4-1-2018

A Review of Standardized Testing Practices and Perceptions in Maine

Janet C. Fairman
Maine Education Policy Research Institute, University of Maine

Amy Johnson
Maine Education Policy Research Institute, University of Maine

Ian M. Mette
Maine Education Policy Research Institute, University of Maine

Garry Wickerd
Maine Education Policy Research Institute, University of Maine

Sharon LaBrie
Maine Education Policy Research Institute, University of Maine

A Review of Standardized Testing Practices and Perceptions in Maine

Janet Fairman, Ph.D.

Amy Johnson, Ph.D.

Ian Mette, Ph.D

Garry Wickerd, Ph.D.

Sharon LaBrie, M.S.

April 2018

Maine Education Policy Research Institute
McLellan House, 140 School Street
Gorham, ME 04038
207.780.5044 or 1.888.800.5044

Maine Education Policy Research Institute
5766 Shibles Hall, Room 314
Orono, ME 04469-5766
207.581.2475

Published by the Maine Education Policy Research Institute in the Center for Education Policy, Applied Research, and Evaluation (CEPARE) in the School of Education and Human Development, University of Southern Maine, and the Maine Education Policy Research Institute in the School of Education and Human Development at the University of Maine at Orono.

The Maine Education Policy Research Institute (MEPRI), is jointly funded by the Maine State Legislature and the University of Maine System. This institute was established to conduct studies on Maine education policy and the Maine public education system for the Maine Legislature.

This study was funded by the Maine State Legislature, and the University of Maine System.

Table of Contents

Executive Summary i

Introduction and Purpose 1

Context 1

Assessment Background 4

What is "Standardized" Academic Assessment? 4

Types of Academic Assessments 6

State Assessments 7

Universal Screening / Benchmark Assessments 9

Progress Monitoring 10

Diagnostic Assessments 10

District-level Common Assessments of Proficiency 11

Commonalities and Differences: eMPowerME & NWEA MAP Assessments 11

Study Methods & Limitations 14

District Assessment Coordinators 15

Teacher Survey 17

Survey Findings 19

Participation in Testing 19

State Assessments 19

Commercially-Developed Assessments 19

Use of District-Developed Tests 21

Self-Reported Test Preparation and Practice Time 22

Perceptions of Usefulness 25

State-Mandated Tests: District perceptions 25

State Assessments: Teacher Perceptions 30

Commercially-Developed Tests: District perceptions 33

Commercially-Developed Tests: Teacher perceptions 35

District-Developed Common Assessments: District perceptions 38

Context Related to Student Intervention Systems 40

Summary of Findings 42

Conclusions and Policy Implications 48

References 52

Appendix A: Federal ESSA Statutory Language 55

Appendix B. Standardized Testing Costs 65

Appendix C: Survey Instruments 67

About the Authors 73

A Review of Standardized Testing Practices and Perceptions in Maine

Executive Summary

In early 2017 the Maine legislature's Joint Standing Committee on Education and Cultural Affairs considered L.D. 573, a bill that would require the Maine Education Policy Research Institute (MEPRI) to "conduct an audit of standardized testing in a random sample of school administrative units statewide for the purpose of understanding the amount, cost and usefulness of standardized testing." The Committee declined to support the bill as proposed, and instead requested a smaller study in the spirit of the bill's intent. This report summarizes the resulting effort, which included a literature scan, document analysis, and surveys of two groups of school practitioners (testing administrators and teachers) to analyze the amount of time Maine students spend on testing, the types of tests administered, and whether the results are perceived as useful for practitioners. The scope of the study did not include viewpoints from policymakers who use test results for accountability purposes (such as superintendents, boards of education, or the Maine Department of Education).

Assessment Background: Academic assessment is any activity intended to assess **student learning**. Asking questions during class, weekly spelling quizzes, book reports, essays, graded homework, and teacher-designed unit exams are all examples of assessments. An academic assessment is said to be **standardized** when it has controlled conditions so that results can be compared across test-takers and over time. This means all test-takers would experience questions of similar type and difficulty level, would have the same supervised test conditions, and would be scored in the same way. Because all assessments have at least some random error in measuring true test-taker ability, high-stakes decisions should use **multiple sources of information** whenever possible. Standardized academic assessments vary according by several factors: who participates, the assessment format, how performance is compared, whether the assessment is used for a formative or summative purpose, and whether the assessment is intended to measure performance at a single point or growth over time.

In Maine, schools are required to participate in annual state testing in grades 3-8 and 11 (e.g., eMPowerME, or SAT). Maine also requires districts to implement a multi-tiered system of supports for students, through a framework such as Response to Intervention (RtI), which requires universal screening. The exams currently in use for state testing cannot also be used for screening, and the most common tool for universal screening does not meet specifications as a state accountability test. School districts administer commercially-developed tests for the purpose of universal screening or benchmark assessment, progress monitoring, or diagnostic purposes. District-developed assessments are typically used to measure proficiency in specific content and grade levels.

Study Methods: To conduct the study, the following research questions were identified and investigated:

- Which standardized tests are used in Maine schools?
- How much instructional time is spent on test preparation and testing?
- To what extent do practitioners find standardized testing results useful to inform instructional decisions?
- What are the pros and cons of potentially using NWEA MAP Growth as a statewide summative assessment?

