Three. We used this structure to organize the presentation of findings. We annotated each summary with all applicable codes, and then we identified the most frequently discussed themes. In Chapter Three, we highlight the findings that appeared with the greatest frequency as well as the greatest consistency.

Framework

We developed a framework to show the role of assessment in influencing instructional practice (see Figure 2.1). The framework was based on similar models that describe the patterns of influence among standards, assessment, accountability, teaching practice, and student outcomes (Goertz, Oláh, and Riggan, 2009; Hamilton et al., 2007; National Research Council, 2001). The studies we reviewed suggested a number of factors that appear to influence the ways in which assessment affects teaching and learning, and we incorporated this evidence into our framework. We used the framework to structure our analysis of the literature and to organize the presentation of our findings. In the following paragraphs, we briefly describe each element, and in Chapter Three we present the literature that supports them.

Figure 2.1.
How Assessment Might Influence Instructional Practices



Curriculum Standards

The standards identify the knowledge and skills students are expected to master. A majority of states have adopted the CCSS as their standards for English language arts and mathematics. Standards influence practices in two ways: directly through district and school policies (like textbook selection) and indirectly by influencing what is included on state assessments.

Assessment

Assessments are usually developed or selected to measure the content represented in the standards. States that adopt the CCSS are likely to adopt one of the consortia or commercial tests that are aligned with the CCSS. Assessments influence practice by signaling to educators which aspects of curriculum will "count." If educators attend to the tests, then features of the test such as item format, the validity/reliability of the test scores for their intended purposes, and methods of score reporting may influence what is taught and how it is taught.

Accountability

Research suggests that attaching incentives to test-based measures of performance gives the tests greater influence on practice than would occur in the absence of incentives. It is in part because of this mediating relationship that most of the research examining the effects of testing on instructional practices focuses on tests with high stakes for educators (and, to a lesser degree, for students), including those that are part of state or local accountability systems.

District/School Policy

Districts and schools respond to the standards and assessment and their use for accountability by setting policies related to curriculum, resources, teacher support, etc. Some of these policies influence instructional practices directly (e.g., new textbooks), while others operate indirectly by changing teachers' knowledge or beliefs or key features of the school in which they work. These policies and practices can act as barriers to, or facilitators of, effective responses to testing.

Educator Background, Beliefs, and Knowledge

Educators' opinions and attitudes (e.g., teacher buy-in or support for the assessment, teacher beliefs about pedagogy or student learning) directly affect their instructional practice. There is also a reciprocal relationship between district/school policies and practices and teacher beliefs and knowledge. For instance, school and district policy can influence the opinions and attitudes of teachers, but those opinions and attitudes can also inform the policies and practices adopted by a given school or district. As in the case of district and school policies, teachers' background, beliefs, and knowledge can serve as barriers to, or facilitators of, effective responses to testing.

School Characteristics

Features of the school, such as grade levels served or instructional philosophy, and the characteristics of the students, such as family income level and parental support, influence teachers' instructional practices directly, or they may have an indirect influence through their effect on school or district policy.

Instructional Practices

Changes or improvements in instructional practices are the ultimate outcome of interest in this report. As discussed in subsequent sections, these instructional modifications may encompass changes in curriculum (e.g., content focus), changes in pedagogical strategies (e.g., approaches to classroom assessment), and changes in how teachers interact with specific students. These changes are ultimately manifested at the classroom level, but they might originate at other levels, such as when a district adopts a new required curriculum in response to changes in state testing policies. Either directly or indirectly, instructional practices may be influenced by all other factors included in the diagram.

Limitations

Time and resource constraints limited the extent of our literature reviews, but we do not think this had a serious effect on our findings. Most importantly, we included all the clearly relevant studies from major sources that were available for electronic searching. In addition, many of the studies we reviewed also included comprehensive reviews of other literature, leading to fairly wide coverage of each body of literature. While much of the relevant research consisted of smaller, qualitative studies, these studies were sometimes more revealing about mediating factors than the more quantitatively oriented studies.

Another complication we faced is that some studies revealed significant diversity among teachers in their responses to the same assessment conditions. For example, in Shepard and Dougherty (1991), slightly more than half of the teachers reported placing more emphasis on direct instruction, while slightly more than one-third reported placing more emphasis on small group instruction. Diversity among teachers in their responses to particular reforms or policies is common in education research, and it reflects in part the fact that teachers often retain a great deal of autonomy in their classrooms even when facing strong external accountability demands (Hamilton et al., 2008). However, these discrepant responses mean that any conclusions we make about the ways that certain conditions influence teaching need to be interpreted in light of the probability that not all teachers will respond in the expected ways.

Diversity in findings across studies was also common. For instance, Au (2007) conducted a meta-synthesis of 49 qualitative studies on the effects of high-stakes testing. While the majority of the studies found that testing led to narrowing of curriculum and increased use of teacher-centered pedagogies, there were also a significant number of studies that found that testing

expanded the curriculum and led to increased reliance on more student-centered pedagogies. Such contradictory findings might be attributable to differences in the structure and nature of the tests themselves or the conditions in which teachers work. This idea will be discussed more in subsequent sections.

Thus, while there are recurring themes in the research literature on the impact of tests, there is not one clear set of findings that apply to all situations. Furthermore, given the qualitative nature of many of the studies, the generally small sample sizes, and the lack of research designs that would support causal inferences, this review does not provide definitive evidence regarding the effects of testing, but instead helps us understand potential consequences (intended or not) of different types of assessment and the conditions surrounding them.

3. How Educators Respond to Assessment

Research indicates that testing is associated with a variety of changes in educators' practices (and, as we will discuss in Chapter Four, these changes are influenced by a number of factors). To understand how the widespread adoption of new, CCSS-aligned assessments would be likely to influence students' experiences in schools and classrooms, it is important to consider the range of common responses to testing and the ways in which these might be shaped by a shift from current testing practices to an approach to assessment that emphasizes deeper learning. In this chapter, we summarize what is known about test-induced changes in instructional practice. In Chapter Four, we summarize the research on factors that mediate the relationship between tests and teachers' reactions to tests.

Educators' responses to assessments take a wide variety of forms (see Kober [2002], Koretz and Hamilton [2006], and Stecher [2002] for discussions of how these responses can be categorized), and we use the general term "responses to testing" to describe them. The commonly used term "test preparation" often carries negative connotations, conjuring images of students engaged in extensive drill and practice or spending time filling in multiple-choice bubbles. We intend for the term "responses to testing" to suggest a much broader range of actions and activities, some of which might be beneficial for promoting student learning and some of which are probably unhelpful or even potentially harmful.

The studies reviewed suggest a variety of ways that testing can influence teachers' and students' activities in the classroom. We have classified these changes into three broad categories of instructional practices: changes in curriculum content and emphasis, changes in how teachers allocate time and resources across different pedagogical activities, and changes in how teachers interact with individual students. Table 3.1 shows our classification, including these main categories and subcategories, and the following sections describe each type in detail.

Changes in Curriculum Content and Emphasis

Research identified changes in *what* is taught, i.e., the curriculum, as a result of assessments. These effects include changes in the order in which content is covered, narrowing the focus of the curriculum to certain subjects or content, and focusing on particular types of skills (e.g., basic skills or facts, higher-order thinking skills or concepts). Remember that shifts in curriculum can reflect actions taken at various levels of the education system. In many schools, principals have become more active in trying to influence time spent on tested subjects (Ladd and Zelli, 2002), and both district and school administrators can take steps such as changing the amount of time devoted to specific subjects during the school day (Rentner et al., 2006; Hannaway, 2007).

Table 3.1.
Changes in Instructional Practices in Response to Assessment

Changes in Curriculum Content and Emphasis
• Changes in the sequence of topics
• Reallocation of emphasis across and within topics
• Focus on basic skills and tasks
• Focus on higher skills and cognitive level
Changes in Pedagogical Activities
• Focus on test preparation
• Changes in instructional strategies
• Changes in classroom assessment practices
Changes in Teachers' Interaction with Individual Students
• Using test results to individualize instruction
• Focus on "bubble kids"

Changes in the Sequence of Topics

It is common for high-stakes tests to be administered in the spring, several weeks or even months before the end of the school year. To ensure that students are exposed to all tested content before they take the test, teachers or administrators sometimes rearrange the sequence in which content is presented. For instance, to accommodate the fact that the Texas state history test that was administered in April covered history up until the 1970s, high school social studies teachers in that state recognized the need to rearrange the curriculum in order to cover all of the tested content before the April test (Salinas, 2006). Similarly, some Maine mathematics teachers reported teaching geometry earlier in the year than they otherwise would have because it was on the test (Firestone, Mayrowetz, and Fairman, 1998). A case study comparing the practices of Kentucky high school social studies teachers in required (and tested) courses versus nontested elective courses revealed that teachers tended to teach the elective courses thematically, whereas they tended to teach required courses chronologically to ensure that students received all of the content they might be tested on and that there was consistency across teachers (Fickel, 2006).

There was also some evidence of the curriculum being shifted across grade levels to accommodate testing. For example, a principal in Maine explained that the history class that focused on Maine history was moved from 9th to 7th grade in his school because the material covered by that class was included on the 7th-grade test (Firestone, Mayrowetz, and Fairman, 1998).

Reallocation of Content Across and Within Subjects

Perhaps the most commonly reported reactions to tests involve reallocation of curriculum content to focus more on tested subjects or topics and less on subjects or topics that are not tested. The tendency for educators to focus more on tested than nontested content could be considered beneficial in the context of a testing program that covers a broad range of skills and knowledge, but would generally be viewed as undesirable if it occurs in response to tests that sample only a subset of the skills and knowledge that are considered important. While reduction in emphasis on social studies, art, and other subjects that are frequently omitted from high-stakes testing programs typically receives the bulk of attention from critics of testing, both forms of narrowing—across and within subjects—have been documented in the literature, and both raise concerns about what students are missing (House of Commons, Children, Schools and Families Committee [United Kingdom], 2008; Yeh, 2005). At the same time, some reallocation might be considered desirable and in fact could be an explicit goal of accountability systems that emphasize specific subjects (Hannaway and Hamilton, 2008). Therefore it is important to understand the specific changes that are made and the extent to which they are consistent with the goals that educators, parents, and other stakeholders have for student learning.

Smith (1991) described the process in which elementary teachers, faced with numerous curricular demands and other programs, began to drop nontested and nonrequired activities to focus on testing and raising test scores. Several other studies found a similar shift toward tested material in response to high-stakes testing policies (Rentner et al., 2006; Jones et al., 1999; Amrein and Berliner, 2012; Yeh, 2005; Au, 2007; Nichols and Berliner, 2005). This phenomenon is not limited to the United States; students in England were given less time with subjects such as physical education, music, and technology as a result of testing (Wiggins and Tymms, 2002).

A RAND study of the implementation of No Child Left Behind (NCLB) traced changes in instructional time over three years of implementation of the law, from 2004 to 2006, in California, Georgia, and Pennsylvania (Hamilton et al., 2007; Stecher et al., 2008). Across the three states, the greatest increases in instructional time reported by teachers were in core subjects tested according to NCLB accountability requirements—English/language arts and mathematics, with elementary teachers more likely than middle school teachers to report reallocating time toward these subjects. This difference probably reflects the higher level of control that elementary teachers in self-contained classrooms have over time allocation as compared with middle schools, where different classes are often taught by different teachers and schedules do not allow for shifting of time across subjects.

