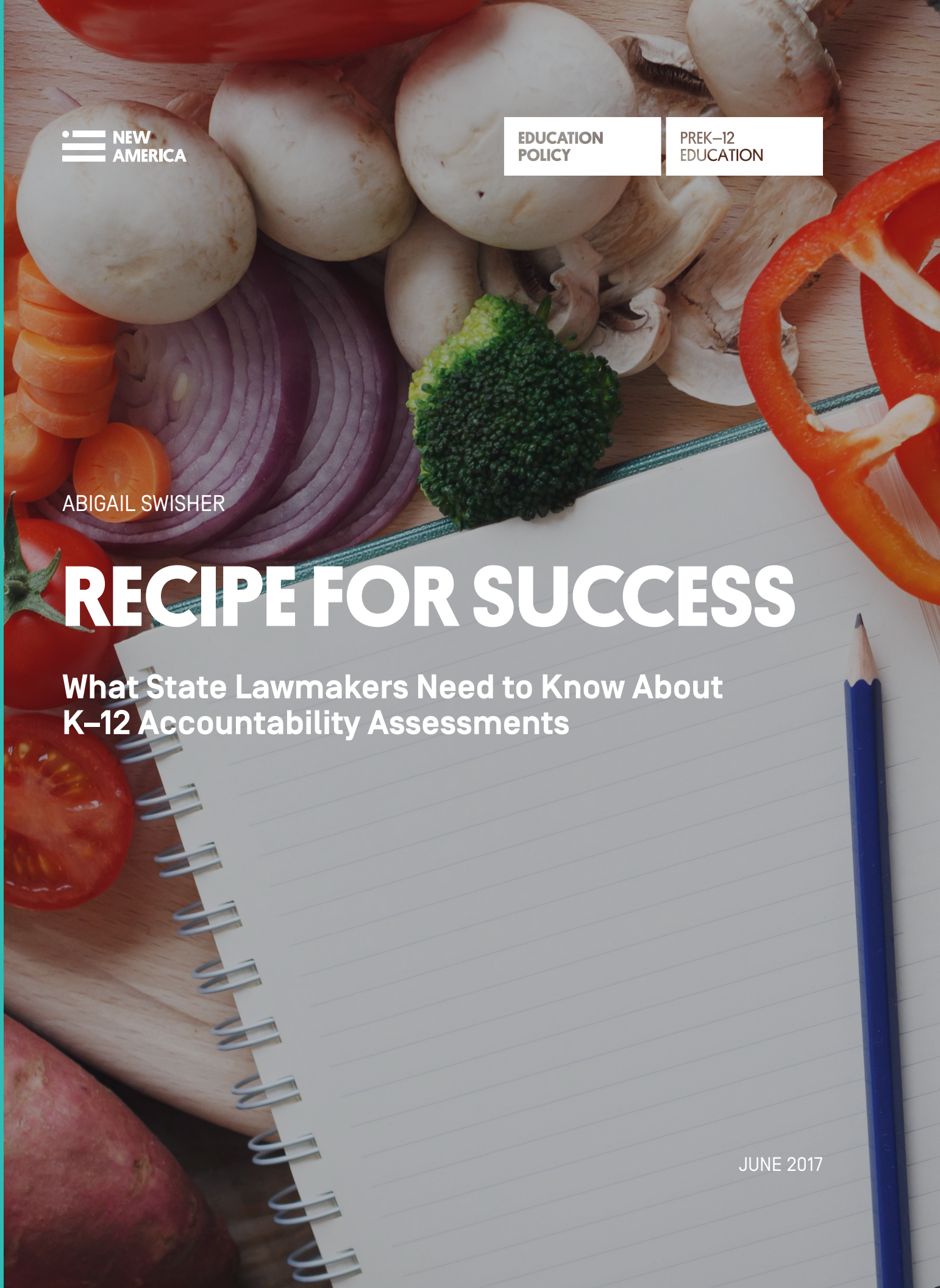

NEW AMERICA

EDUCATION POLICY | PREK–12 EDUCATION

ABIGAIL SWISHER

RECIPE FOR SUCCESS

What State Lawmakers Need to Know About K–12 Accountability Assessments

JUNE 2017

About the Author

Abigail Swisher is a program associate with the Education Policy program at New America. She is a member of the PreK-12 team, where her work focuses on college- and career-ready policies, with specific attention to both assessment and the transition from high school to higher education. Prior to joining New America, Swisher taught elementary school science and engineering in Granville County, North Carolina.

Acknowledgments

This work would not have been possible without the support of the Charles Stewart Mott Foundation. Thanks also to the Gates Foundation for its support of our 2016 Mapping College and Career Ready Policies scan, which helped shape the concept for this paper.

About New America

New America is committed to renewing American politics, prosperity, and purpose in the Digital Age. We generate big ideas, bridge the gap between technology and policy, and curate broad public conversation. We combine the best of a policy research institute, technology laboratory, public forum, media platform, and a venture capital fund for ideas. We are a distinctive community of thinkers, writers, researchers, technologists, and community activists who believe deeply in the possibility of American renewal.

Find out more at **newamerica.org/our-story**.