To investigate these questions, two online surveys were conducted with practitioners. First, a statewide survey of school district testing coordinators was conducted in fall 2017. A total of 123 out of 179 coordinators completed surveys for a response rate of 68.7%. Designated testing coordinators may hold a variety of professional roles including: superintendent, assistant superintendent, curriculum coordinator, or other positions. Another survey with similar items was administered in January 2018 to a random sample of 2,000 Maine public school classroom teachers representing all grade levels and content areas. A total of 614 teachers responded out of 1,981 teachers with successful email addresses for a response rate of 31.0%. Both surveys used scaled items as well as open-ended items for comments. A limitation of the study is the absence of perspectives from other stakeholders in the state testing system: school boards, superintendents, the State Board of Education, and the Maine Department of Education. These groups may hold different views about the value or usefulness of the current state testing system.

Survey Findings: Key findings related to the research questions:

Use of Standardized Tests—

- All districts responding to the district survey reported administering at least one state-mandated assessment., most commonly the eMPowerME assessment (91.1%) followed by the Science assessment for 5th and 8th grades (88.6%), SAT (76.4%), and the Science assessment for 11th grade (74%).
- The overwhelming majority of school districts—over 99%—also opt to administer additional commercially-developed standardized assessments to supplement the state test results. There were only a few assessment products that had high rates of use.
- Of the commercially-developed assessments used, districts indicated most frequent use of the NWEA Reading and Mathematics tests (65.9% and 65% respectively). The teacher survey also confirmed the predominant use of the NWEA, and less frequent use of other kinds of commercially-developed tests.
- About half (50.4%) of the responding districts indicated their districts administer a district-developed writing prompt, 30.1% districts indicated use of other common assessments, and 30.1% indicated they do not administer any district-developed common assessments.

Time on Testing or Preparation—

- Districts indicated they are more likely to encourage teachers to spend class time for students to prepare or practice for the state tests than for commercially-developed or district-developed assessments.
- Teachers estimated more time spent on test preparation for state assessments in tested grades and subjects than for similar subjects in untested grades. Teachers indicated much less time was spent on test prep at the secondary grade level for both tested and untested grades than the elementary grade level. Estimates of test prep for tested subjects in grades 3-8 generally ranged from one hour to over 10 hours showing considerable variation across respondents.
- Teachers estimated that students in grades 3-8 spend the most class time taking tests, with estimates of the median time for a typical student ranging from 12-21 hours. Teachers estimated that students in lower and higher grades spend from 4-8 hours in testing.

Perceptions of Usefulness of Testing—

- Districts and classroom teachers perceived the state-mandated tests to be least useful for informing decisions at district, school and classroom levels. They perceived commercially-developed and district-developed tests to be more useful for these decisions.
- Teachers reported specifically that universal screening assessments were more useful than the state tests, and found them helpful for informing: feedback to students and parents, guiding instructional practices, informing changes in school or district curricula, identifying students who need additional support, and determining a student's proficiency. Teachers also perceived that the state tests were **not** useful for informing the community about school performance.
- Districts and teachers indicated several reasons why they perceive the state tests to be less useful at the school and classroom levels. These included: lack of timely reporting of results, lack of multiple data points to measure growth over time, and lack of fine-grained results to allow for diagnostic information. Practitioners also noted the frequent changes in the state's testing system, making comparisons over time difficult, and the inability to disaggregate data for further analysis. While practitioners indicated a desire for stability in the testing system, they also expressed the need for assessments to be more useful and timely. Some teachers expressed low levels of confidence in the accuracy and validity of standardized test results, and felt their own classroom assessment was more useful.
- Most of the responding teachers believed their schools had adequate assessment systems in place to identify students who need additional support, and that they have adequate knowledge to use test results to inform teaching. However, teachers wanted more time to discuss student progress with their colleagues by grade level or content area.

Potential for Meeting Multiple Purposes of Assessment—

- The study considered the potential for identifying a single assessment that could serve a dual role of both end-of-year and universal screening. It was found that the current eMPowerME reporting system cannot fulfill the purpose of universal screening because the information provided to schools is not specific to discrete content standards. The assessment is also not administered frequently enough for this purpose. The eMPowerME exam would need to be retooled to serve both needs.
- The study also considered the potential for using the current NWEA MAP Growth assessment to meet the purpose of federal accountability as well as universal screening. The current reporting system cannot fulfill the purpose of state accountability because students are given computer-adapted questions to discern their current level of achievement and learning growth, and are not necessarily assessed specifically against grade-level expectations. The resulting data do not meet federal accountability criteria.

Testing Costs—

- Researchers were unable to isolate spending on standandardized testing within the districts' expenditure data reports. A full audit of each district would be required to assess expenditures as well as total costs of testing, which would demand considerable time and expense. A report appendix compiles pricing information on selected standardized assessments for a rough sense of costs. The per-student costs for the state assessments in AY 2018 are about $26 per Maine student, of which $9 is from general funds and the remainder is covered through federal funds. The most commonly used commercial testing products cost school districts about $13 to $16 per student.

Conclusions: Overall, district leaders and teachers agreed in their view that the state tests are less useful for school and classroom decisions than the commercial or district tests. The inability to obtain timely results on the state tests, the lack of multiple data points to measure growth, and the inability to disaggregate data and obtain finer-grained results were key factors reducing the usefulness of the state tests. However, current federal and state requirements largely dictate the current level of testing. There are no readily-available options for substantially reducing testing time without a change in policy or new developments in available assessments.