Many studies also mentioned narrowing curriculum within content areas. Interviews with 13 middle school language arts teachers and case studies of two of them revealed that end-of-grade high-stakes test required teachers to focus their writing instruction on structure and elaboration, giving them less time to focus on literature appreciation, collaborative work, and engaging in

writing activities oriented toward the real world (Watanabe, 2007). In other instances, the impact of the test on the narrowing of content was more direct. In the case of a group of high school social studies teachers in Kentucky, the high-stakes test and lack of alignment with the state curriculum framework led teachers to analyze the content from released test items to identify the specific content they would teach (Fickel, 2006).

Our review of professional certification and licensure literature suggests that assessments' influence on curriculum content is not exclusive to the K–12 education sector. There has been a debate in the legal education field as to whether the bar exam drives curriculum decisions (Trujillo, 2007). The Society of American Law Teachers (SALT) released a statement in 2002 contending that law schools offer "bar courses" at the expense of clinical or more specialized courses. This sentiment is echoed by Howarth (1996), who argued that bar exams determine the curricula that law schools teach by creating, "a canon of legal education, making certain courses central and exiling others to the periphery. The 'core' courses in a law school's curriculum are very likely to be the courses tested on a jurisdiction's bar exam" (p. 928). However, others argued that such criticism was unfounded and that there was little statistical evidence that the bar exam actually influenced decisions around upper-level course offerings (Darrow-Kleinhaus, 2004; Carpenter, 2005).

While many studies noted that focusing on tested subjects meant that nontested subjects were given less emphasis, a few reported that this led to more opportunities for cross-curricular integration. Science teachers in Minnesota, for example, acknowledged that their high-stakes test led to an increased focus on math and reading, but they thought that integrating these subjects into their science instruction was beneficial (Yeh, 2005). However, such integration can also lead to an imbalance in curricular focus. For example, in a study exploring the impact of federal legislation designed to improve the quality of vocational education programs, Stasz et al. (2004) found that integrating academics into vocational education (by raising vocational education standards, a core performance indicator in the legislation) reduced time and focus on actual vocational tasks because of the increased time on academic requirements and test preparation.

In other cases, testing affected non-academic activities, as well. A national survey of teachers found that while time spent on tested subjects increased, time spent on both nontested subjects and other activities (e.g., student free time, field trips, assemblies) decreased (Pedulla et al., 2003). Nichols and Berliner (2005) identified instances of naptime, recess, and lunch being cut or given less time in order to provide more time focused on tested subjects and test preparation, and the RAND NCLB studies found similar reductions in activities such as field trips (Hamilton et al., 2007; Stecher et al., 2008).

Focus on Basic Skills and Facts

Another way that curriculum might change in response to testing is in terms of cognitive depth, either by focusing more on basic skills (this section) or by emphasizing higher-order skills (next section), depending on the perceived emphasis of the test. There is a large amount of research

suggesting that high-stakes testing leads to increased focus on basic skills (Jones et al., 1999; Herman and Golan, 1991; Shepard and Dougherty, 1991) or facts (Johnston and McClune, 2000; Gallagher and Smith, 2000). Watanabe (2007) noted how middle school English teachers in North Carolina tend to frame their questioning in terms of right and wrong answers to parallel the multiple-choice state test. Even higher-order thinking skills, such as making inferences, were turned into a series of specific answers that led to the ultimate "right" answer. The House of Commons Committee (2008) recognized that such a focus tends to lead to "shallow learning" and short-term knowledge retention. This can occur even when the standards themselves emphasize deeper learning. A study of alignment between state standards and tests found that even when standards included many higher-order skills and competencies, the tests that were designed to measure those standards focused more heavily on easy-to-measure constructs, such as procedural knowledge and computational fluency in mathematics (Rothman et al., 2002).

While focusing solely on basic skills would generally be identified as a negative consequence of testing, there is also a recognition that addressing basic skills is not as problematic if that is not the sole emphasis. Yeh (2005) noted that Minnesota teachers and principals were concerned that state-mandated testing focused on basic skills would lead teachers to focus only on helping students achieve those basic skills. However, they also recognized that they could have a positive effect on learning if basic skills are not the exclusive focus and if they are integrated throughout the curriculum both across and within grade levels.

Focus on Higher Skills and Cognitive Levels

Refocusing curriculum in response to tests is not always problematic (Smith and O'Day, 1991), and this may be particularly true when the focus is on raising the content, skill, and cognitive levels addressed through classroom curricula. This is one of the promises of performance assessments, which may encourage an increased focus on higher-order thinking skills, and on activities that can hold long-term value for students, such as research and writing (Darling-Hammond and Adamson, 2010).

While interpretations of what a performance assessment is and how it is structured can vary widely, one of its defining characteristics is that student responses to performance tasks are unconstrained by a pre-specified set of options, as in multiple-choice tests. Both multiple-choice tests and performance assessments contain some sort of stimuli or prompt serving as a basis for student response, but the unconstrained nature of performance task responses allows these assessments to include a wider and more complex range of stimuli (Stecher, 2010). Whether this, combined with the other characteristics of a given performance assessment, can lead to an instructional focus on an expanded and more complex range of skills has been a research question behind several studies in recent decades.

The statewide portfolio assessment system adopted by Vermont in the 1990s provides a good example of a large-scale performance assessment. In research by Koretz et al. (1994), teachers reported increasing their emphasis on mathematical problem solving and representations in

response to the emphasis on these activities in the portfolio system. When mathematics tests require students to explain their answers rather than simply select a response, mathematics teachers reported increased emphasis on explanation in their classes (Taylor et al., 2003). Similarly, when language arts assessments included a component that required students to write essays, teachers often responded by increasing the amount of class time devoted to writing (Koretz, Barron, et al., 1996; Koretz and Hamilton, 2003; Stecher et al., 1998).

Lane, Parke, and Stone (2002) explored the impact of the Maryland School Performance Assessment Program (MSPAP) and the Maryland Learning Outcomes (MLOs) on instructional practices and found that the majority of teachers credited the MSPAP with having a moderate or great amount of impact on the content they presented. In particular, a majority of teachers in the study reported an increased emphasis on mathematical problem solving, reasoning, and communication since the introduction of the MSPAP. In a small study of classroom-based performance-assessment-driven instruction, Fuchs et. al (1999) also found that teachers implementing performance assessments shifted their curriculum away from basic, isolated, and routine content toward mathematical problem solving and communication.

While a focus on higher cognitive levels may be the goal of a given performance assessment, these effects are not always as potent as intended. In a follow-up study regarding student views of MSPAP, Parke and Lane (2007) found that students most frequently reported short answer and textbook tasks occurring in their classrooms. Tasks involving real-life application were least reported by students. Furthermore, Stecher (2010) cautioned that, while performance assessments may reduce the curriculum-narrowing effects of high-stakes testing, they are "not immune" to these effects. For example, Stecher and Mitchell (1995) found that teachers in Vermont were engaging in "rubric driven instruction," meaning that rather than focusing on problem solving in the larger sense, aspects of problem solving that led to higher scores on the state rubric were emphasized instead.

Changes in Instructional Activities

Testing not only influences *what* teachers teach, but in some cases it can affect *how* they teach. Several studies identified changes in the ways teachers convey content in their classrooms. These include engaging in test-preparation activities, adopting new instructional strategies, and changing assessment practices.

Focus on Test-Taking Strategies

Some of the actions educators take to prepare students for tests focus less on the content and skills assessed but more on the format and structure of the test, explaining strategies students can use to perform well on certain types of test items. Common activities include coaching students using similar items to the test, having students take a sample test or work through released items, and using commercial test preparation materials (Nichols and Berliner, 2005; Amrein and

Berliner, 2012; Rentner et al., 2006; Firestone, Mayrowetz, and Fairman, 1998; Burger and Krueger, 2003). Teachers' propensity to engage in these activities can be influenced by the actions of district or school administrators. In a RAND study of NCLB, for example, sizable majorities of principals in the three participating states reported distributing commercial test-preparation materials and copies of released test items for use by teachers (Hamilton et al., 2007). Although time spent on these strategies can detract from instructional time that could be spent on engagement with higher levels of academic content or more complex tasks, it is important to recognize that some effort to help students become comfortable with the testing format might be necessary to ensure that students are able to display their knowledge and skills on the test; therefore, it might actually enhance the validity of scores.

Changes in Instructional Strategies

Teachers have to make decisions about how to present content to their students—for instance, whether to adopt a lecture-style format and whole-class discussion, or whether to take a more student-centered approach that relies on small-group discussion and student-initiated projects. While there is no definitive evidence that one instructional approach is more effective than others in all contexts, it is important to understand how teachers' pedagogical strategies might be shaped or altered by assessments. A variety of other factors might also influence teachers' instruction, including the curriculum materials that they have adopted and the characteristics of the students in their classes, but there is evidence that testing can have an effect on teachers' choice of instructional strategies. The evidence suggests that instruction changes to emphasize the kinds of skills measured by the test—be they disaggregated, basic skills or more integrated performances.

Many studies identified instances of teachers using more traditional teaching practices, such as lecturing, in response to high-stakes tests, and the use of these practices frequently overlaps with a focus on basic skills. Within the United Kingdom, the House of Commons, Children, Schools and Families Committee (2008) reviewed many studies and reports that found that teachers, in the face of high-stakes tests, tended to focus on promoting basic skills while devoting less attention to helping students develop creativity and imagination. Harlen and Crick (2002) conducted a systematic review of the literature on summative assessments and student learning and found that instructional activities tended to be highly structured and emphasize transmission of content where there was a strong emphasis on summative assessment.

Researchers found that teachers frequently used direct instruction and lecture in the context of emphasizing tested facts and basic skills (House of Commons, Children, Schools and Families Committee, 2008; Harlen and Crick, 2002; Johnston and McClune, 2000; Assessment Reform Group, 2002; Fickel, 2006; Smith, 2006; Au, 2007). Other changes documented in the literature include increased emphasis on whole-class instruction (McNess et al., 2001) and worksheets (Smith, 1991), and decreased emphasis on inquiry and collaborative learning (Smith, 2006; Cimbricz, 2002; Watanabe, 2007) in response to testing.

Vogler (2006) conducted a survey of high school social studies teachers in Mississippi to explore the extent to which teachers use more traditional, teacher-centered practices and tools over student-centered ones in response to a high-stakes multiple-choice test. Five of the six most frequently used teaching practices were teacher-centered (e.g., textbooks, multiple-choice questions, visual aids, lecturing, textbook-based assignments), and six of the seven least used teaching practices were student-centered (e.g., journals, role playing, group projects, project-based assignments, computers/educational software, problem-solving activities).