About the Education Policy Program

New America's Education Policy program uses original research and policy analysis to solve the nation's critical education problems, serving as a trusted source of objective analysis and innovative ideas for policymakers, educators, and the public at large. We combine a steadfast concern for low-income and historically disadvantaged people with a belief that better information about education can vastly improve both the policies that govern educational institutions and the quality of learning itself. Our work encompasses the full range of educational opportunities, from early learning to primary and secondary education, college, and the workforce.

Our work is made possible through generous grants from the Alliance for Early Success; the Buffett Early Childhood Fund; the Foundation for Child Development; the Bill & Melinda Gates Foundation; the Evelyn and Walter Haas, Jr. Fund; the Heising-Simons Foundation; the William and Flora Hewlett Foundation; the Joyce Foundation; the George Kaiser Family Foundation; the W.K. Kellogg Foundation; the Kresge Foundation; Lumina Foundation; the McKnight Foundation; the Charles Stewart Mott Foundation; the David and Lucile Packard Foundation; the Pritzker Children's Initiative; the Smith Richardson Foundation; the W. Clement and Jessie V. Stone Foundation; and the Berkshire Taconic Community Foundation.

Find out more at **newamerica.org/education-policy**.

Contents

INTRODUCTION

Assessment policy takes up a far larger share of K–12 education policy in the United States than it did even a few decades ago. Since clear and specific federal reporting and accountability requirements for states' yearly assessments began in earnest with the 2001 passage of No Child Left Behind (NCLB), controversies over these tests have only continued to grow.[1] In the years since states began programs of statewide testing in schools, state education agencies (SEAs) traditionally called the shots on what assessments are used. But as more Americans question how well K–12 assessments are serving students and schools, they are bringing those concerns to elected officials. In Indiana, dissatisfaction with the state's I-STEP assessment program became a focal point of the governor's race. In Colorado, a group of concerned parents, known to lawmakers as "The Moms," pushed their state legislators to take an active role in reforming high school tests they were unhappy with.

Legislators nationwide have taken a particularly active interest in accountability assessments, those state-selected tests that all students are required to take from grades three through twelve (see **Types and Levels of Assessment**). Arguably the most widely discussed—and bemoaned—tests students take, accountability assessments are also the most challenging assessments to select. Planning for a statewide system of accountability assessment is an inherently complex process, one that requires a deep understanding of both the science of assessment and the needs of assessment users (see **Who Are K–12 Assessment Users?**). Weighing these considerations, which are sometimes in conflict with one another, does not lend itself to hasty decision-making.

Who Are K–12 Assessment Users?

K–12 assessment users are any groups that use data from K–12 accountability assessments, though they may use that data in very different ways. This generally includes (but is not limited to) students, families, teachers, school administrators, the public, local education agencies (LEAs), state education agencies (SEAs), researchers, and the U.S. Department of Education.

Types and Levels of Assessment

Types of Assessments

The assessments used in K–12 schools can be categorized in many ways. Below is a selection that policymakers are likely to encounter in conversations on accountability assessment.

Summative assessments: tests used to measure students' cumulative learning over a set period—typically, over the span of a course or school year.

Interim assessments: sometimes referred to as benchmark assessments, used to test students' knowledge at different points throughout a course or school year.

Growth assessments: tests administered at multiple points throughout a learning period in order to measure how students' knowledge or skills grow over time.

Formative assessment practice: a process that teachers and students use to adjust teaching and learning on the spot during instruction. Often confused with interim or growth assessment, formative assessment is a practice integrated into instruction when teachers ask probing questions or observe students completing a math problem, for example. It is not a planned assessment separate from instruction like a quiz.

College entrance exams: assessments designed and traditionally used for entry into higher education. These tests measure students' general academic readiness for college-level coursework in key subject areas such as reading, writing, mathematics, and scientific thinking.

Technology-delivered assessments: tests administered using a technology other than paper and pencil. Often referred to as online assessment, many technology-based assessments do not actually operate online, but rather via computer server. Technology-delivered assessments can be designed like traditional multiple-choice and essay assessments, or use a variety of innovative approaches. One popular use of technology-delivered assessments, for instance, is making tests computer adaptive—assessing students' academic level as they test and adapting questions accordingly.

Levels of Assessment

Though assessment always occurs at the classroom level, the requirement to test can come from several levels of organization.

Classroom: Assessments chosen or created by classroom teachers can include (but are not limited to) formative assessment practices, interim and growth assessments, and other periodic assessments of content such as unit tests. Results may stay within the classroom or be shared with school leadership and students' families.

District/LEA: Assessments chosen by an LEA for classroom teachers to administer to students can include diagnostic assessments and interim and growth assessments. Results are shared with the school district, and may or may not be shared with students' families.

State & Federal Accountability: Assessments required by federal education law are selected by the state. Individual results are shared with students' families, teachers, and school leadership. Each school's results (not identifiable by student) are shared with the LEA, the SEA, the U.S. Department of Education, and the public. In addition to meeting federal requirements, states may also set additional accountability policies that use assessment data.