Policy Implications: Some broad considerations for future policy include the following:

- Reducing the number of assessments that students participate in or the time they spend in testing presents considerable challenges. The current state tests meet the purpose of federal and state accountability, but cannot fulfill the purpose of universal screening. Some possible options might include:
    - Developing state tests for administration at multiple points in the year to allow for a measure of growth in learning.

- Exploring revisions to the NWEA MAP Growth reports so they can meet federal accountability purposes, which may require an increase in test length and therefore not necessarily reduce testing time.
- Developing concordance tables based on the NWEA MAP Growth assessment to estimate student-level proficiency for state-level expectations and accountability. Approval by the U.S. Dept. of Education would be needed to ensure this approach would satisfy the ESSA accountability requirements. This could eliminate the need for separate state tests. A less likely policy proposal would be to propose a statewide sampling system (similar to NAEP) to satisfy the need to evaluate grade-level proficiency at each school for accountability purposes, combined with NWEA MAP Growth results to measure individual student achievement.
- Modifying the eMPowerME Benchmark assessments to also meet the need for a final screening assessment. It is unclear if districts would be willing to adopt this in place of their current use of NWEA MAP Growth.
- Instead of combining universal screening with state accountability assessment, Maine could combine standardized end of course assessment with state accountability by developing state exams for key subjects.

- Some improvements could be made in the state testing system to increase perceptions of usefulness for schools and classroom teachers. For example, finding ways to speed up the reporting of results, providing more fine-grained results that align with content standards, and providing data in a format that districts could access and disaggregate would all improve the perceived usefulness of state test results for schools and teachers.
- While stability in the state assessment program is strongly desired, the current system is not viewed as very useful to schools and teachers. Planning for a better system in the future should begin with knowledge of what is working well and what is not working well with the current system. Policymakers may wish to pursue a goal of reducing the number of tests or time students spend in testing as well as improving the usefulness of the data resulting from state tests. To increase the likelihood that a new system would meet diverse needs and purposes, educational leaders and practitioners should be consulted in a meaningful way throughout the development process.
- Any major change in the state assessment system would require significant investment of time and funding. Some adjustments and shifts would be less costly than others.
- Further study is needed to investigate emerging accountability systems currently in development to identify new options. New Hampshire is piloting a district-level performance-based assessment system.
- Additional data could be collected on the status of district implementation of a MTSS/RtI system to obtain a more complete picture of that effort.

A Review of Standardized Testing Practices and Perspectives in Maine

Introduction and Purpose

In early 2017 the Maine legislature's Joint Standing Committee on Education and Cultural Affairs considered L.D. 573, a bill that would require the Maine Education Policy Research Institute (MEPRI) to "conduct an audit of standardized testing in a random sample of school administrative units statewide for the purpose of understanding the amount, cost and usefulness of standardized testing." After weighing the time and expense that would be required to conduct a formal audit, the Committee declined to support the bill as proposed. However, a task was added to the annual MEPRI work plan to conduct a study in the spirit of the bill's intent. This report summarizes the resulting effort, which included a literature scan, document analysis, and surveys of two groups of school practitioners (testing administrators and teachers) to analyze the amount of time Maine students spend on testing, the types of tests administered, and whether the results are perceived as useful for practitioners.

Context

National

Since the implementation of the No Child Left Behind Act (NCLB) in 2002, standardized tests have played a central role in the school accountability movement. In a quest to evaluate and improve student achievement of state standards, teacher quality, and school accountability, school districts have increased the frequency and quantity of time spent on testing student learning.

During this same time period, school districts saw a marked increase in the availability of standardized tests developed by educational business partners. Schools have adopted these commercially-developed tools because of their advertised efficiency, quality, ease of scoring, and accessible reports. They serve various purposes such as: assessing learning aligned to a specific curriculum, screening, measuring growth, progress monitoring, and diagnosing specific learning challenges.

However, the national focus on test scores produced unintended consequences. Many schools devoted increasing time and resources to test preparation and test administration, particularly in the tested content areas of reading, writing, and mathematics, at the expense of other important aspects of school curricula (Koretz, 2017). FairTest—an organization founded to coordinate efforts to reduce the use of standardized tests in schools—reported several reductions in testing across the US (Koretz, 2017):

- Since 2012 the number of states with high school graduation exit examinations has dropped from 25 to 13.
- Some states and districts have reduced the amount testing or time used for testing. New Mexico and West Virginia eliminated the number of tests for 9th and 10th grades students. Other states have eliminated some tests in some grades. In response to pressure, the Partnership for Assessment of Readiness for College and Careers (PARCC) cut the length of its exams by 90 minutes.
- Seven states have ended the use of student test scores as a means to evaluate teachers. Other states such as New Mexico have reduced the weight of test scores for teacher evaluations.
- In New Hampshire, half the school districts are replacing grade-level standardized tests with teacher-generated performance assessments.

Based on FairTest's report it appears that states and school districts across the US have been re-evaluating their reliance on standardized tests, Maine included.