While much of the research discussed above suggests that teachers may respond to high-stakes multiple-choice testing by relying on traditional or teacher-centered instructional practices, the literature around performance assessments reveals a potential shift in a different direction. For example, in a study of portfolio use in 24 teacher education programs, Anderson and DeMeulle (1998) found that 92 percent of teacher educators surveyed reported that portfolio use had an impact on their teaching, including making their practice more student-centered. In fact, this type of shift is one of the rationales that early advocates of performance assessment offered—the idea that well-designed tests could be "worth teaching to" and could therefore promote more cognitively demanding instruction (see, e.g., Resnick and Resnick, 1992). Some studies have found that teachers reported responding to performance assessments with an expanded repertoire of teaching strategies and techniques (Falk, Ort, and Moirs, 2007; Fuchs et al., 1999; Adair-Hauck et al., 2006). In contrast to his 2006 study, Vogler (2002) found that since the public release of student results on a high-stakes performance assessment, teachers had reported an increase in their use of open-response, creative, and critical thinking questions in the classroom, an increase in inquiry and investigation activities, and a decrease in the use of textbook-based assignments and lecturing.

Formative assessment—instructionally embedded assessments intended to inform teacher practice—may also influence the pedagogical strategies employed in the classroom. For example, some studies provide evidence that developing teachers' formative assessment practices can in turn influence their questioning and feedback practices in the classroom (Harrison, 2005; Black and Wiliam, 2005). However, even when an assessment is intended to alter the scope or focus of instructional strategies in the classroom, this does not always happen with the depth or breadth that is expected. In a qualitative study of classroom-based, formative Assessment for Learning (AfL) practices, Marshal and Drummond (2006) found that many teachers enacted AfL practices in their classroom only superficially. Video analysis revealed that only about one-fifth of the 27 recorded lessons captured the "spirit" of AfL's foundational principle of student autonomy, while the remaining lessons stuck to "the letter of the rules" of AfL without embodying that same principle on a deeper level.

Moreover, although the research on how testing affects the content of instruction has consistently shown that teachers alter what is taught in response to what is tested, the evidence regarding shifts in *how* that content is presented is mixed. Firestone, Mayrowetz, and Fairman (1998) looked at how middle school math teachers in Maine and Maryland responded to their

state performance assessments. While the teachers placed more emphasis on the test through focusing on tested content and test-taking strategies, the study found that the teachers' pedagogies that would support the development of higher-order thinking were essentially unchanged. In a study by Diamond (2007), teachers reported that testing influenced the content of their instruction but was not a major influence on the strategies they selected for presenting that content, and in particular whether they used strategies that emphasized interaction, communication, and discussion.

Changes in Classroom Assessment Practices

In addition to the potential impact that high-stakes assessment may have on teachers' day-to-day instructional strategies, some research suggests that teachers' classroom-based assessment practices may be influenced as well. Some studies found that teachers changed their classroom assessments to mirror the format of the high-stakes test (Grant, 2001; Yeager and Pinder, 2006; Fickel, 2006; van Hover, 2006; Ehren and Star, 2013). Other studies found that teachers tended to shift their focus toward using summative assessments that emphasize scores over using the test results to inform the learning process (Assessment Reform Group, 2002; Harlen and Crick, 2002).

There is some evidence that the implementation of performance assessments where more complex student work is scored using rubrics that describe the features of performance at different levels may encourage teachers to incorporate similar rubrics as an assessment device in their classrooms (Adair-Hauck et al., 2006; Vogler, 2002). Other studies suggest that teachers may respond to formative assessment efforts by using peer- and/or self-assessment in the classroom (Black and Wiliam, 2005; Frohbieter et al., 2011; Harrison, 2005).

Changes in Teachers' Interactions with Students

In addition to changing teachers' instructional strategies and curricular focus, testing sometimes influences the ways in which teachers allocate their time, resources, and attention among their students. Research suggests that testing can encourage teachers to focus on meeting specific students' needs by individualizing instruction for all students or by shifting attention toward students whose results "count" more in the accountability system.

Using Test Results to Individualize Instruction

In the current accountability context, student-level data has become abundant, as has an emphasis on *using* these data to inform instruction (Hamilton et al., 2009).

Several studies have found that teachers use the information gained through assessment to identify student needs and misconceptions. Shepard and Dougherty (1991), for example, found that many teachers believed that results from their districts' high-stakes tests helped them identify student strengths and weaknesses, and that the results also helped attract resources for

the students who needed them. Similarly, Falk, Ort, and Moirs (2007) described research indicating that New York teachers found that the Early Literacy Profile—a large-scale, classroom-based performance assessment—informed their instruction by providing immediate feedback on student learning, giving them a clearer and richer sense of student knowledge and progress. Many teachers also reported relying more on the evidence gained through the assessments than their own subjective feelings as the basis for instructional decisions. At the higher education level, surveys administered in a study of the statewide implementation of the Performance Assessment for California Teachers (PACT) revealed that many teacher educators and teacher training programs provide more support and guidance in candidates' weak spots identified by the performance assessments (Pecheone and Chung, 2006).

Teachers often reported that scores from large-scale, end-of-year tests were less useful for instructional planning than scores on interim or classroom-based tests that provided more frequent information that is more closely aligned to their curriculum (Marsh, Pane, and Hamilton, 2006). One might think that formative and interim assessments would have a stronger effect on instruction than do end-of-year accountability tests because the former are more frequent and more closely tied to curriculum, and research suggests that the use of such assessments does indeed give teachers insight into their students' skill and understanding (Black and Wiliam, 2005; Oláh, Lawrence, and Riggan, 2010; Goertz, Oláh, and Riggan, 2009; Shepard, Davidson, and Bowman, 2011). However, Goertz, Oláh and Riggan (2009) found that while teachers did access and analyze interim assessment data, beyond helping them decide what to reteach and to whom, the information garnered from assessments did not really change *how* specific content or students were taught. Indeed, this gap between identifying student needs and changing instructional practices to *address* student needs is one of the most consistent themes throughout the assessment literature (Hamilton et al., 2009; Oláh, Lawrence, and Riggan, 2010; Shepard, Davidson, and Bowman, 2011; Heritage et al., 2009; Lanting, 2001).

How teachers use information gained through student assessment may be influenced by how teachers interpret the information (Coburn and Turner, 2011; Knapp et al., 2006). Frohbieter et al. (2011) found that teachers gained varying levels of information from formative assessments, ranging from binary judgments about skills and knowledge, to highly nuanced insights into student understanding. In a study exploring the use and impact of interim assessment data in elementary schools in the School District of Philadelphia, Christman et al. (2009) found that teachers were involved in three different types of "sense-making" as they discussed and interpreted assessment data. Most common was *strategic* sense-making, in which teachers identified short-term strategies to help schools reach adequate yearly progress (AYP) targets. *Affective* sense-making focused on teachers' sense of agency and collective responsibility, and their personal beliefs. Least common was *reflective* sense-making, which involved questioning and evaluating instructional practices and what teachers needed to learn in order to help their students succeed. The authors found this final form of sense-making to be particularly promising in terms of improving actual instructional practice.

Focus on "Bubble Kids"

The "strategic" use of assessment data to focus on students who are likely to count more in an accountability context is one of the potentially negative ways in which information gained through assessment may impact teachers' instructional choices. Booher-Jennings (2005) identified this as part of the process of "educational triage" in which Texas elementary school teachers focused resources on "bubble" students thought to be on the threshhold of passing the test. She described this process as one in which teachers divert resources toward students who are most likely to increase pass rates. Large-scale teacher surveys suggest that emphasis on bubble kids became a fairly widespread phenomenon in response to the use of performance categories or levels rather than scale scores (i.e., reporting that places students into a category such as "basic" or "proficient" rather than assigning a score along a numerical continuum) (Hamilton et al., 2007; Stecher et al., 2008; Pedulla et al., 2003), and Amrein and Berliner (2012) identified a similar phenomenon in reviewing qualitative studies of the effects of high-stakes testing. They referred to this as focusing on "borderline" students, who are on the border of passing or failing a test.

Amrein and Berliner also reported that the focus on borderline students came at the expense of students deemed likely to fail the test regardless of the support they received. Wiggins and Tymm (2002) noted a similar occurrence in schools in England where borderline students were given extra resources at the expense of those students who were not likely to pass at all, or who were likely to pass without any additional support. They noted that this behavior might increase the average test score, but at the expense of certain students.

In other cases, specific students were targeted for extra support, but they were not always the bubble or borderline students. Ferman (2004) found that high school teachers in Israel provided more intensive support and allocated more teaching time to the lowest level students to increase the chances they would pass the test. Jacob's (2005) study of the effects of high-stakes testing on student achievement among high school students in Chicago found that low-achieving students made greater gains, relative to their high-achieving peers, in high-stakes subjects than in low-stakes subjects. He inferred that for the lowest-achieving students teachers shifted their resources away from low-stakes subjects to high-stakes subjects.

4. Conditions That Influence Educators' Responses to Assessment

As Chapter Three shows, research suggests that educators frequently alter their practices in response to assessments, but research also indicates that those changes are mediated by a number of factors. We remarked earlier in the report that much of the research on testing has occurred in the context of accountability, i.e., where there are important consequences associated with test results. In this section we describe the conditions research indicates affect the impact assessments might have on practice. Table 4.1 summarizes the mediating factors identified in our literature review, roughly organized in terms of the conceptual framework discussed earlier in this report (Figure 2.1).

Table 4.1.
Factors That Mediate the Relationship Between Assessment and Instructional Practices

Attributes of the Tests and Testing Programs
• Purpose and use of test
• Quality of the assessment
• Testing format
Accountability Context
• Nature of consequences
• Accountability metrics and decision rules
Educator Background, Beliefs, and Knowledge
• Domain knowledge
• Teacher beliefs
• Familiarity with assessment
• Endorsement of assessment
School and Student Characteristics
• School characteristics
• Aggregate student performance
District/School Policy
• Use of time
• Professional development
• Collaboration
• Curriculum

Attributes of the Tests and Testing Programs

Our review of several different bodies of assessment literature sheds some light on ways in which the attributes of a given assessment might mediate the effects of assessment on instructional practices. However, the amount of research on test attributes is limited, and the research has been conducted in a wide variety of contexts involving a wide variety of tests. Thus, while the findings are interesting, few have been replicated.

Purpose and Use of Test

Research has found that the purpose of an assessment may influence the ways in which the assessment affects teacher practice. Tests that are explicitly intended to shape instructional practice may be more likely to promote changes in instruction than tests that are used for other purposes, such as placing students in programs. Before exploring relationships between intended purpose and instructional practices, it is important to understand the kinds of decisions and inferences tests are designed to support. Perie, Marion, and Gong (2009) differentiated between summative, interim, and formative assessments in terms of purposes. They claimed that large-scale summative assessments serve primarily evaluative purposes and do not necessarily lend themselves to being instructionally useful. In contrast, classroom-based formative assessment is intended to be used specifically for the purpose of diagnostic teaching. Interim assessments fall in between, and can be designed to both inform instructional practice at the classroom level and to tell a broader story of assessment results at an aggregated level. Furthermore, Perie, Marion, and Gong (2009) asserted that few assessments can serve several purposes well and that

> if policy makers want an assessment to help educators improve instruction, they should look for one that ties directly to the classroom instruction . . . actually, if this is the sole goal of the assessment, we argue that resources would be better spent helping teachers learn formative assessment techniques. (p. 13)

Several features of assessments influence their utility for instructional decisionmaking. Research has found that, to be useful to teachers, assessments should be tightly aligned with the curriculum and, ideally, should be linked to guidance to help teachers identify strategies for responding to the data they produce—e.g., sample lesson plans that focus on specific content strands in the test (Perie, Marion, and Gong, 2009; Hamilton, Stecher, and Yuan, 2012). The timeliness of score reporting is also critical to a test's instructional impact: Wright (2002) found that teachers in California could not use assessment results from the summative state test to inform their instruction because they did not receive the results until the end of the school year or after the year ended.