Planning for a statewide system of accountability assessment is an inherently complex process, one that requires a deep understanding of both the science of assessment and the needs of assessment users.

Legislators around the U.S. are encountering significant challenges as they strive to help improve assessment systems. Quick fixes to these systems can be costly and poorly aligned with the needs of users. There is an important balance to be struck between responding to constituent concerns and weighing the complex requirements and tradeoffs in selecting assessments for state and federal accountability.

Assessment might be thought of like meal preparation. While assessment is an immensely complex topic, there are central components that any state policymaker should consider before helping choose an assessment or system of assessments for accountability. First, understand the policy requirements of state accountability assessment. Next, determine specific goals and create a set of aligned "claims," and translate those claims into uses. Finally, weigh the options for purchasing the assessments needed. Legislators armed with this basic framework are more likely to make recommendations that save time and money, better meet the needs of assessment users, and cultivate buy-in from a broader group of stakeholders.

1. UNDERSTAND THE REQUIREMENTS

What are the federal requirements for annual testing? What new flexibilities are available to states under ESSA?

To begin, policymakers must have a clear understanding of the basic parameters of their state assessment system for accountability. Since the passage of NCLB, states have been required to test students in English language arts and mathematics yearly between the third and eighth grades and once in high school. Under NCLB, states were also required to assess students in science three times: once between grades 3–5, once between grades 6–8, and once in high school.

NCLB was replaced with the Every Student Succeeds Act (ESSA) in late 2015. With the new law, the requirement to test yearly in grades three through eight and once more in high school remains the same. ESSA states clearly that those assessments must be aligned with the challenging academic standards adopted by each state, and must give clear information about whether or not students are attaining those standards for their grade level. Notably, ESSA gives states considerably more freedom to determine how much test scores will weigh in school accountability, along with additional flexibility in how they design and administer those assessments. While the law remains firm that states must administer accountability assessments in the required grades and subjects, states now have the flexibility to:

- administer assessments in "chunks" (interim or benchmark assessments) throughout the school year, rather than all at once, as long as those assessments produce a single score at the end of the year;
- allow individual local education agencies (LEAs) to administer a "nationally recognized high school assessment" (such as a widely-used college entrance examination) to students in that district in lieu of the high school assessment adopted by the rest of the state; and
- use "innovative" assessment practices such as computer-adaptive technology or the incorporation of student portfolios, projects, and performance tasks.

In addition to the broad strokes of federal policy, there will be other considerations for each state in order to meet the technical specifications laid out in ESSA and its accompanying regulations,[2] factoring in any state legislation that may set additional requirements for accountability assessments.

2. DETERMINE GOALS AND TURN THEM INTO CLAIMS

What are the state's goals for its accountability assessment system, and how do they build on one another? How do those goals translate into claims?

Assessment does not occur for its own sake, but as a means to an end. That end is the same for accountability assessments across the U.S.—ensure that all students succeed in a given subject area—but may differ in the details according to state context.

Take, for instance, a state which decides that one of its goals is to ensure students show mastery of academic standards through performance tasks modeled after a pilot program in New Hampshire, which administers performance-based assessments in several districts every year.[3] This is a specific goal for improvement, and one that should impact how the state goes about selecting the most appropriate assessment to meet its goal for students. A state has lot to gain by ensuring that its goals for improvement are clearly delineated at the start of the process. Otherwise, it may choose assessments that are poorly designed to help accomplish those goals.

States must also view their goals as part of a coherent system that spans kindergarten through 12th grade.[4] Goals for student learning build on one another, and assessments must do the same. For example, one state might want to ensure that students are digitally literate enough to demonstrate their mastery of academic standards using computerized assessment tasks. Technology-delivered assessments can provide students complex opportunities to show learning, so the state

Validity and Reliability

States are required to guarantee that assessments are "valid and reliable." **Validity**, as defined by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, is "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests," or, simply, whether the claim meets use."[5] **Reliability** concerns whether an assessment will give the same or roughly similar results if given multiple times without a change to the test subject's knowledge.

selects a system of assessments for grades three through eight that are aligned with this goal. In response to public pressure to reduce testing time in the upper grades, however, the state selects a college entrance exam to test in high school.[6] The entrance exam it selects is only offered in a paper and pencil format. So, while the assessment system across grades three through eight is aligned with the goal of demonstrating digital literacy, it fails to do so for high school students.

After establishing the goals of their assessment systems, states must work to translate their goals into claims, statements that describe what an assessment measures about a student, teacher, or school.[7] For example, "this student is not yet proficient in fourth grade language arts" is a claim about a student's attainment of the knowledge and skills taught in the grade level subject area. In order to make that claim, the assessment must accurately measure the student's demonstrated knowledge of and skill in that subject. Put simply, when policymakers identify a claim, they are identifying exactly what they would like their assessment to measure. Stating this clearly is the first step to ensuring the selection of an assessment that is a valid measure of student learning (see **Validity and Reliability**).

3. TRANSLATE CLAIMS INTO USE

Does the assessment's use meet the claim the state wants to make? How many uses can one assessment reasonably support?