The discomfort with some aspects of NCLB helped shape federal policy when its replacement, the Every Student Succeeds Act (ESSA), was passed in 2015. Under ESSA, states have been asked to develop their own accountability systems and consequences for schools identified as underperforming, rather than using one-size-fits-all federal approaches. There remains a continued emphasis at the federal level on standardized testing, including mandatory standardized tests to evaluate whether students in grades 3-8 and high school are meeting grade-level expectations of achievement. The specific criteria for federal accountability systems are important for understanding the specifications that state standardized tests must meet. These requirements are explained in more detail in the assessment section that follows.

Maine

Maine has administered statewide assessments to public school students in selected grades for over two decades. The Maine Educational Assessment (MEA) program has changed over the years, as state and federal policies and priorities have changed. Prior to the passage of the No Child Left Behind (NCLB) Act, Maine's testing program assessed student proficiency in six of the Maine Learning Results subjects in grades 4, 8, and 11. Literacy (reading and writing) and health education were tested in late fall, and mathematics, science & technology, social studies, and visual and performing arts were tested in spring. After the advent of NCLB, Maine altered its assessment program to test math and literacy in grades 3 through 8 and 11, and also science in grades 5, 8, and 11, to meet the federal mandate. This system remains in place under the state's ESSA accountability plan. The state contracts with Measured Progress, a company based in New Hampshire, to administer its eMPower assessments in mathematics and literacy in grades 3-8, and uses the SAT test from CollegeBoard as the 11th grade summative exam. Measured Progress also administers a science assessment for students in grades 5, 8, and 11.

In addition to these state-required summative assessments, a large majority of Maine school districts opt to purchase additional standardized tests to serve other purposes. These uses are described more fully in the Assessment Background section that follows; they are included in the scope of this report. Moreover, in response to a state requirement to implement proficiency-based diploma systems, many Maine districts have also begun or intensified their use of locally-developed "common assessments." These are used school- or district-wide—for all students in a given grade level—to evaluate student learning in specified knowledge areas. These assessments can also be considered to be standardized when they are administered and scored in similar ways so that results across different teachers can be combined and compared.

Perceptions of Usefulness

An important consideration for the investment of time and resources into standardized testing is its acceptability to teachers, administrators, students, and parents. The concept of acceptability in a school setting is defined by whether school staff, students, or parents regard an intervention or assessment as "appropriate, fair, and reasonable for

the problem or the client" (Kazdin, 1977, p. 493). In terms of standardized assessment, all stakeholders' judgements about an instrument should be considered. For instance, minimizing the amount of time spent testing is important to all stakeholders. The usability of the results is very important to teachers. Easily interpreted scores are important for students and parents. The ease of implementation is crucial to all school staff and students. When acceptability is not considered carefully, research indicates that low levels of acceptability are related to low levels of implementation integrity and utilization (Elliott, Witt, Kratochill, & Stoiber, 2002). Given that acceptability is tied closely with implementation integrity and utilization, it is vital that acceptability be a priority when selecting standardized assessments. The principles of acceptability guided the development of survey items on perceptions of usefulness.

Assessment Background

What is "Standardized" Academic Assessment?

Academic assessment is any activity intended to assess **student learning**. Asking questions during class, weekly spelling quizzes, book reports, essays, graded homework, and teacher-designed unit exams are all examples of assessments. Some assessments are commercially developed and sold to schools, either as part of a curriculum package or as stand-alone products, to measure certain skills or knowledge areas.

Schools also use assessments to evaluate non-academic dimensions such as student dispositions, study skills, and other attributes. Assessing student behaviors is also an increasingly common practice for identifying students who may benefit from behavioral intervention programs. However, non-academic assessments are beyond the scope of the current study.

An academic assessment is said to be **standardized** when it has controlled conditions so that results can be compared across test-takers and over time. This means all test-takers would experience questions of similar type and difficulty level, would have the same supervised test conditions, and would be scored in the same way.

By controlling conditions, it is easier for an assessment process to be **reliable**--that is, to produce consistent results. If two students with similar knowledge levels on the

content being tested achieve widely different test scores, then the testing process is not suitably reliable for comparing those students. Several different factors could produce unreliable results, such as: test versions that contain questions of varying difficulty levels (so that one test is easier than the other), significantly different testing circumstances (such as a fire alarm occurring during a testing process), or subjective scoring systems that make it possible for different test graders to give widely different scores to test answers of similar quality. Ensuring reliability of a testing process requires rigorous pilot testing, statistical analysis of test responses to ensure repeatability, and training of test scorers to ensure they produce consistent and accurate grades.

In addition to reliability, it is also important to gauge the **validity** of an assessment. Validity is an evaluation of whether a test measures what it is intended to measure. A well-designed algebra test can be a valid measure of algebra knowledge, but is not a valid assessment of calculus knowledge. There are myriad ways in which an assessment can have challenges that affect validity. The questions may not fully or adequately capture the scope of knowledge being assessed; questions may be poorly worded and lead to a pattern of incorrect answers when students truly do know the content; or the test may work well when piloted with one group of students, but not work well when used with students of different ages or language abilities.

As with reliability, establishing that an assessment is valid requires time and expense. It is rare for non-commercial standardized tests to undergo the level of scrutiny required to verify that an assessment is both reliable and valid. Before choosing to purchase a commercially-developed assessment, schools should ensure that the product has been evaluated to ensure validity and reliability.