Research suggests that the clear and consistent communication of an assessment's purpose to educators is also important. Shepard, Davidson, and Bowman (2011) suggested that different understandings between districts and teachers about the goals of interim assessments may be a limiting factor in teachers' deeper use of assessment data. Similarly, Goertz, Oláh, and Riggan

(2009) highlighted the importance of setting strong expectations for the instructional use of interim assessment data. In a case study of a middle school teacher, Buck and Trauth-Nare (2009) suggested that framing formative assessment as being one component of a larger effort to prepare students for high-stakes testing may have alleviated the sense of pressure the teacher felt to forgo time adjusting instruction to respond to student needs in favor of more time spent on test preparation.

It is also important to note that in the accountability context, the specific ways in which accountability decisions are linked to test scores can influence instruction. One of the purposes of NCLB testing has been to identify students who are below a certain threshold (called "proficient") so that these students can be given additional support, and so that schools with large percentages of these students can be subjected to interventions and sanctions. The "bubble kid" phenomenon described earlier reflects this decision to require states to use proficiency as a threshold for determining consequences. The location of the "proficient" cut score, and the decision to use such a cut score at all, influence educators' decisions regarding whether and how to reallocate attention from some students to other students.

Quality of the Assessment

The extent to which the test measures the intended constructs and produces trustworthy scores for all students has implications not only for the quality of information it produces, but for the ways that it influences teachers to shape students' educational experiences (Kober, 2002; McNess et al., 2001; Burger and Krueger, 2003). Test quality mediates the impact of testing on practice in two ways. First, low-quality tests—i.e., tests that have large amounts of error, that don't support appropriate inferences, or that unfairly favor certain groups of students—produce misleading information that can lead to poor instructional decisions. Second, low-quality tests have the potential to reduce educators' confidence in them and reduce their impact (Burger and Krueger, 2003). Watanabe (2004) noted the importance of "face validity," the extent to which the test appears to be an authentic reflection of classroom practice. He posited that the presence of face validity is more likely to lead teachers to change their instruction in positive ways. Face validity is also relevant to teacher buy-in, discussed below.

Testing Format

As we discussed in the sections on changes in curriculum and instruction, the format of the test is an important mediating factor: Multiple-choice tests have often been accompanied by increased emphasis on basic skills, for example, whereas performance-based assessment has been associated with greater focus on problem solving and inquiry (Cimbricz, 2002; Smith, 1991; Vogler, 2006; Looney, 2009). Ehren and Star (2013) found that elementary school teachers in Boston and New York targeted their instruction based on differences in item format within the same test. Teachers looked at how students performed on multiple-choice and open-response sections and provided various instructional strategies to target those format types (e.g.,

identifying distractors for multiple-choice tests, explaining answers in math journals for open-response questions). The format of the test items, and the relative emphasis given to different item formats within a test, is clearly an important consideration when trying to predict how it will influence instruction.

The format of the test sends a message regarding the kinds of tasks in which students are expected to engage, and therefore it can influence teachers' choices regarding curriculum and instruction. For example, the idea that more authentic assessment tasks can promote instructional improvement is one of the driving forces behind the development and implementation of performance assessment. Advocates of performance assessments argue that they signal to teachers that more complex skills and more authentic tasks should be part of their instruction. A similar sort of logic is used to encourage more formative assessment, i.e., assessment that is embedded within a given learning activity and linked directly to the current instructional context of the classroom (Perie, Marion, and Gong, 2009). Advocates of formative assessment argue that building assessment directly into instruction makes the information more immediately actionable and provides insights teachers can incorporate directly into instructional planning. However, individual performance assessments and formative assessments vary widely in design and in quality, and they can have widely different impacts on instruction (Shepard, Davidson, and Bowman, 2011; Bennett, 2011; Black and Wiliam, 1998a; Stecher, 2010; Perie, Marion, and Gong, 2009).

It is also important to note that some researchers have found contrary evidence, i.e., that changing the testing format in and of itself is not sufficient to change teacher practice (Watanabe, 2004; Cheng, 2004; Levinson, 2000). Cheng (2004) looked at the effects of a high-stakes test in Hong Kong and found that teachers made few changes in their teaching practice as a result of the test. The researcher suggested that while teachers respond to high-stakes tests, changing the nature of the exam (e.g., multiple-choice, performance-based) alone will not necessarily enable teachers to teach differently. This implies that other supports must be provided to achieve desirable changes in instruction. Several of these supports are discussed elsewhere in this section.

Accountability Context

Nature of Consequences

Research shows that the consequences that are attached to test performance influence teachers' reactions to tests; in particular, educators respond to high-stakes assessments differently than to lower-stakes assessments (Firestone, Mayrowetz, and Fairman, 1998; Pedulla et al., 2003). However, the extent to which stakes are considered to be "high" is subject to the interpretation of teachers, administrators, parents, and students in the system. Merely publishing school-level scores might not generally be considered a high-stakes situation, but for the school administrator

who anticipates angry parents' responses to low scores, the stakes might in fact feel quite significant. Teachers in Maryland, where test scores were published but no consequences were attached, changed practice as much as teachers in Kentucky, where test results were tied to financial rewards or penalties for schools (Koretz, Mitchell, et al., 1996; Koretz, Barron, et al., 1996; Koretz, 2000). After 2001, NCLB required that states attached significant consequences for schools to test scores, and since Race to the Top, a growing number of states are requiring that student test scores be factored into measures of teacher effectiveness. Thus, increasing numbers of educators operate in an environment where test results have high stakes.

Accountability Metrics and Decision Rules

Even in the high-stakes environment of NCLB, specific features of the accountability system, including the grade and subjects tested, who is held accountable, and what types of metrics are used for decisionmaking (e.g., cut scores versus continuous scale scores, gain scores versus status scores) mediate how the district, school, and teachers will respond to the assessment. For example, the setting of the performance threshold will mediate teachers' reactions to tests. As we noted in Chapter Three, there is evidence that teachers focus more attention on the students who are on the borderline between achieving proficiency or not (Booher-Jennings, 2005; Hamilton et al., 2007; Stecher et al., 2008; Pedulla et al., 2003; Amrein and Berliner, 2012). Similarly, we would expect teachers of subjects that are not included in the accountability system to be influenced by the test in a different way than teachers of subjects that are included, and research has found teachers' reactions are based in part on which subjects are included in accountability computations at different grade levels (Stecher et al., 1998).

Many states and districts are adopting new teacher evaluation systems that require all teachers to be evaluated, based in part on their students' test scores. However, the features of the tests used in the evaluation system vary across grades and subjects (e.g., typically only a subset of teachers have students who take the state tests, whereas others may use district- or teacher-developed assessments). In addition, the weight given to test results relative to other sources of evidence, such as classroom observations or student feedback, can vary from site to site. As a result, the responses of teachers are likely to vary depending on the types of tests that are administered in a particular grade and subject and on how the test scores are factored into the effectiveness metric.

Educator Background, Beliefs, and Knowledge

Not surprisingly, the literature suggests that the characteristics of the teachers themselves may affect whether and how assessments influence their instructional practices. In a review of research on testing, Cimbricz (2002) found that the relationship between state tests and teacher practice is mediated by a number of teacher characteristics, including content knowledge, approaches to teaching, beliefs about teaching and learning, and prior experience. The most

frequent mediating factors we found in our review of the literature were the depth and breadth of domain knowledge, beliefs about curriculum and instruction, familiarity with the assessment, and endorsement of the assessment.

Domain Knowledge

Research suggests that teachers who have greater content and domain knowledge are more likely to make positive changes in instruction in response to assessment results. This finding was particularly strong in the context of formative assessment where researchers found a relationship between strong domain knowledge and effective use of assessment data to inform teaching (Bennett, 2011; Frohbieter et al., 2011; Goertz, Oláh, and Riggan, 2009; Heritage et al., 2009). For example, Jones and Moreland's 2005 case study of New Zealand teachers who teach technology curricula explored the connections between teachers' pedagogical content knowledge and their classroom practices. Through a series of interventions (such as professional development and classroom support) intended to increase teachers' domain knowledge of technology, the authors observed improvement in teachers' use of formative assessment practices, including the provision of more targeted feedback to students in technology classes. The high-stakes assessment literature also suggests that content knowledge plays a large role in influencing instruction. Salinas (2006) observed that the pedagogical content knowledge of high school social studies teachers in Texas mediates between the assessment and teachers' changes in instruction, as teachers' knowledge about strategies for teaching history influence their instructional choices in light of assessment results.

Teacher Beliefs

Research suggests that alignment among teachers' beliefs about learning, content, pedagogy, and assessment can reinforce or encourage instructional change, and that misalignment may impede or alter change processes. Formative assessment research has also found that the personal beliefs, interests, and assumptions of teachers may mediate the influence of assessments on instructional practices (Buck and Trauth-Nare, 2009; Marshal and Drummond, 2006). For example, Lee, Feldman, and Beatty (2012) explored factors that may impede teachers' use of a technology-enhanced formative assessment (TEFA) and found that teachers must reconcile their use of TEFA with their perspectives and philosophies about teaching and learning, their attitudes and confidence, and their resistance to change. In their seminal literature review of formative assessment research, Black and Wiliam (1998a) asserted that because of the tight tie between embedded, continuous formative assessment practices and other elements of a teacher's pedagogy, the effective implementation of this type of formative assessment calls for, "deep changes . . . in teachers' perceptions of their own role" (p. 20). Such changes, which are elaborated in Black and Wiliam (1998b), include movement toward teaching through ongoing interaction with students (rather than through a one-way transmission of knowledge from teacher to student).

The high-stakes testing literature also mentions the importance of teachers' beliefs about content and pedagogy as another mediating factor (Nichols and Berliner, 2005; Watanabe, 2007; Fickel, 2006). Pedulla et al. (2003) found that a majority of teachers, particularly at the elementary level, felt that high-stakes tests lead them to teach in ways that conflict with their conception of sound instruction. The RAND studies of NCLB produced similar findings for the sample of teachers included: Fewer than a third of teachers in California and Pennsylvania, and about half in Georgia, agreed that the state accountability system supported their personal approach to teaching and learning (Hamilton et al., 2007). Borko and Eliot (1999) conducted a case study of the response of two elementary teachers in Kentucky to a high-stakes portfolio assessment. They found that the teachers' pedagogical beliefs and testing requirements conflicted and required teachers to reconcile the two in their instructional practice.

Familiarity with Assessment

Teachers' familiarity with the assessment appears to be a mediating factor between assessments and classroom practice, which can potentially lead to beneficial or detrimental practices in different situations. For example, Frohbieter et al. (2011) observed that teachers' familiarity with formative assessment systems accompanied greater integration of formative assessment practices into their instruction. Burger and Krueger (2003) reported that as teachers become increasingly knowledgeable about testing and testing issues, they were better able to apply that knowledge to more effective instructional strategies. However, at the same time, they cautioned that such knowledge can lead to practices that improve test scores without improving student learning. The number of years that an assessment has been used in a school, district, or state may also be a potential mediating factor in its usefulness. Parke and Lane (2007) suggested that the fact that a statewide performance assessment had been in place for over a decade may have enhanced that assessment's positive impact.