After it has identified a claim the assessment should be able to make, a state must identify whether or not that claim is appropriate for how the data will be used. All states, at minimum, must use assessment data to determine if schools are helping all students master challenging academic standards, but they may wish to extend the use of that data for other purposes. This complicates the process of choosing an assessment system, and may force policymakers to evaluate competing priorities. It also requires a strong understanding of the limits of each assessment's valid and reliable use.

Throughout this process, it is important to remember what claims an assessment can support, and how those claims limit what any one assessment can be used for. Take, for example, a college entrance exam designed to support claims about a student's readiness for college-level academics. The exam makes claims about readiness for college-level math, reading, and writing, respectively, along with a general readiness score. A significant number of states have recently selected college entrance exams as their high school accountability assessment. Some cite efforts to reduce testing time, and others

the perceived popularity of tests like the ACT and SAT with students and parents.[8]

One state, hoping to meet its goal of increasing the number of high school students ready to attend college directly after graduation, considers adopting such a college entrance exam for high school accountability. But this additional use of the assessment presents a number of challenges related to the central claims of the exam. First, the exam supports a claim about *readiness* (a prediction of how well students will do in subsequent work), but not *proficiency* (a measurement of how well students have mastered certain academic content). While those goals might be closely linked, they are not necessarily the same, particularly if the knowledge and skills measured to predict "readiness" differ from the academic standards students are learning.

States must also remember that each assessment can only meet so many uses, and they may have to settle for only meeting a few. Take, for example, a state which decides that one of the central uses of its assessment will be to provide detailed information on how well students have mastered different "strands" of academic standards. In addition to providing a total score for how well a student mastered eighth grade science, the data would be broken down into mastery of physical science, life science, earth science, and so on.

The biggest challenge here is that while summative assessments are often able to return accurate data to teachers on how their whole class mastered a particular academic strand, a body of educational psychometric research suggests that they are not generally designed to provide that data for individual students.[9] Because the eighth grade science assessment measures a full year of content, the exam only asks a select number of questions about each academic strand. If it asks 80 multiple-choice questions, for instance, perhaps 15 cover all of the course's physical science content. While 80 questions might be enough to give a valid and reliable idea of how well an individual student understands eighth grade science as a whole, 15 questions are not enough to say whether or not that student understands eighth grade physical science specifically.[10]

To get reliable and valid information on each student's performance on individual strands of content, therefore, students would have to take a significantly longer exam. Complicating this picture, however, is another concern heard around the state: a need to reduce time spent testing. The state must weigh its priorities: using summative assessment to provide more detailed feedback versus minimizing time spent testing.

States must remember that each assessment can only meet so many uses, and they may have to settle for only meeting a few.

In another instance, a state wants its summative assessment to give teachers feedback to improve instruction throughout the year. As states are newly permitted to use interim assessments towards accountability requirements, some have expressed interest in using this strategy to support ongoing feedback. Logistically, however, the model presents certain challenges and unknowns. As *Education Week*'s Catherine Gewertz has documented, assessment experts caution that though these tests would hypothetically provide uniform and consistent interim data to educators for instructional purposes, using them for accountability would also significantly increase time spent testing and limit educator flexibility in tailoring the sequence of the curriculum to their own classrooms.[11]

Every assessment is designed with a discrete purpose or set of purposes in mind. When policymakers either mismatch claims and uses or attempt to extend the functions of an assessment without understanding what uses it can support, they run the risk of misapplying the information generated by the assessment and potentially undermining important systems that rely on assessment data.

4. SHOPPING WISELY: WEIGH OPTIONS FOR PURCHASING ASSESSMENTS

How much can the state afford to spend to meet its goals for assessment? How much autonomy does the state want over its assessment system, and how much is it willing to pay for that level of control?

With a clear vision for their accountability system in mind, states must decide the most efficient way to procure needed assessments. Much like grocery shoppers planning a meal, states seeking new tests have essentially three options: store-bought, from scratch, or partially prepared.

Store-Bought

Many states opt to use assessments or systems of assessments that come fully created. Often called "off the shelf" or "prepackaged" assessments, these tests are shared among multiple states with common academic standards (for example, the Common Core), or states which have adopted standards very similar to those of others. Two widely used examples of prepackaged assessments are those created by the non-profit PARCC and Smarter Balanced consortiums.[12] States can also purchase prepackaged assessments through private vendors, such as ACT, which administers both the ACT college entrance exam and a series of assessments for lower grades. Much in the way that shoppers give up control of the ingredients in store-bought meals, this option generally gives states the least input into how the assessment is designed, implemented, scored, and reported to users.[13]

Not all prepackaged options will necessarily save states money—prices for prepackaged assessments vary significantly—but for some, sharing an assessment could result in significant savings. Participation in non-profit assessment consortiums, for instance, allows states to share the fixed costs of testing and invest in test enhancements that might have been too costly to build alone.[14] For example, both PARCC and Smarter Balanced have built in technology-enhanced questions that go beyond traditional multiple-choice questions, allowing students to drag and drop, write equations on screen, and so forth.