The principle of **fairness** dictates that the higher the stakes are for the test-taker, the more stringent the expectations should be for validity and reliability. This is critically important if the results of the assessment will be used to make consequential decisions about a student's future, such as course placement, awarding course credit (passing grades), progression to the next grade, attainment of a high school diploma, college admission, placement in an intervention program, etc. Also, because all assessments have at least some random error in measuring true test-taker ability, high-stakes decisions should use **multiple sources of information** whenever possible.

Types of Academic Assessments

Standardized academic assessments vary by several key characteristics:

- Who participates: tests can be administered to a class of students all at the same time, to all students in a class tested one at a time, or only to selected students (rather than to the entire class).
- Format: computer based, paper based, orally administered, or expert observation. The format plays a key role in determining how quickly the tests can be scored and when results can be made available, and impacts costs.
- Comparison: Norm-referenced tests compare each student's performance to that of other tested students; student scores follow a bell-curve pattern and the comparison is typically to the average (middle) performance level. Criterion-referenced tests measure knowledge against certain key topics (learning standards) and comparison is to pre-established numbers of correct responses that are deemed to represent benchmark levels of knowledge.
- Summative vs. formative use of results: **Summative** assessment provides test results that reflect learning during an entire unit, course, or school year. Results are graded formally and produced after opportunities for learning have ceased (Harlen & James, 1997). Results may weigh heavily in student grades, grade level promotion or retention (including graduation), or teacher evaluation. **Formative** assessment is used to determine how much a student has learned, establish goals of student learning, and determine what students must learn to arrive at the goal (William & Thompson, 2007). The process of formative assessment involves regular feedback to the learner that is designed to facilitate meeting curricular objectives rather than producing a final high stakes evaluation of student learning. Teaching practices focus on clarifying objectives, activating a sense of ownership over the learning process, and providing knowledge of resources (Black & William, 2009).
- Growth vs. status: assessment results can be compared to benchmarks to evaluate how well students are doing against objective expectations at a given point in time (**status**), or they can be compared to prior test results to determine whether students are increasing their knowledge in certain areas (**growth**), either individually or as a group. Assessment design should be appropriate for each purpose; status measures should be aligned to the standards being assessed, while growth measures need to ask similar content on each test to determine whether student knowledge is increasing in that area.

Across the above categories, several distinct types of tests have arisen to suit specific needs. Different test purposes demand different formats, comparisons, and levels of validity and reliability to fulfill their requirements. The purposes most commonly seen in use in K-12 schools are: statewide assessments used for federal school accountability

purposes; universal screening (or "benchmark") assessments; tools for monitoring student progress in intervention programs; assessments to diagnose specific student learning challenges; and district-wide assessments of student proficiency. These test purposes are described in more detail below, including common examples of each test purpose, typical characteristics, whether they are used formatively and/or summatively, and whether they are used as growth or status measures.

State Assessments

Examples in Maine: *eMPowerME, SAT*

The federal Elementary and Secondary Education Act (ESEA), as most recently reauthorized by the Every Student Succeeds Act (ESSA) of 2015, stipulates that states must have assessment systems that meet certain criteria. Excerpted statutory language is included in Appendix A. In order to receive federal funds, states must implement accountability systems that use standardized testing results to identify schools that are underperforming by state-established expectations.

Among the requirements, there are two particular criteria that drive the design of state assessment systems. States must assess *each student* in grades 3-8 and once in high school for knowledge of *grade-level learning standards* specified by each state. First, to assess *each student* on the full gamut of math and literacy expectations and provide valid and reliable results, the tests must include multiple items for each knowledge area in the state standards. This results in lengthy testing processes for each individual. Second, the requirement to assess students against *grade-level standards* results in designating a narrow scope of content to be tested in each grade. This emphasizes a status assessment approach rather than formative tests that are optimized to measure growth. For example, all fourth-grade students are assessed only on fourth-grade expectations. A child who began the school year reading at a third-grade level and made no progress will be designated the same as one who began at a first-grade level and progressed to a third-level throughout the year. Both would score as not proficient, regardless of the individual progress made. Conversely, a child who reads at a sixth-grade level and one reading at a high school level would both score as exceeding expectations. The results are useful for providing consistent and comparable results of fourth grade achievement across all

schools, but are less useful for informing achievement for individual students well above or below grade-level expectations. This grade-level approach also presumes a strictly age-based approach for organizing learning, which may create tensions in schools pursuing personalized learning systems based on proficiency level.

Maine's current state assessments do not provide results before the end of the school year, or even before the beginning of the next school year. This eliminates the potential for the results to be used by teachers to inform instruction or assess proficiency in course-related learning standards. Data on patterns of student learning in prior years can generally inform curriculum improvements, but are not available quickly enough to provide feedback on whether a program or intervention is working as intended.

To meet the federal criteria, Maine's state assessments are primarily designed to report on status, not growth. However, because Maine stakeholders also value feedback on whether schools are contributing annually toward student learning, the state accountability system also uses information about changes in student proficiency levels from one testing year to the next to calculate a growth measure from the status assessment data. Because the tests only measure a narrow band of content specific to each grade span, they are not well-suited to assess a full range of learning growth. As a result, school-level accountability ratings based on status are moderately to highly correlated with the growth metric (Johnson & Fairman, 2017). This diminishes the ability to evaluate patterns of growth in student learning.