Endorsement of Assessment

Teacher familiarity with a given assessment does not necessarily equate to teacher endorsement of the utility and intent of that assessment. In fact, research suggests that "buy-in" for an assessment supports changed instructional practices (Darling-Hammond and Rustique-Forrester, 2005; Buck and Trauth-Nare, 2009; Jones and Moreland, 2005; Parke and Lane, 2007). This is true for teachers as well as other educational stakeholders, including school leaders, students, parents, and policymakers.

Not surprisingly, teachers appear to have strong opinions about externally mandated, high-stakes assessments, and these opinions may influence their practices. Some research suggests that teachers find standards and assessments beneficial to their practice by helping them focus their instruction and obtain feedback about the effects of that instruction (Mabry et al., 2003). Other studies, however, indicate that some teachers do not share the perception that tests are instructionally useful (Clarke et al., 2003; Taylor et al., 2003), and that many teachers question

whether large-scale tests are accurate measures of the skills and knowledge and of their students, particularly in the case of special education students, minority students, and English language learners (Pedulla et al., 2003). Falk, Ort, and Moirs (2007) found that while teacher dislike of or misconceptions about a large-scale performance assessment (intended to both monitor student progress and to provide teachers with instructionally useful information) hindered the successful implementation of that assessment in the classroom, slower roll-out to increase teacher buy-in was an enabling condition for that implementation.

School and Student Characteristics

Researchers have posited that characteristics of the schools and of the students they serve play a role in mediating the influence of testing on classroom instruction.

School Characteristics

Perhaps the most frequently researched school characteristic in this context is grade configuration. Some studies found variation in the effects of state tests across grade levels, with stronger effects in elementary school than in secondary school (Smith, 1991; Au, 2007; Yeh, 2005; Pedulla et al., 2003; Lane, Parke, and Stone, 2002). Pedulla et al. (2003) found that elementary and middle school teachers reported much larger effects of testing programs on their instruction than did high school teachers. Additionally, elementary teachers reported feeling that the state test was less compatible to their curricula than did high school teachers. The authors posited that this may be related to the fact that elementary school teachers teach multiple subjects, which leaves more room for discrepancies between the test and curriculum. In a metasynthesis of the literature on the influence of testing on instruction, Au (2007) found that certain changes in teacher instruction, such as the narrowing of content, were most prevalent at the secondary level. In another instance, Yeh (2005) interviewed 61 teachers and administrators across four Minnesota school districts about their experiences with state-mandated testing in general (i.e., not specific to any subject). While interviewees reported that the effects of the test on instruction varied by grade level, this was more related to the attributes of the test at each grade than the actual grade level. While elementary teachers focused their instruction on critical thinking and 8th-grade teachers focused on basic skills, the author noted that this was likely related to the types of content and skills that their respective tests emphasized. Thus, the relationship between testing effects and grade configuration may be driven by the characteristics of the test and the way instruction is organized rather than grade level (or age) per se.

Additional school characteristics that have been identified as potential mediators of testing effects are urbanicity and governance (i.e., whether the school is a traditional public school, a charter school, or a private school). Wiggins and Tymms (2002) hypothesized that differential effects of testing on Scottish and English schools may have been related to the school's location. More specifically, they suggested that parental choice in urban versus rural areas may affect the

extent to which the schools believe the national tests have dysfunctional consequences on schools and instruction. Hayes and Read (2004) looked at the effects of an English language proficiency test used for university admissions in New Zealand on preparation courses for the exam. They posited that differences in the courses may be attributed in part to whether a school is private or public. The researchers suggested that private schools may feel more pressure for students to pass the test and change their instruction accordingly. It is important to recognize that the literature on how school characteristics, such as urbanicity and governance, affect educators' responses to testing is sparse. Strong conclusions about these effects are not warranted, and the effects are likely to vary depending on the features of the accountability system (e.g., which grade levels are held accountable and whether private schools are included in the system).

Aggregate Student Performance

Teachers in schools with higher percentages of low-performing students seem to react to high-stakes testing differently than teachers in higher-performing schools. For example, a Government Accountability Office analysis of data from the RAND NCLB studies showed that a variety of responses, including focusing more on tested topics and using test-score data for decisionmaking, were more commonly reported among teachers at high-poverty and high-minority schools than at other schools (GAO, 2009). These differences are likely attributable to the fact that high-poverty and high-minority schools are often at greater risk of failing to meet their accountability targets than are other schools. Along these lines, researchers have found some evidence that teachers at low-performing schools change their instruction more as a result of the test than do teachers at higher-performing schools (Jones et al., 1999; Amrein and Berliner, 2012). These changes might reflect the greater pressure felt by teachers in low-performing schools, but could also stem from the need to engage in greater amounts of remediation when students lack the skills necessary to do grade-level work. Wright (2002) provided an example from a study of a California elementary school where teachers of low-performing students reported having to teach more content and skills in a shorter period of time than they would have if their students had been higher-performing. The need to engage in remediation is not, however, simply a result of the test; even without the test, teachers would need to provide such remediation if they hope to help these students catch up to their higher-performing peers.

Policy and Practice

Many of the factors we have identified as mediators are themselves influenced by school or district policies, including the use of time, policies related to professional development (both in-service and pre-service training), the collaboration that it can foster, and choices related to curriculum.

Use of Time

Time manifested itself as a mediating factor in a variety of ways in the literature on testing. Many researchers noted that changes in instruction are more obvious in the weeks and months before the test (Jones et al., 1999; Smith, 1991; Amrein and Berliner, 2002). During this period, teachers tended to focus on instruction that is focused on raising test scores as opposed to promoting long-term student learning. Additionally, many researchers noted that a lack of instructional time played a large role in mediating the effects of testing on instruction. Lack of sufficient teaching time manifested itself in terms of focusing only on tested subjects (McNess et al., 2001; Koretz, 2005) or narrowing or condensing of content (Wright, 2002; van Hover, 2006; Fickel, 2006). Additionally, teachers felt that they had to take time away from other activities—including classroom-based assessments—in order to engage in test preparation that doesn't always increase student learning (Boardman and Woodruff, 2004; Jones et al., 1999; Buck and Trauth-Nare, 2009; Falk, Ort, and Moirs, 2007). Preparation time came up as an influential factor in how teachers used data to inform their instruction. Teachers reported needing sufficient time to interpret, reflect on, and then act on the data gathered through formative assessments (Goertz, Oláh, and Riggan, 2009; Bennett, 2011).

Professional Development

Professional development was consistently highlighted as an enabling condition for assessment to influence teaching practice, particularly in the literature on performance assessment and formative assessment (Darling-Hammond and Rustique-Forrester, 2005; Hamilton, Stecher, and Klein, 2002; Perie, Marion, and Gong, 2009; Hamp-Lyons, 2007; Kober, 2002; Commission on Instructionally Supportive Assessment, 2001). Though professional development was not always the focal point of research, authors of empirical studies on responses to testing frequently pointed to teacher training, support, and capacity-building (or lack thereof) as an important determinant of changing practice (Shepard, Davidson, and Bowman, 2011; Fuchs et al., 1999; Buck and Trauth-Nare, 2009; Furtak, 2012; Watanabe, 2004; Yeager and Pinder, 2006). Teachers themselves also pointed to professional development as an important support in their efforts to improve their teaching practices in response to assessments (Vogler, 2002; Dekker and Feijs, 2005). However, although high-quality professional development is generally considered a critical factor in promoting effective implementation of education reforms, there is little empirical evidence that provides guidance on the amount and types of professional development that would promote constructive responses to assessment. Some researchers have reported that the duration of professional development may influence its impact on classroom assessment practices. Black and Wiliam (2005) found that sustained commitment for at least two years is needed to develop teachers' formative assessment practices. Jones and Moreland (2005) reported similar findings—that the three-year timeframe of professional development activities intended

to improve teachers' assessment for learning practices positively influenced the ultimate effects of those activities on instruction.

Another noteworthy finding is that testing itself can provide a form of professional development when it includes opportunities for teachers to score open-ended assessments, and this can influence practice (Darling-Hammond and Ducommun, 2010; Falk and Ort, 1997; Kitchen et al., 2002; Parke and Lane, 2007). However, some have criticized such findings as being overstated (Goldberg, 2012; Goldberg and Rosewell, 2000). In a review of literature from the past two decades, Goldberg (2012) summarized the benefits most often cited by teacher-participants in studies of scoring as professional development including:

> The clarification of standards, identification of desirable instructional practices based on examination of student work, increased assessment literacy that can inform classroom assessment practice, and deeper appreciation of the manifold ways that students might successfully demonstrate what they understand and what they can do. (p. 39)

However, Goldberg also argued that what teachers learn from scoring experiences "has tended to center around the assessment itself, rather than on broader implications for instructional content areas and domains being assessed" (p. 44). For example, in a small study investigating the impact of the Maryland School Performance Assessment Program (MSPAP) scoring experience on teacher practice, Goldberg and Rosewell (2000) found that while teachers generally endorsed the scoring experience and that teacher-scorers were more likely than their nonscoring colleagues to engage in some classroom activities, such as cueing for multidisciplinary thinking, such changes in classroom practice were limited and frequently did not connect to the state-mandated learning outcomes. The authors suggested that scoring may not be sufficient professional development on its own, and should perhaps be coupled with sustained and robust professional support that encourages teachers to see how their testing and scoring practices fit into the larger context of the performance standards they seek to assess.

Collaboration

Professional development that includes teacher collaboration around testing was frequently mentioned as a positive influence on teacher practice. For example, in studying a project intended to develop teachers' formative assessment practices, Harrison (2005) found that professional dialogue between teachers and the opportunity for professional development in a peer-supported environment were two enabling conditions supporting improvements in teachers' instructional strategies. Similarly, Dekker and Feijs (2005) found that formal and informal contact with colleagues helped to sustain the effects of a professional development program designed to change teachers' instruction via formative assessment practices. In other formative assessment research, teacher collaboration was identified as a positive factor helping teachers transform assessment information into instructional improvement. For example, Christman et al. (2009) found that reflective conversations among teachers helped focus their attention "away

from students' failures and toward analyzing and strategizing about their own practices" (p. 48). The literature on high-stakes testing also points to teacher collaboration as an important factor in mediating changes in teacher instruction (McNess et al., 2001; Swanson and Stevenson, 2002). Along these lines, Fickel (2006) found that the collaborative nature of a high school social studies department helped teachers make sense of reforms related to testing and implement them in their classrooms.

Similarly, some studies have found that scoring open-ended responses to tests promotes collaboration, which in turn supports changes in practice aligned with tests. In a study exploring the role of teacher as scorer in a large-scale, standards-based performance assessment, Falk and Ort (1997) found evidence that teacher involvement in scoring offers teachers a space in which they can collaborate and learn from each other. Providing teachers with forums for collaboration was also one of the conditions that Goldberg (2012) cited as being identified by teachers as critical to ensuring the effectiveness of scoring as professional development.

Some research suggests that collaboration between teachers and outside organizations about assessments can also foster positive responses to assessments. In particular, projects exploring collaboration between schools and research centers suggest that these relationships may strengthen assessments' role in productively informing school and classroom processes (Ancess, Barnett, and Allen, 2007; Niemi, Baker, and Sylvester, 2007; Harrison, 2005; Jones and Moreland, 2005; Buck and Trauth-Nare, 2009). In other instances, researchers suggested that interactions between test developers and teachers related to the assessment design process will have a positive influence on both assessments and how the assessments influence instruction (Qi, 2004; Runte, 1998).