From Scratch

Many states opt for the "by us, for us" approach to assessment, where they pay national vendors

to develop tests unique to their state. Much like cooking a meal from scratch, approach gives the most control over test construction. Texas, for instance, has consistently opted to develop its own assessment system, the STAAR exams. For states like Texas that have their own academic standards, it may be the only option available, as shared assessments and test item banks are designed to measure standards used by more than one state.

As Edward Roeber, director of the Michigan Assessment Consortium, noted in an interview,[15] building a system from scratch is generally the most expensive, as everything from test questions to the technological infrastructure that underpins computerized assessments must be built from the ground up.[16]

In addition to being the costliest option, it is also the most time-consuming. According to Marianne Perie, an assessment expert with the University of Kansas, states should expect to take roughly three years from the time that they decide to change their assessment system to the time that the assessment can first be implemented for accountability purposes. These years, according to Perie, include selecting an assessment, going through the multiple layers of review and approval required to award a state contract, and then working with the vendor to build any additional technology infrastructure needed to accommodate the assessment.[17]

Partially Prepared

States which want greater flexibility and control while still retaining the benefits of a shared assessment have new options for purchasing content. Like a partially prepared meal, some pieces will come ready-made and others will have to be created from scratch. Members of both the PARCC and Smarter Balanced assessment consortiums, for instance, can purchase blocks of assessment questions and access to test item banks, which states can adapt. Michigan created a "hybrid" exam using a combination of questions from the Smarter Balanced assessments and test questions developed by in-state educators. As PARCC has expanded options for purchasing blocks of test questions, states in that consortium have also made plans to create hybrid assessments, as New America has previously reported.[18] States that do not participate in consortiums also use this approach, though some additional caution should be taken to ensure that the assessment fits the state context. Florida, for instance, leased questions for its 2015 assessments from the state of Utah. As National Public Radio's *StateImpact* project reported, this approach garnered significant criticism after an independent evaluation suggested that test questions were not fully aligned with Florida's standards.[19]

What Do Americans Think About Testing?

What, in general, do Americans think about the assessments used for state education accountability? Recent opinion polls on the subject abound, but they paint a confusing and contradictory picture.

Support for assessment in general

- A 2016 poll by *Education Next* found a generally high level of expressed support for the current federal requirements for yearly testing: 69 percent of the general public and 68 percent of parents said that they supported the requirements. Those numbers were significantly lower for teachers, 50 percent of whom expressed support for the requirement to test yearly.[20]
- In contrast, when asked in a 2015 poll from Education Post whether they believed standardized tests were fair, only 44 percent of people responded that they believe tests are fair. When asked to rate whether tests "have a positive impact on education overall," 44 percent rated tests as positive.[21]
- Though debate about the use of multi-state assessments has become heated in recent years, 63 percent of the public, 62 percent of parents, and 53 percent of teachers said they supported "using the same standardized tests in every state." [22]

Differences by demographic group

- *Education Next*'s poll found little significant difference in support by racial and ethnic group, though Hispanic respondents were most supportive. Though authors separated responses among self-identified Democrats and Republicans, there was very little difference, and a generally high level of support, for the federal requirement to test yearly.[23]
- In contrast, a review of recent polling data by Randy Bennett of the Educational Testing Service found that African American and Hispanic parents were markedly "more supportive of testing than white parents." Bennett suggests a relationship between these parents' support for assessment and their experiences with school improvement strategies that use test data.[24]

Support for assessment based on how results are used

- In Education Post's poll, parents were divided on how tests should be used. More than half agreed that tests should be used to identify students who need help, to see if "students are meeting learning standards," and to help parents identify areas to help their children. African American parents were much more likely to say that tests should be used to identify schools that need help than parents of other races.[25]
- Teachers were divided as to how state accountability assessments should be used, but the largest share (36 percent) said that they should be used "to determine if students are meeting critical benchmarks." [26]
- Of principals polled by NWEA, 37 percent said that state accountability assessments were either useful or very useful to teachers, parents, students, and administrators. A slightly lower share of parents (32 percent) polled believed that state accountability assessment results were useful or very useful to parents in general, though they were not asked to rate how useful those assessments were to them specifically.[27]

Perceptions about amount of testing in schools and test anxiety

- Three-quarters of students and one-half of parents surveyed believed that students are spending the right amount of time or too little time taking tests. By contrast, a much higher fraction of teachers, principals, and superintendents (7 out of 10) said they believe students spend too much time testing.[28]

- Education Post found that 49 percent of the public believed students spend too much time testing. Forty percent said students spend the right amount of time testing, while 8 percent said they spend too little time. African American parents were less likely than other parents to say students spend too much time testing.[29]

- Then again, Phi Delta Kappa International (PDK) and Gallup's 2015 poll found that 64 percent of the public and 67 percent of parents believed that there is too much standardized testing.[30]

- Only 29 percent of parents polled by Education Post said they believe tests "put too much stress on my child and cause anxiety." More, 43 percent, said that their child experiences some stress, but that it is manageable, and 23 percent reported tests were not stressful for their child. Of the 29 percent who cited too much stress and anxiety in their child, roughly a third said that they believe the stress comes from teachers, administrators, school staff, and the school district.[31]

Opting-out (for more information, see section to the right)