By contrast, a system for school-level accountability could theoretically be designed to use a sampling process to determine average student performance, whereby each student is tested on only selected standards, but the full range of standards would be assessed once data from all students is combined. (This is the basic method used at the federal level to measure and compare state performance in the National Assessment of Educational Progress (NAEP) program.) Within some margin of error, this could produce similar depictions of school performance with less testing time per student. However, it would not result in meaningful information about each individual student's learning across all of the grade-level standards. By requiring this level of individual assessment, the federal specifications thus dictate a substantial threshold minimum of testing time.

Universal Screening / Benchmark Assessments

Common Examples: *NWEA MAP Growth, STAR, AIMSweb, DIBELS, Teacher's College Reading Assessments, PSAT, ACCUPLACER*

Maine requires that all school districts implement a multi-tiered system of supports (MTSS) for students, per MRSA Title 20-A, Sect. 4710. In a MTSS framework, such as the Response to Intervention (RTI) system commonly used in Maine, schools conduct universal screening in math and literacy (reading and writing) three times per year. Universal screening instruments consist of items designed to assess specific academic skills, such as phonemic awareness, as well as general outcomes comprised of multiple skills, such as oral reading fluency. The goal of universal screening (sometimes referred to as benchmark assessment) is to detect which students may have skill deficits compared to their grade-level peers so that they may receive appropriate academic support (Jenkins, Hudson, & Johnson, 2007).

In order for screening assessments to serve their intended formative purpose, the results must be available soon after administration and be provided in a format that teachers can readily use to make instructional decisions. Sufficient detail should be provided about each individual student as well as patterns of overall class performance in order to guide instruction. If results are not quickly available, or if teachers lack sufficient data tools or expertise to make use of the results, then the opportunity for formative assessment is lost.

Screening assessments come in a variety of formats. Most are computer-based with preliminary results available almost immediately and final reports available within days or weeks, after verification. Paper-based versions may be machine-scored or scorable by the teacher for quick results. They are shorter in duration than state assessments (typically 30-45 minutes or less per subject). Some tests emphasize status results, with achievement rated against peer group performance. Others, including NWEA, are designed to emphasize growth in learning over time.

Progress Monitoring

Common examples: *AIMSweb, NWEA MAP Skills, Teacher's College Reading & Writing Assessments*

Once a student has been identified as at-risk through universal screening, school personnel apply additional evidence-based interventions in an effort to correct skill deficits. During this process, at-risk students are assessed weekly or bi-monthly to determine whether they are responding to the interventions (Fuchs & Fuchs, 2006). To be effective, progress monitoring assessments must have comparable alternate forms that allow for multiple administrations of the same skills or general outcomes without improved performance resulting from exposure to previous testing (Fuchs, Compton, Fuchs et al., 2008). Progress monitoring is formative by its nature—it is used to inform instructional "next steps" and not to determine grades or academic placements. Progress monitoring assessments are typically quite short – 10 minutes or less – and are often scored by teachers or educational specialists for quick results.

Diagnostic Assessments

Common examples: *Woodcock, Fountas & Pinnell, DIBELS*

Diagnostic assessments are most often utilized to deeply analyze specific skill difficulties detected through universal screening. For example, if a student screened at-risk for phonemic awareness deficits, a follow-up diagnostic assessment would be administered to determine the specific phonemes the student needed to learn. Unlike universal screening and progress monitoring assessments, diagnostic assessments are generally administered once, require more time and skill, and directly inform skill interventions (Johnson, Pool, & Carter, n.d.).

Diagnostic assessments are typically administered by educational specialists and not by classroom teachers (unless they have received specific training). These tests do not neatly fall into either the formative or summative category; they can be considered formative since their results guide next steps in student instructional options, but could also be considered summative since their results may lead to meaningful changes in a child's educational program.

District-level Common Assessments of Proficiency

In order to inform instruction, schools need information on how well students are learning the information in the curriculum. This is even more of a priority in Maine's current state context of proficiency-based diploma systems. Because instruction and assessment practices vary from teacher to teacher, it has become increasingly common for school districts to use common assessments to determine student proficiency in certain knowledge or skills that the district has identified as critically important. Schools sometimes use commercially-produced assessments for this purpose, such as those included in a packaged curriculum. In other cases, schools or districts develop their own common assessments that are administered to all students at a specified point in their studies. Some of these common assessments can be said to be standardized because they are administered under the same conditions for all students and graded using the same criteria – sometimes even by teams of teachers to ensure consistent results. However, such district-developed assessments are rarely studied for reliability, and thus cannot be sanctioned as valid from a research perspective. Nonetheless, they are typically used for summative purposes.

Commonalities and Differences: eMPowerME & NWEA MAP Assessments

In an effort to increase the efficiency of standardized testing and reduce the time students spend taking assessments, the question is often raised: "Can we consolidate tests and use them for more than one purpose?" To address this question, this section of the report analyzes whether and how well the state exams (eMPowerME and SAT) and the commercially-developed exam most commonly used by individual districts, the NWEA MAP Growth, could serve as tools for both federal accountability and universal screening.

eMPowerME

The eMPowerME assessment measures student progress according to Maine grade-level standards in grades 3-8 for math and English language arts. It produces norm-based cut scores as a means to determine student proficiency (2016-2017 eMPowerME ELA/Literacy and Mathematics Technical Report, n.d.). This construction makes it suitable for meeting the mandatory federal reporting requirements specified in ESSA (Appendix A).