Curriculum

The choice of curriculum is another policy that has been frequently identified in the literature as a mediating factor between assessment and practice. Earlier in the report, we discussed curriculum as an aspect of instructional practice that can be influenced by testing, but curriculum also serves as a mediator of the relationship between assessment and teachers' instructional practices. Research suggests that adequate curriculum resources and materials may support the effective use of assessments to improve instruction (Falk, Ort, and Moirs, 2007; Dekker and Feijs, 2005; Bennett, 2011; Segall, 2006), but that teachers sometimes reported that they lack such resources (Adair-Hauck et al., 2006; Guskey, 1994; Vitali, 1993). In particular, there is a need for curriculum resources and other instructional materials that are aligned with the standards and assessments (Kober, 2002; Wright, 2002; House of Commons, Children, Schools and Families Committee, 2008). Much of the discussion surrounding accountability testing has focused on alignment between the tests and the standards, positing that more complex standards such as the CCSS will be accompanied by the development of assessments of deeper learning, but in fact achieving full alignment between tests and standards is difficult because of limitations in the kinds of skills and competencies that can be assessed through available means and because

standards tend to include more content than can be assessed in a reasonable amount of time. Consequently, studies of alignment suggest that in many cases, even when alignment is found to be high because test items can be mapped back to standards, the test actually samples only a subset of the standards and overrepresents the standards that are easiest to test (e.g., those that focus on facts or basic skills) while underrepresenting more complex skills and reasoning (Rothman et al., 2002). In addition, a lack of specificity in the standards can make it difficult for teachers to determine what to teach, and this uncertainty could increase the likelihood that teachers will narrow the curriculum to focus on what is on the test.

One way to address these problems is by providing teachers with curricula that cover a broader range of content that is aligned with the standards, so that teachers who follow the curriculum will expose their students to the full complexity of the standards while also preparing them to perform well on the tests (Hamilton, 2011). The critical factors are that the curriculum be well matched to the standards and the test, and that teachers believe that it is well matched so that they can be confident that teaching the curriculum will promote improved test scores.

There is evidence that teachers' beliefs about the alignment between the test and curriculum influence teachers' decisions about what content to cover. Segall (2006) found that Michigan high school social studies teachers responded to the alignment between the state test and standards by changing the content they taught and their beliefs about how they taught social studies more so than they changed their pedagogies. Firestone, Mayrowetz, and Fairman (1998) found that some high school teachers in Maryland and Maine felt that there was not complete overlap between tested content and the curriculum they taught and that their state tests required them to create additional lessons and units to address tested content that wasn't in their given curriculum. Teachers' perceptions about the extent to which the curricula and tests covered the same content influenced what teachers taught and the amount of time and the relative emphasis they gave those topics.

Other researchers have tried to clarify the meaning of alignment and its implications for teaching. Yeh (2005) suggested that well-aligned state tests and curriculum that cover both basic facts and higher-order thinking skills can help avoid narrowing of curricula. He pointed to Minnesota's state test and curriculum as an example of this. In another case, Looney (2009) pointed out that perfect alignment among tests, standards, and curricula is not always ideal because it can lead to a lack of innovative, high-quality teaching if such instruction does not fit into the aligned standards and assessment. In such cases, defining alignment more broadly in terms of rubrics and exemplars may help to mitigate such problems.

5. Conclusions

The purpose of our review was to summarize what researchers have learned about the impact of testing on classroom practice and determine what lessons can be drawn about the likely impact of deeper learning assessment. The investigation yielded complex findings. Most of the research examined changes that accompanied test-related policies after they were implemented, limiting causal conclusions. Furthermore, most of the research took place in complex contexts where many factors influenced practice, so it was difficult to attribute changes specifically to some aspect of testing. In addition, variability in how educators responded to tests was evident across different studies as well as within individual studies; like most education policies, a specific approach to testing does not induce change uniformly across affected populations. Nevertheless, the research does suggest relationships between test-related policies and classroom practice, as well as mediating factors that influence these relationships, and it provides some guidance for thinking about the ways that new CCSS-aligned tests might affect practice. Some of this guidance is applicable primarily to those who are responsible for developing new tests and the accountability systems in which they are embedded, and some is more relevant to administrators who will influence school- and district-level policy and practice in response to new testing and accountability policies.

Specifically, the literature suggests that new, CCSS-aligned assessments are likely to promote desirable changes in practice when the following conditions are met.

Conditions Relating to the Tests and the Testing Programs

Test content and format should mirror high-quality instruction. A large body of research documents unanticipated and often undesirable changes in practice in response to high-stakes multiple-choice tests, such as excessive emphasis on tested skills and item formats. The format of the tests signals to teachers the kinds of skills they should emphasize and the kinds of assessments they should use. In contrast, some evidence suggests that high-quality performance assessment can encourage teachers to increase their emphasis on the kinds of higher-order skills and processes that are embodied in the CCSS. Yet, performance assessment has been a two-edged sword—it can be more difficult to obtain high-quality results (i.e., reliable and valid scores) using performance assessments because the assessments can be difficult to score consistently and they usually take more time to complete. For a test to have any chance of promoting deeper learning, it is critical that at least a portion of the test reflects learning activities that are consistent with the goals of deeper learning. Some sacrifice in reliability may be appropriate to represent more demanding content and signal its importance to teachers, but the

extent to which reliability and other aspects of technical quality can be compromised depends in large part on the stakes attached to the scores.

Tests should be used only for purposes for which they were designed and validated. Because of financial constraints and a desire to avoid spending large amounts of instructional time on testing, many schools and districts have relied on external tests as a source of information for a variety of decisions, such as how students should be grouped or which students should get promoted to the next grade. Professional testing guidelines caution against using tests for purposes for which they were not designed, and note the need to gather validity evidence that is specific to a particular use and interpretation (see, e.g., American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Using a test for an unintended purpose can lead to subpar decisions and can diminish the utility of the test for its intended purposes, so it is important that users of tests understand what those tests were and were not designed to do. Those who design assessment policy or use test scores for decisionmaking should also monitor the uses and consequences of testing programs over time so that they can identify and address inappropriate uses or unintended consequences. At the time they are launched in 2014–2015, there will be little evidence of the validity of PARCC and Smarter Balanced for particular purposes, but validation research is high on the agenda of both consortia. Users should not make important decisions on the basis of these tests until their validity can be demonstrated.

Score reporting should be optimized to foster instructional improvement. Teachers and other educators must make frequent decisions about what to teach, how to teach it, and how to address the needs of individual students within their classrooms. If tests are intended to support these decisions, they should provide score reports that are tailored to the needs of educators. Important features of score reporting systems in this context include the rapid provision of results, score reports that are clear and accessible to educators who are not technically oriented, and a reporting mechanism that can be tailored to provide information about performance for individual students and relevant groups of students (e.g., English language learners). Reporting mechanisms should also display information about specific skills and knowledge in a way that aligns with the curriculum teachers are using, which requires not only a sophisticated reporting system but also a test that measures these constructs with adequate reliability and validity. Ideally, the curriculum and testing system would be tightly linked so that teachers receive frequent feedback related to the material that they are teaching. For example, users of Smarter Balanced and PARCC formative assessments will receive standards-linked data on student performance periodically during the year.

Conditions Relating to Educator Capacity and Beliefs

Teachers should receive training and support to interpret and use test scores effectively. The research on educators' data use, discussed earlier, suggests that the mere provision of data to

teachers is not sufficient for promoting effective use of that information. Teachers need guidance on how to interpret and respond to the data, and this guidance should be provided on an ongoing basis rather than as a one-shot professional development workshop. If the tests assess skills that are unfamiliar to the teachers, as might be the case with tests aligned to the CCSS and tests of deeper learning, then teachers will need support to improve their own subject-matter knowledge in these areas as well as their skills for using the test-score data to impart this learning to students. Such changes may not occur quickly, so teachers will need time to learn and time collaborate in reviewing student test scores and designing interventions to address them. Even when teachers understand how to interpret test results, they don't always have access to guidance on how to respond in ways that will address the needs of all of their students, so resources such as sample lesson plans that focus on specific knowledge and skills included in the standards or tests can be helpful.

Conditions Relating to the Accountability Context

The test scores should "matter," but important consequences should not follow directly from test scores alone. Research on the ways that high-stakes tests influence instruction suggests that if the test is unimportant or irrelevant to students, teachers, administrators, and parents, it is unlikely to have an effect on instruction. On the other hand, if there are very high stakes attached for schools, teachers, or students, there may be severe "teaching to the test" that does not promote real deeper learning but focuses on superficial features of items. This type of overemphasis on test content or item format can occur even when tests are designed to measure higher-order skills and when they include open-ended or performance-based tasks, particularly if item format and content is predictable from one administration to the next. Thus, undue stakes corrupt practice, which undermines the validity of the scores. Thoughtful planners build in mechanisms to deflect distortion. These mechanisms might include multiple measures that emphasize important outcomes or processes that are not measured by the test, so that the test does not become the sole incentive or source of guidance. Many of the teacher evaluation systems that states and districts are adopting illustrate this principle; they include test scores in addition to direct measures of practice (e.g., through classroom observations) and stakeholder feedback (e.g., through student surveys). If thoughtfully designed, these nontest measures can serve as a check on a tendency to focus excessively on tested content.

If there are externally mandated, high-stakes tests, they should be part of an integrated assessment system that includes formative and summative components. A focus on using data to inform instruction has become common in schools, districts, and charter management organizations across the United States. If the only "data" that are available are annual, externally mandated tests, the system will lack utility for instructional guidance, and many educators might place undue emphasis on that single set of tests. A comprehensive assessment system should provide timely and consistent information that can be used for instructional improvement by

teachers, self-reflection by students, mastery certification, and system monitoring. In such a system, different assessments would address different purposes, but would all be implemented in support of each other (rather than in competition or conflict with each other). The Smarter Balanced and PARCC assessment designs reflect a realization of the importance of integrating formative and summative assessments.

Accountability metrics should value growth in achievement, not just status, and should be sensitive to change at all levels of student performance, not just changes at a single cut point. Under NCLB, schools were judged based on the percentage of students whose test scores met or exceeded the proficient cut score. This approach had a number of negative consequences, as noted earlier in this paper. Accountability indices can be constructed differently to ameliorate these problems. For example, indices based on growth in achievement that also take into account performance all along the achievement scale (rather than just whether a student is above or below "proficient") should provide better information about performance, result in higher levels of buy-in from educators, and be the basis for a set of incentives that may be more consistent with public goals for education than the current system. Most new state and district accountability systems, including those that measure the performance of individual teachers, emphasize achievement growth rather than status. Some tests support measures of growth better than others, and there are many ways to model growth using test scores. Consequently, it is important that the designers of these systems understand what types of measures lend themselves to the measurement of growth, and that they select a growth modeling approach that is well suited to their assessments and to the purposes of the evaluation and accountability system.