- When asked by *Education Next* whether they supported or opposed a parent's right to opt out of state assessments, 60 percent of the public and 49 percent of parents were opposed.[32]

- By contrast, PDK and Gallup's results from 2015 found a lower percentage opposed a parent's right to opt out of state assessments—only 41 percent of the public and 47 percent of parents.[33]

One major challenge to understanding what Americans think about accountability assessment is that much of the research conducted does not differentiate it from any other forms of assessment. Most of the polls reviewed above only ask questions using the terms "testing" or "standardized testing," either of which could refer to any number of tests students take at school. It is not hard to imagine that parents might have appreciably different opinions about their child's weekly spelling quiz versus their end-of-year state accountability assessment.[34]

Perhaps some measure of the difference between these poll results lies in how the researchers frame assessment issues, as was suggested with discrepancies between *Education Next* and PDK's data on support for opt-out in prior years.[35] The complexity of stakeholders' opinions about assessment should, at the very least, refute simplistic narratives about how members of the public and users view assessment.

Opt-Out

Students opting out of state assessments has been perhaps the most visible expression of dissatisfaction with assessment. Students (often with the encouragement and/or permission of their parents) will either remain home on test day or refuse to take state assessments when administered. The exact objections of different groups vary, but media coverage reviewed by Randy Bennett found that a vocal contingent reported concern about time spent testing and perceptions that assessments carry unduly high stakes for students, schools, and educators.[36]

Whatever the motivation of those who opt out, the fact that some students do so has potentially significant consequences for their schools, districts, and state. A high opt-out rate can compromise the reliability of assessments, rendering them useless for their intended purposes.

In order to ensure that data submitted to the government for accountability purposes remains reliable, federal K–12 education law requires that states, districts, and schools have 95 percent of students participate in state assessment. In the 2014–15 school year, 13 states failed to meet this requirement, according to the U.S. Department of Education.[37] Since then, ESSA has made states responsible for dispensing consequences for low participation and ensuring that participation rates improve. It remains to be seen how states will handle efforts to combat opt-out, though. *Education Week* recently reported that early ESSA plans submitted in April suggest that many plan to punish schools with under 95 percent participation rates by lowering their school rating.[38]

LOOKING FORWARD

Choosing the right assessments—those that best meet a state's goals for its K–12 education system—is an inherently challenging process. It may be tempting to approach the problem with a single consideration (reducing testing time, giving better instructional feedback, etc.) in mind, but policymakers would do well to take a panoramic view. Key ingredients cannot be left out if assessments are to help drive educational improvement. A state must select claims that meet its goals and uses that meet its claims if it wants a valid system of assessment.

Just as there are steps to a recipe that must be followed, there are steps to the process of choosing an assessment system that cannot be done out of order. Rather than, for example, exploring purchasing options (considering suggestions about adopting existing assessments or conferring with vendors, for example) before understanding assessment goals, states should be sure to establish goals and claims first.

Legislators must also weigh the most effective route to improve their state's assessment systems. While passing legislation may seem the most direct way for lawmakers to impact student assessment, it is also time- and labor-intensive. Legislation can become outdated quickly, and, in the event that it does become irrelevant, is cumbersome to repeal. Legislation may also create unnecessarily burdensome red tape for the SEA, and in some circumstances, could actively prevent the SEA from carrying out assessment as law and best practice prescribe.

In Indiana, for instance, legislators wanted to assuage concerns about the accuracy of test scoring and how it impacted school A–F quality ratings. In an effort to ensure accuracy, the legislature passed a requirement that mandated that student test scores be checked by test vendors multiple times. The year following, stakeholders across the state complained of the increased wait time for scores. State Superintendent Glenda Ritz pointed the finger at legislation, which she said placed an extra burden on the assessment vendor and ultimately slowed the return of test results to districts, schools, teachers, and families.[39]

> **Just as there are steps to a recipe that must be followed, there are steps to the process of choosing an assessment system that cannot be done out of order.**

It is difficult to envision a future, at least in the near term, in which state legislators will not feel public pressure to remain involved in state assessment decisions. For states that have continued to shift their assessment systems from year to year,

establishing equilibrium should be a priority, because year-to-year changes have significant costs. There are monetary costs to continually paying to create or contract new tests and help schools adapt to them. There are educational costs to asking educators and students to adapt to new assessments every year, and depriving the state of year-to-year data to help schools improve. Finally, there are political costs to continuous change: the state's assessment systems (and the people who make decisions about them) can be significantly undermined in the eyes of the public. Given the high cost and low satisfaction of a hastily prepared meal, lawmakers would do well to carefully consider the ingredients when helping to replace assessments.

Acknowledgments

This project would have taken a very different shape if not for the organizers and participants of both the Education Commission of the States' 2016 National Forum and the Center for Assessment's 2016 Reidy Interactive Lecture Series. At the former, assessment users—from teachers to state commissioners—shared the widespread debates in their states about how to improve K–12 tests. They voiced concerns about their own assessment systems, shared potential solutions, and challenged one another to think critically about creating systems of assessment that best met their states' needs. Several months later, at the Center for Assessment's annual lecture series, a group composed in large part of psychometricians and district and state K–12 data managers and directors gathered to discuss assessment literacy. One consistent theme mentioned throughout was the pressing need to build this literacy in a group with growing influence in choosing K–12 assessments: state legislators. The robust conversations had at these convenings, together with the quality, detailed coverage of state assessment debates provided by many education reporters, built a narrative about the unmet needs of state policymakers that was difficult to ignore.