It is theoretically possible that eMPowerME could be retooled to serve a dual purpose as a year-end universal screening instrument, because it assesses individual student proficiency in specific learning areas. However, the current reporting system cannot fulfill this purpose. Individual student reports are not specific enough to discrete standards, nor available with sufficient frequency (three times per year), to use with the RtI framework. Moreover, since eMPowerME is administered toward the end of the school year and results are not provided until the late fall of the subsequent school year, it is not responsive enough to serve as screening instrument for the school year in which it is administered. By the same token it cannot monitor student progress for students participating in intervention programs. Measured Progress offers a companion assessment product, including a test bank of items that can be used for formative assessment at the classroom level and its *Benchmarks* assessments, that meet apparent criteria for use in universal screening and progress monitoring in an RtI system. This would be a separate and additional assessment processes from the eMPowerME year-end summative test.

In addition, school districts seek feedback on student growth as well as information on student status. As described above, eMPowerME is primarily a status measure and only assesses grade-level content. Students who are working on learning content that is either behind or ahead of their grade-level peers are not within the scope of the tested material. Based on website descriptions, the companion *Benchmarks* products appear to also focus only on grade-level content and do not emphasize growth. Maine has developed a methodology for calculating student growth from the annual test data, but it is not optimal. Maine school districts did not report using these additional tools. Additional investigation of eMPowerME and its companion products would be required to assess its potential to yield a robust measure of growth in student learning.

NWEA MAP Growth

The Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) Growth assessment is marketed to offer a universal screening tool appropriate for use with RtI and MTSS systems. (Progress monitoring and diagnostic assessment are also available using the companion MAP Skills tools). According to NWEA, MAP Growth can be

administered up to three times per school year as a universal screening tool for grades K-12; available test subjects measure proficiency in math, reading, language usage, and/or science. It is a computer adaptive test, so that the difficulty level of questions is geared to individual student performance. This allows assessment of a student's knowledge level along the entire K-12 spectrum, rather than looking more narrowly at performance within one assigned grade level. Special normative scores, called RIT Scores, allow teachers to view individual student progress continuously across grade levels. These are tracked over time to depict student learning in between testing intervals.

In the past, there have been questions raised about the reliability and validity of the NWEA MAP assessment. Critics have claimed that the tests were not designed to assess student progress in the way states envision. This criticism appears to center on using test data to evaluate teacher performance, which is not the intended purpose of the assessment (Shaw, 2013). NWEA has produced responses to several of the criticisms and they have produced technical reports describing their sampling and norming processes (Thum & Hauser, 2015; NWEA, 2013). Based on these explanations and their technical reports, there appears to be little evidence to support reliability and validity concerns with the assessment when used for its intended purpose. However, additional analysis would be appropriate if the test were to be used in a new way for accountability purposes.

NWEA claims that MAP is aligned to all state standards and Common Core State Standards (NWEA MAP Suite, 2018). Currently, it does not produce reports that align to Maine's grade-level standards. This makes the test unsuitable as a state accountability measure, because it does not meet the federal criteria for assessing students against grade-level expectations (see Appendix A).

Summary

Maine school districts are required to participate in annual state testing in grades 3-8 and 11. They also have a de facto requirement to administer universal screening assessments because they are expected to implement MTSS support systems, which are based on a foundation of testing. The exams currently in use for state testing cannot also be used for screening, and the most common tool for universal screening does not meet specifications as a state accountability test.

While this analysis specifically investigated the NWEA MAP Growth exam, the same conclusions can me made about other commercially-developed universal screening assessments commonly used in Maine. The STAR exams from Renaissance Learning, AIMSweb, and ACCUPLACER are also not suitable to meet federal accountability requirements as currently designed.

Other than the eMPowerME test suite including the additional *Benchmarks* exams, the other test suite that has the potential to be suitable for both state assessment and universal screening is the Smarter Balanced Assessment Consortium (SBAC) system. The SBAC test suite includes an annual summative end-of-year test as well as a system of formative assessments that appear to meet criteria for use in universal screening. Maine discontinued use of the Smarter Balanced summative tests in 2015 as a result of legislative action (see Title 20-A MRSA, §6211); this policy change also prohibits use of any test developed by a consortium to assess common core state standards, which disqualifies both the Smarter Balance and the PARCC consortium tests from statewide use. However, even if Maine were to reconsider using the SBAC test suite, or to expand eMPowerME to provide the Benchmarks screening tests, it is unlikely to result in a reduction in testing time over the current practice of the combined use of NWEA and eMPowerME. Substituting eMPower Benchmarks or SBAC interim assessments for NWEA MAP Growth as a universal screening tool would not substantially change total testing time. As long as schools need to administer assessments for both state accountability and universal screening, the only way to materially reduce testing time would be to develop or identify a system where the assessments can serve both purposes.

Study Methods & Limitations

To conduct the study, the following research questions were identified and investigated:

- Which standardized tests are used in Maine schools?
- How much instructional time is spent on test preparation and testing?
- To what extent do practitioners find standardized testing results useful to inform instructional decisions?
- What are the pros and cons of potentially using NWEA MAP Growth as a statewide summative assessment?

An additional research question, "What do districts spend on standardized testing?", was initially included in the research plan but found to be unfeasible due to the lack of specificity in expenditure data reporting. Contextual information regarding standard testing pricing is included in Appendix B.