Conditions Relating to District/School Policy

Assessment should be one component of a broader systemic reform effort. In isolation, tests can send strong signals to educators about what they should focus on and how they should teach. One way to reduce teachers' tendency to overemphasize tests as a source of instructional guidance is to adopt a coherent system of reforms that starts with the standards and aligns other elements to those standards. These elements include curriculum and instructional materials, professional development and support for teachers, data systems, accountability policies, and strategies for community engagement. Such efforts are unlikely to happen quickly, so policymakers and educators need to devote adequate time to making systemic reform. The Smarter Balanced and PARCC assessments are aligned to the CCSS, a first step toward an integrated systemic approach; users will have to coordinate the other elements of the system to achieve this goal.

Summary

We undertook this study to find out the extent to which new assessments might influence instructional practices and which factors might mediate this relationship, i.e., what changes in

policies or context might make new assessments, particularly assessments of deeper learning, have a great influence on practice. By themselves, tests of deeper learning are likely to have some impact on classroom instruction, particularly if they are adopted as part of an accountability system that involves consequences for educators or students. However, research suggests that the role of tests will be enhanced by policies ensuring that the tests have features to make them helpful for instructional improvement, are part of a larger, systemic change effort, and are accompanied by specific supports to help teachers increase their relevant knowledge and skills and modify their practices.

References

Adair-Hauck, B., Glisan, E. W., Koda, K., Swender, E. B., and Sandrock, P. (2006). The Integrated Performance Assessment (IPA): Connecting assessment to instruction and learning. *Foreign Language Annals*, 39, 359–382.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC.

Amrein, A. L., and Berliner, D. C. (2012). *An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing*. Tempe, AZ: Arizona State University Education Policy Studies Laboratory, Education Policy Research Unit (EPRU).

Ancess, J., Barnett, E., and Allen, D. (2007). Using research to inform the practice of teachers, schools, and school reform organizations. *Theory into Practice*, 46(4), 325–333.

Anderson, R., and DeMeulle, L. (1998). Portfolio use in twenty-four teacher education programs. *Teacher Education Quarterly*, Winter.

Assessment Reform Group (2002). *Testing, Motivation and Learning*. Cambridge, UK: University of Cambridge, Assessment Reform Group.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher,* 36(5), 258–267.

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25.

Black, P., and Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, 5(1), 7–74.

Black, P., and Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa International,* October.

Black, P., and Wiliam, D. (2005). Changing teaching through formative assessment: Research and practice. In Organisation for Economic Co-operation and Development (OECD), *Formative Assessment: Improving Learning in Secondary Classrooms*. Paris: OECD.

Boardman, A. G., and Woodruff, A. L. (2004). Teacher change and "high-stakes" assessment: What happens to professional development? *Teaching and Teacher Education*, 20(6), 545–557.

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

Borko, H., and Eliot, R. (1999). Hands-on pedagogy versus hands-off accountability: Tensions between competing commitments for exemplary mathematics teachers in Kentucky. *Phi Delta Kappan*, 80(5), 394–400.

Buck, G., and Trauth-Nare, A. (2009). Preparing teachers to make the formative assessment process integral to science teaching and learning. *Journal of Science Teacher Education*, 20, 475–494.

Burger, J. M., and Krueger, M. (2003). A balanced approach to high-stakes achievement testing: An analysis of the literature with policy implications. *International Electronic Journal for Leadership in Learning*, 7(4).

Carpenter, C. (2005). Recent developments in law school curricula: What bar examiners may want to know. *Bar Examiner*, 81(2).

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe, and A. Curtis (Eds.), *Washback in Language Testing* (147-170). Mahwah, NJ: Lawrence Erlbaum Associates.

Christman, J., Neild, R., Bulkley, K., Blanc, S., Liu, R. Mitchell, C., and Travers, E. (2009). *Making the Most of Interim Assessment Data: Lessons from Philadelphia*. Philadelphia, PA: Research for Action.

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives*, 10(2).

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., and Li, J. (2003). *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from Interviews with Educators in Low-, Medium-, and High-Stakes States.* Boston, MA: National Board on Educational Testing and Public Policy.

Coburn, C. E., and Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective, 9*(4), 173-206.

Commission on Instructionally Supportive Assessment (2001). *Building Tests to Support Instruction and Accountability: A Guide for Policymakers.* Washington, DC: American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, and National Middle School Association.

Darling-Hammond, L., and Adamson, F. (2010). *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., and Ducommun, C. E. (2010). *Performance Counts: Assessment Systems That Support High-Quality Learning.* Washington, DC: Council of Chief State School Officers.

Darling-Hammond, L., and Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education*, 104(2), 289-319.

Darrow-Kleinhaus, S. (2004). A response to the Society of American Teachers statement on the bar exam. *Journal of Legal Education*, 54, 442.

Dekker, T., and Feijs, E. (2005). Scaling up strategies for change: Change in formative assessment practices. *Assessment in Education: Principles, Policy & Practice*, 12(3), 237–254.

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education,* 80(4), 285–313.

Ehren, M. C. M., and Star, J. (2013). *Strategies Teachers Use to Coach Students on the Math State Test*. Paper presented at the 94th annual meeting of the American Educational Research Association (AERA), San Francisco, April 26.

Falk, B., and Ort, S. (1997). *Sitting Down to Score: Teacher Learning Through Assessment*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago, March 24–28.

Falk, B., Ort, S., and Moirs, K. (2007). Keeping the focus on the child: Supporting and reporting on teaching and learning with a classroom-based performance assessment system. *Educational Assessment*, 12(1), 47–75.

Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng, Y. Watanabe, and A. Curtis (Eds.), *Washback in Language Testing* (191–210). Mahwah, NJ: Lawrence Erlbaum Associates.

Fickel, L. M. (2006). Paradox of practice: Expanding and contracting curriculum in a high-stakes climate. In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (75–103). Greenwich, CT: Information Age Publishing.

Firestone, W. A., Mayrowetz, D., and Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, 20(2), 95–113.

Frohbieter, G., Greenwald, E., Stecher, B., and Schwartz, H. (2011). *Knowing and Doing: What Teachers Learn from Formative Assessment and How They Use the Information*. Los Angeles: University of California, Los Angeles (UCLA), Center for Research on Evaluation, Standards, and Student Testing (CRESST), CRESST Report 802.

Fuchs, L., Fuchs, D., Karns, K., Hamlett, C., and Katzaroff, M. (1999). Mathematics performance assessment in the classroom: Effects on teacher planning and student problem solving. *American Educational Research Journal*, 36(3), 609–646.

Furtak, E. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, 49(9), 1181–1210.

Gallagher, T., and Smith, T. (2000). *Main report: The Effects of the Selective System of Secondary Education of Northern Ireland*. Belfast: Queen's University.

Goertz, M. E., Oláh, L., and Riggan, M. (2009). *From Testing to Teaching: The Use of Interim Assessments in Classroom Instruction*. Philadelphia, PA: Consortium for Policy Research in Education (CPRE), Research Report #RR-65.

Goldberg, G. (2012). Judgment-based scoring by teachers as professional development: Distinguishing promises from proof. *Educational Measurement: Issues and Practice*, 31(3), 38–47.

Goldberg, G., and Rosewell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on the performance based instruction and classroom practice. *Educational Assessment*, 6, 257–290.

Government Accountability Office (2009). *Student Achievement: Schools Use Multiple Strategies to Help Students Meet Academic Standards, Especially Schools with Higher Proportions of low-Income and Minority Students.* Washington, DC, GAO-10-18.

Grant, S. G. (2001). An uncertain lever: Exploring the influence on state-level testing on teaching social studies. *Teachers College Record*, 103(3), 398–426.

Guskey, T. R. (1994). What you assess may not be what you get. *Educational Leadership*, 51(6), 51–54.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education*, 27, 25–68.

Hamilton, L. S. (2011). Testing what has been taught: Helpful, high-quality assessments start with a strong curriculum. *American Educator*, 34(4), 47–52.

Hamilton, L., Stecher, B., Marsh, J., McCombs, J., Robyn, A., Russel, J., Naftel, S., and Barney, H. (2007). *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States.* Santa Monica, CA: RAND Corporation, MG-589-NSF. As of August 15, 2013: www.rand.org/pubs/monographs/MG589.html

Hamilton, L. S., Stecher, B. M., Russell, J. L., Marsh, J. A., and Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. In B. Fuller, M. K. Henne, and E. Hannum (Eds.), *Strong States, Weak Schools: The Benefits and Dilemmas of*

Centralized Accountability (Research in the Sociology of Education, Vol. 16, 31–66). United Kingdom: Emerald Group Publishing Limited

Hamilton, L. S., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., and Wayman, J. (2009). *Using Student Achievement Data to Support Instructional Decision Making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hamilton, L., Stecher, B., and Klein, S., (Eds.) (2002). *Making Sense of Test-Based Accountability in Education.* Santa Monica, CA: RAND Corporation, MR-1554-EDU. As of August 15, 2013: http://www.rand.org/pubs/monograph_reports/MR1554.html

Hamilton, L. S., Stecher, B. M., and Yuan, K. (2012). Standards-based accountability in the United States: Lessons learned and future directions. *Education Inquiry, 3*(2), 149–170.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching. In J. Cummins and C. Davison (Eds.), *International Handbook of English Language Teaching*, 15, 487–504.

Hannaway, J. (2007). Unbounding rationality: politics and policy in a data rich system. Mistisfer Lecture, the University Council of Education Administration (November 17).

Hannaway, J., and Hamilton, L. S. (2008). *Accountability Policies: Implications for School and Classroom Practices.* Washington, DC: Urban Institute

Harlen, W., and Crick, D. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review). *Research Evidence in Education Library, Issue 1*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harrison, C. (2005). Teachers developing assessment for learning: Mapping teacher change. *Teacher Development*, 9(2), 255–263.

Hayes, B., and Read, J. (2004). Test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, and A. Curtis (Eds.), *Washback in Language Testing* (97–111). Mahwah, NJ: Lawrence Erlbaum Associates.

Heritage, M., Kim, J., Vendlinski, T., and Herman, J. (2009). From evidence to action: A seamless process in formative assessment. *Educational Measurement: Issues and Practice*, 28(3), 24-31.

Herman, J. L., and Golan, S. (1991). *Effects of Standardized Testing on Teachers and Learning—Another Look*. Los Angeles: University of California, Los Angeles (UCLA), Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., and Linn, R. L. (2013). *On the Road to Assessing Deeper Learning The Status of Smarter Balanced and PARCC Assessment Consortia*. Los Angeles: University of California, Los Angeles (UCLA), Center for Research on Evaluation, Standards, and Student Testing (CRESST).

House of Commons, Children, Schools and Families Committee (United Kingdom) (2008). *Testing and Assessment: Third Report of Session 2007–08,* Volume 1, HC 169-1.

Howarth, J. (1996). Teaching in the shadow of the bar. *University of San Francisco Law Review*, 31, 927.

Jacob, B. (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.

Johnston, J., and McClune, B. (2000). *Selection Project SEL 5.1: Pupil Motivation and Attitudes: Self-Esteem, Locus of Control, Learning Disposition and the Impact of Selection on Teaching and Learning*. Belfast: Queen's University. As of August 16, 2013: http://www.deni.gov.uk/22-ppa_gallagherandsmith_selproj5-1_pupilsmotivationandattitudes.pdf

Jones, M., Jones, G., Brett, D., Hardin, B., Chapman, L., Yarbrough, T., and Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan*, 81(3), 199–203.