Particular thanks to those who contributed their perspective and expertise to my research, including Susan Lyons at the Center for Assessment, Nancy Segal and Heather Weems at the Educational Testing Service, Madeline Webster and Michelle Exstrom at the National Conference of State Legislatures, Sharmila Mann and Julie Woods at the Education Commission of the States, Marianne Perie of the Achievement and Assessment Institute at the University of Kansas, Ed Roeber of the Michigan Assessment Consortium, Shaina Cavazos and Nicholas Garcia of Chalkbeat, and Representative Robert Behning, Chair of the Indiana State House Education Committee. Thanks go to New America colleagues Elena Silva, Lindsey Tepe, and Amy Laitinen for editorial guidance and insight. Sabrina Detlef copyedited and Tyler Richardett provided layout and communications support.

Notes

[1] The terms "test" and "assessment" are often used interchangeably in K–12 policy discussions, and, for the sake of consistency, in this paper as well. It is nevertheless useful to note that their technical definitions vary slightly. Assessment is the process of measuring something (whether it be knowledge, skills, etc.). A test is a type of assessment standardized to allow comparison across test-takers.

[2] For complete regulations on state assessments for federal accountability, see "Title I—Improving the Academic Achievement of the Disadvantaged—Academic Assessments; Final Regulations," *Federal Register* 81, no. 236 (December 8, 2016): 88886–88938.

[3] "Performance Assessment of Competency Education (PACE)," New Hampshire Department of Education, accessed May 31, 2017, https://www.education.nh.gov/assessment-systems/pace.htm.

[4] Assessment for accountability begins in the third grade, but it is important to remember that goals for student success set for earlier grades will affect success in later grades.

[5] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *The Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association, 2014).

[6] This particular type of misalignment has become increasingly common as states that previously participated in assessment consortiums like PARCC and Smarter Balanced (which use technology-delivered assessments) have retained those assessments in earlier grades and moved to college entrance exams like the SAT and ACT for high school. Notable recent examples include states like Delaware (a PARCC state) and Connecticut and Oregon (both Smarter Balanced states).

[7] The concept of claims, as they apply to educational testing, was explained particularly succinctly by Erika Hall in "Assessment Literacy: Test Purpose and Use" (presentation at Reidy Interactive Lecture Series, Center for Assessment, Portsmouth, NH, September 29, 2016).

[8] For more on this trend, see Abigail Swisher, "Worry Less About Your High Schoolers' Testing Time and More About Their Tests' Quality," *New America Weekly*, edition 116, March 24, 2016, https://www.newamerica.org/weekly/116/worry-less-about-your-high-schoolers-testing-time-and-more-about-their-tests-quality/; and Kate Zernicke, "Rejected by Colleges, SAT and ACT Gain High School Acceptance," *New York Times*, April 5, 2016, https://www.nytimes.com/2016/04/06/us/act-and-sat-find-a-profitable-market-as-common-core-tests.html.

[9] For more on the challenges of valid and reliable subscores, see: Sandip Sinharay, Gautam Puhan, and Shelby J. Haberman, "An NCME Instructional Module on Subscores," *Educational Measurement: Issues and Practice* 30, no. 3 (2011): 29–40; Sandip Sinharay, "How Often Do Subscores Have Added Value? Results from Operational and Simulated Data," *Journal of Educational Measurement* 47, no. 2 (2010): 150–74.

[10] The actual number of questions needed for an assessment to give accurate and reliable information is highly dependent on the information being assessed and the items themselves.

[11] Catherine Gewertz, "ESSA's Flexibility on Assessment Elicits Qualms from Testing Experts," *Education Week*, December 18, 2015, http://www.edweek.org/ew/articles/2015/12/21/essas-flexibility-on-assessment-elicits-qualms-from.html?cmp=SOC-SHR-FB. For more on the considerations involved in using interim assessments for accountability, see Nathan Dadey and Brian Gong, *Using Interim Assessments in Place of Summative Assessments? Consideration of an ESSA Option* (Washington, DC: Council of Chief State School Officers, April 2017).

[12] The model of these consortiums differs from that of traditional assessment vendors. PARCC (Partnership for Assessment of Readiness for College and Careers) and Smarter Balanced are non-profit consortiums of states, originally funded through the Race to the Top grant program. Both are governed by a board of states. Member states in both consortiums provide input and guidance for ongoing assessment improvement.

[13] For more on this, particularly in the context of pre-packaged high school assessments, see Erin O'Hara, "Choices and Trade-Offs: Key Questions for State Policymakers When Selecting High School Assessments," High Quality Assessment Project, May 2016.

[14] For more on the costs and benefits of sharing assessments, see: Matthew M. Chingos, *Strength in Numbers: State Spending on K–12 Assessment Systems* (Washington, DC: Brown Center on Education Policy at Brookings, November 2015); Barry Topol, John Olson, and Edward Roeber, *Getting to Higher-Quality Assessments: Evaluating Costs, Benefits, and Investment Strategies* (Stanford, CA: Stanford Center for Opportunity Policy in Education, 2013).

[15] Edward Roeber (assessment director, Michigan Education Consortium) telephone interview with author, December 9, 2016.

[16] Cost will vary depending on the state's specifications for the assessment and, in the case of technology-delivered assessments, according to the state's existing technology infrastructure;

[17] Marianne Perie (director, Center for Assessment and Accountability Research and Design, University of Kansas) telephone interview with author, November 29, 2016.

[18] Abigail Swisher, "New Flexibilities on PARCC, New Opportunities to Lower the Bar for Students," *EdCentral* (blog), New America, November 17, 2016, https://www.newamerica.org/education-policy/edcentral/parcc-flexibility.

[19] John O'Connor, "Test Review Raises Questions About Florida Standards Assessments Results," *State Impact Florida*, September 1, 2015, https://stateimpact.npr.org/florida/2015/09/01/test-review-raises-questions-about-florida-standards-assessments-results/.

[20] Paul E. Peterson, Michael B. Henderson, Martin R. West, and Samuel Barrows, "Ten-year Trends in Public Opinion from the EdNext Poll," *Education Next* 17, no. 1 (Winter 2017): 8–28.

[21] "2015 Parent Attitudes Survey," Education Post, 2015, http://educationpost.org/wp-content/uploads/2015/10/Ed-Post-2015-Parent-Attitudes-Survey.pdf.

[22] Paul E. Peterson, Michael B. Henderson, Martin R. West, and Samuel Barrows, "Ten-year Trends in Public Opinion from the EdNext Poll," *Education Next* 17, no. 1 (Winter 2017): 8–28.

[23] Ibid.

[24] Randy E. Bennett, *Opt Out: An Examination of Issues*, ETS Research Report 16-13 (Princeton, NJ: Educational Testing Service, April 2016); *Making Assessment Work for All Students: Multiple Measures Matter* (Washington, DC: Northwest Evaluation Association and Gallup, 2016).

[25] "2015 Parent Attitudes Survey," Education Post, 2015, http://educationpost.org/wp-content/uploads/2015/10/Ed-Post-2015-Parent-Attitudes-Survey.pdf.

[26] *Making Assessment Work for All Students: Multiple Measures Matter* (Washington, DC: Northwest Evaluation Association and Gallup, 2016).

[27] Ibid.

[28] Ibid.

[29] "2015 Parent Attitudes Survey," Education Post, 2015, http://educationpost.org/wp-content/uploads/2015/10/Ed-Post-2015-Parent-Attitudes-Survey.pdf.

[30] Joan Richardson, ed., "Testing Doesn't Measure Up for Americans," *Kappan Magazine*, September 2015.

[31] "2015 Parent Attitudes Survey," Education Post, 2015, http://educationpost.org/wp-content/uploads/2015/10/Ed-Post-2015-Parent-Attitudes-Survey.pdf.

[32] Paul E. Peterson, Michael B. Henderson, Martin R. West, and Samuel Barrows, "Ten-year Trends in Public Opinion from the EdNext Poll," *EdNext* 17, no. 1 (Winter 2017): 8–28.

[33] Joan Richardson, ed., "Testing Doesn't Measure Up for Americans," *Kappan Magazine*, September 2015.

[34] Northwest Evaluation Association (NWEA) and Gallup's yearly poll is a notable exception—their poll clearly differentiates between types of assessments and provides a detailed window into users' perceptions of each.

[35] It should be noted that a similar discrepancy between results from these two polls had been found in the organizations' 2015 polls as well, and was attributed by researchers during an interview with NPR's Anya Kamenetz to differences in question phrasing. For more, see Anya Kamenetz, "2 Polls Span 2 Poles on Testing in Schools," *National Public Radio*, August 25, 2017, http://www.npr.org/sections/ed/2015/08/25/434347544/two-polls-span-two-poles-on-testing.

[36] Randy E. Bennett, *Opt Out: An Examination of Issues*, ETS Research Report 16-13 (Princeton, NJ: Educational Testing Service, April 2016).

[37] Andrew Ujifusa, "Education Department Asks 13 States to Address Low Test-Participation Rates," *Education Week*, December 23, 2015, http://blogs.edweek.org/edweek/campaign-k-12/2015/12/twelve_states_asked_to_address.html.

[38] Andrew Ujifusa, "Here's How Some States' ESSA Plans Address Testing Opt-Outs," *Education Week*, April 5, 2017, http://blogs.edweek.org/edweek/campaign-k-12/2017/04/state_essa_plans_testing_opt_outs.html.

[39] Scott Elliot, "3 Reasons Why Indiana's ISTEP Test and School A-F Scores Come Out So Late Now," *Chalkbeat*, October 14, 2016, http://www.chalkbeat.org/posts/in/2016/10/14/3-reasons-why-indianas-istep-test-and-school-a-f-scores-come-out-so-late-now/.