Jones, M., and Moreland, J. (2005). The importance of pedagogical content knowledge in assessment for learning practices: A case-study of a whole-school approach. *The Curriculum Journal,* 16, 193–206.

Kitchen, R., Cherrington, A., Gates, J., Hitchings, J., Majka, M., Merk, M., and Trubow, G. (2002). Supporting reform through performance assessment. *Mathematics Teaching in the Middle School*, 8(1), 24–30.

Knapp, M. S., Swinnerton, J. A., Copland, M. A., and Monpas-Huber, J. (2006). *Data Informed Leadership in Education.* Seattle: Center for the Study of Teaching and Learning.

Kober, N. (2002). Teaching to the test: The good, the bad, and who's responsible. *Test Talk for Leaders*, 1. Washington, DC: Center on Education Policy (CEP).

Koenig, J. A. (2011). *Assessing 21st Century Skills: Summary of a workshop*. Washington, DC: National Academies Press.

Koretz, D. (2000). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, 37(4), 752–777.

Koretz, D. (2005). *Alignment, High Stakes, and the Inflation of Test Scores*. Los Angeles: University of California, Los Angeles (UCLA), Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D., Barron, S., Mitchell, K., and Stecher, B. (1996). *The Perceived Effects of the Kentucky Instruction Results Information System (KIRIS)*. Santa Monica: RAND Corporation, MR-792-PCT/FF. As of August 15, 2013: www.rand.org/pubs/monograph_reports/MR792.html

Koretz, D., and Hamilton, L. (2003). *Teachers' Responses to High-Stakes Testing and the Validity of Gains: A pilot stuDy* (CSE Tech. Rep. 610). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., and Hamilton, L. S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., 531–578). Westport, CT: American Council on Education/Praeger.

Koretz, D., Mitchell, K., Barron, S., and Keith, S. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice,* 13(30), 5–16.

Koretz, D., Mitchell, K., Barron, S., and Keith, S. (1996). *The Perceived Effects of the Maryland School Performance Assessment Program*. CST Technical Report No. 409. Los Angeles: Center for the Study of Evaluation, University of California.

Ladd, H., and Zelli, F. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, 38(4), 494–529.

Lane, S., Parke, C. S., and 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.

Lanting, A. (2001). *An Empirical Study of a Districtwide K–2 Performance Assessment Program: Teacher Practices, Information Gained, and Use of Assessment Results*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Seattle, April 10-14.

Lee, H., Feldman, A., and Beatty, I. (2012). Factors that affect science and mathematics teachers' initial implementation of technology-enhanced formative assessment using a classroom response system. *Journal of Science, Education, and Technology*, 21(5), 523–539.

Levinson, C. Y. (2000). Student assessment in eight countries. *Educational Leadership*, 57(5), 59-61.

Looney, I. (2009). *Assessment and Innovation in Education*. OECD Education Working Paper No. 24, EDU/WKP(2009)3.

Mabry, L., Poole, J., Redmond, L., and Schultz, A. (2003). Local impact of state testing in southwest Washington. *Education Policy Analysis Archives*, 11(22).

Marsh, J., Pane, J., and Hamilton, L. (2006). *Making Sense of Data-Driven Decision Making in Education: Evidence from Recent RAND Research.* Santa Monica, CA: RAND Corporation, OP-170-EDU. As of August 15, 2013: http://www.rand.org/pubs/occasional_papers/OP170.html

Marshal, B., and Drummond, M. J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, 21(2), 133–149.

McNess, E., Triggs, P., Broadfoot, P. Osborn, M., and Pollard, A. (2001). The changing nature of assessment in English primary classrooms: Findings from the PACE project 1989–1997. *Education 3-13*, 29(3), 9-16.

National Research Council. (2001). *Investigating the Influence of Standards: A Framework for Research in Mathematics, Science, and Technology Education*. Washington, DC: The National Academies Press.

Nichols, S. L., and Berliner, D. C. (2005). *The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing.* Tempe, AZ: Arizona State University Education Policy Studies Laboratory, Education Policy Research Unit (EPRU).

Niemi, D., Baker, E. L., and Sylvester, R. M. (2007). Scaling up, scaling down: Seven years of performance assessments development in the nation's second largest school district. *Educational Assessment*, 12(3–4), 195–214.

Oláh, L., Lawrence, N., and Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, 85(2), 226–245.

Parke, C. S., and Lane, S. (2007). Students' perceptions of a Maryland state performance assessment. *The Elementary School Journal*, 107(3), 305–324.

Partnership for Assessment of Readiness for College and Careers (No date). PARCC Assessment Design. As of August 16, 2013: http://www.parcconline.org/parcc-assessment-design.

Pecheone, R., and Chung, R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education*, 57(1), 22–36.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., and Miao, J. (2003). *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers*. Boston: Boston College, National Board on Education Testing and Public Policy.

Perie, M., Marion, S., and Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5–13.

Qi, L. (2004). Has a high-stakes test produced the intended changes? In L. Cheng, Y. Watanabe, and A. Curtis (Eds.), *Washback in Language Testing* (171–190). Mahwah, NJ: Lawrence Erlbaum Associates.

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., and Zabala, D. (2006). *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act.* Washington, DC: Center on Education Policy (CEP).

Resnick, L. B., and Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for education reform. In B. Gifford and M. O'Conner (Eds.) *Changing Assessments: Evaluation in Education and Human Services*, 30, 37–75, Springer.

Rothman, R., Slattery, J., Vranek, J., and Resnick, L. (2002). *Benchmarking and Alignment of Standards and Testing* (CSE Tech. Rep. 556). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Runte, R. (1998). The impact of centralized examinations on teacher professionalism. *Canadian Journal of Education*, 23(2), 166–181.

Saavedra, A. R., and Opfer, D. (2012). *Teaching and Learning 21st Century Skills: Lessons from the Learning Sciences*. New York: Asia Society.

Salinas, C. (2006). Teaching in a high-stakes testing setting: What becomes of teacher knowledge? In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (177–193). Greenwich, CT: Information Age Publishing.

Segall, A. (2006). Teaching in the age of accountability: Measuring history of measuring up to it? In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (105–132). Greenwich, CT: Information Age Publishing.

Shepard, L., and Dougherty, K. (1991). *Effects of High-Stakes Testing on Instruction.* Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago.

Shepard, L., Davidson, K., and Bowman, R. (2011). *How Middle School Mathematics Teachers Use Interim and Benchmark Assessment daata.* Los Angeles: University of California, Los Angeles (UCLA), Center for Research on Evaluation, Standards, and Student Testing (CRESST), CRESST Report 807.

Smarter Balanced Assessment Consortium (No date). *Computer Adaptive Testing.* As of August 16, 2013:

http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Smarter-Balanced-CAT.pdf

Smith, M. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5), 8–11.

Smith, A. M. (2006). Negotiating control and protecting the private: History teachers and the Virginia standards of learning. In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (221–247). Greenwich, CT: Information Age Publishing.

Smith, M. S., and 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* (233–67). New York, NY: The Falmer Press

Society of American Law Teachers (SALT) (2002). Society of American Law Teachers statement on the bar exam. *Journal of Legal Education*, 52(3), 446–452.

Stasz, C., Bodilly, S., Remes, S., Oyadomari-Chun, T., McCaffrey, D., Kaganoff, T., and Barnes-Proby, D. (2004). *Efforts to Improve the Quality of Vocational Education in Secondary Schools: Impact of Federal and State Policies*. Santa Monica, CA: RAND Corporation, MR-1655-USDE. As of August 15, 2013: www.rand.org/pubs/monograph_reports/MR1655.html

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. In L. Hamilton, B. Stecher, and S. Klein (Eds.), *Making Sense of Test-Based Accountability in Education* (79–100). Santa Monica, CA: RAND Corporation, MR-1554-EDU. As of August 15, 2013: http://www.rand.org/pubs/monograph_reports/MR1554.html

Stecher, B. (2010). *Performance Assessment in an Era of Standards-Based Educational Accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stecher, B., Barron, S., Kaganoff, T., and 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. 482). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B., Epstein, S., Hamilton, S., Marsh, J., Robyn, A., McCombs, J., Russel, J., Naftel, S. (2008). *Pain and Gain: Implementing No Child Left Behind in Three States, 2004–2006*. Santa Monica, CA: RAND Corporation, MG-784-NSF. As of August 15, 2013: http://www.rand.org/pubs/monographs/MG784.html

Stecher, B., and Mitchell, K. (1995). *Portfolio Driven Reform: Vermont Teachers' Understanding of Mathematical Problem Solving* (CSE Tech. Rep. 55). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Swanson, C., and Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, 24(1), 1–27.

Taylor, G., Shepard, L., Kinner, F., and Rosenthal, J. (2003). *A Survey of Teachers' Perspectives on high-Stakes Testing in Colorado: What Gets Taught, What Gets Lost* (CSE Tech. Rep. 588). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Trujillo, L. (2007). The relationship between law school and the bar exam: A look at assessment and student success. *University of Colorado Law Review*, 78, 69–114.

van Hover, S. D. (2006). Teaching history in the old dominion: The impact of Virginia's accountability reform on seven secondary beginning history teachers. In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (195–219). Greenwich, CT: Information Age Publishing.

Vitali, G. (1993). *Factors Influencing Teachers' Assessment and Instructional Practices in an Assessment-Driven Educational Reform*. Doctoral dissertation, University of Kentucky.

Vogler, K. (2006). The impact of a high school graduation examination on Mississippi social studies teachers' instructional practices. In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (273–302). Greenwich, CT: Information Age Publishing.

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teacher's instructional practices. *Education*, 123(1), 39–56.

Watanabe, M. (2007). Displaced teacher and state priorities in a high-stakes accountability context. *Education Policy*, 21(2), 311–368.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, and A. Curtis (Eds.), *Washback in Language Testing* (129–146). Mahwah, NJ: Lawrence Erlbaum Associates.

Webb, N. L., Alt, M., Ely, R., and Vesperman, B. (2005). Web alignment tool (WAT): Training manual 1.1. As of August 16, 2013: http://wat.wceruw.org/Training Manual 2.1 Draft 091205.doc

Wiggins, A., and Tymms, P. (2002). Dysfunctional effects of league tables: A comparison between English and Scottish primary schools. *Public Money & Management*, January–March, 43–48.

William and Flora Hewlett Foundation (April, 2013). Deeper Learning. Web page. As of August 16, 2013:
http://www.hewlett.org/deeperlearning

William, D. (2001). *Level Best? Levels of Attainment in National Curriculum Assessment*. London: Association of Teachers and Lecturers.

Wright, W. E. (2002). The effects of high stakes testing in an inner-city elementary school: The curriculum, the teachers, and the English language learners. *Current Issues in Education*, 5(5).

Yeager, E. A., and Pinder, M. (2006). "Does anybody really understand this test?" Florida high school social studies teachers' efforts to make sense of the FCAT. In S. G. Grant (Ed.), *Measuring History: Cases of State-level Testing Across the United States* (249–272). Greenwich, CT: Information Age Publishing.

Yeh, S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives,* 13(43).

Yuan, K., and Le, V. (2012). *Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items Through the State Achievement Tests*. Santa Monica, CA: RAND Corporation, WR-967-WFHF. As of August 15, 2013:
http://www.rand.org/pubs/working_papers/WR967.html