The scope of the study was limited to practitioners (district-level testing coordinators and classroom teachers). District school boards, superintendents, the State Board of Education, and the State Department of Education were not included. These stakeholders also use the results of standardized assessments and have different needs than practitioners. Thus, the study should not be construed as a complete representation of the value of state summative assessments.

To learn which standardized tests Maine school districts currently conduct, the amount of time spent on testing, and the extent to which practitioners feel those testing results are useful, MEPRI conducted two separate surveys. The first was a survey of district-level testing coordinators, and the second was a survey of classroom teachers. The methods for each survey are described separately below. The final research question about the possibility of using the NWEA assessment as a state summative measure was investigated through document analysis, including NWEA technical reports, website information, and analysis of state and federal policy.

District Assessment Coordinators

MEPRI first developed a statewide survey with input from district administrators and the MDOE. The MDOE Director of Assessment and Accountability sent an email message to district assessment coordinators which explained the purpose of the MEPRI survey. MEPRI obtained an email list of district assessment coordinators from the MDOE and invited members of this professional group to participate in the survey through an emailed message. This emailed message included information about the project and assured participants that the survey was confidential and that only aggregate results would be shared with the MDOE and the Maine State Legislature. The email also contained a link to the online survey, which was estimated to take 10 to 15 minutes to complete.

The survey was conducted over a period of one month during October-November 2017 and there were four reminders sent by email to non-respondents. All public school

districts were included in the survey. The only schools or districts excluded were: "tuition only" schools (SAUs without schools); state operated or special purpose schools; and private schools. In total, 123 out of 179 district assessment coordinators who were surveyed completed surveys, for an overall response rate of 68.7%. Of the 123 assessment coordinators who responded, seven individuals coordinate assessment for two districts and some of these individuals may have completed a survey for each district they serve.

Three sections of the survey asked districts about testing practices and perceptions related to state mandated tests, commercially-developed tests, and district-developed tests for students in grades K-12. For each section, participants were asked to check boxes indicating which assessments their districts conduct. They also had the option to write in "other". Second, districts were asked whether or not "teachers are encouraged to provide time for students to prep or practice for any of these tests in your district". Third, districts were asked to indicate their perception of how useful testing results were for various levels of decision making, using a six-point Likert scale ranging from (1) "not at all useful" to 6 "extremely useful". Fourth, districts were provided the option to write comments regarding any of the three types of testing they conduct. The survey concluded with demographic questions for respondents to indicate an enrollment range for the district, whether or not the district includes secondary grades 9-12, and the county location. In total, the survey included 15 questions.

All data in this report were compiled from participants' responses to a survey administered through *Qualtrics*, a web-based survey tool. Quantitative data were de-identified prior to analysis using *SPSS Statistics Version 24*. To establish internal reliability of the survey, a Cronbach alpha coefficient was calculated on the nine questions that comprised each of the three subscales, which included: testing practices and perceptions related to state-mandated tests (0.92); testing practices and perceptions related to commercially-developed tests (0.94); and testing practices and perceptions related to district-developed tests (0.81). The scaled items were analyzed using descriptive statistics (frequencies, percentages, means, and standard deviations) and inferential statistics (independent *t*-tests and oneway ANOVAs). Tables 1-12 report a majority of these findings, while Appendix C offers more in-depth statistical analyses.

For the three open-ended survey questions, a total of 69 of 118 (58.5%) participants wrote comments in response to a question about whether their district encourages teachers to provide time to students to prepare or practice for state-mandated tests, while 44 of 115 (38.3%) participants wrote comments to a similar question about commercially-developed tests, and 17 of 74 (23%) participants wrote comments regarding time for preparation or practice of district-developed tests. In total, the three open-ended questions produced 130 written responses. These comments were analyzed qualitatively through an open coding process that identified the themes and subthemes that are shared in this report.

Teacher Survey

Building on the results of the district testing coordinator surveys, a second questionnaire was administered in January 2018 to a random sample of 2,000 Maine public school classroom teachers representing all grade levels and content areas. Teachers were identified from the publicly-available state directory listing of teachers by position type, selecting "classroom teachers." Position types that included special education teachers, literacy specialists, or English Language Learner specialist teachers were not included in the survey, as these roles tend to have a different role with respect to standardized tests than that of the traditional classroom teacher. For similar reasons, teachers in charter schools were not included in the sample because those environments typically have additional testing expectations built into their charter agreements beyond state accountability tests.

The classroom teacher position type includes those teaching any of the Learning Results content areas at any grade level. Teachers' interactions with standardized tests varies substantially depending on the grade level and subject matter they teach. For example, a fifth-grade teacher responsible for the four core subjects of math, literacy, social studies, and science would have a very different connection to state test results than would a high school physical education, art, music, or health teacher. Thus, the survey was designed to gather feedback from all types of classroom teachers, but to target the types of questions asked based on the grade level (pK-2, 3-8, or high school) and subject matter taught. This allowed disaggregation of results by teacher type. Because the directory

information on classroom teachers did not provide detail on the grades or subjects taught, it was not possible to further refine the sample to specific types of teachers.

A total of 614 teachers responded to the teacher survey out of the 1,981 teachers with successful email addresses (31.0%). The survey was administered online using SurveyMonkey and invitations were sent by email. The responses were deemed sufficiently representative of the target population of classroom teachers. Teachers of non-tested subjects (e.g. music, art, physical education, etc.) were underrepresented based on known proportions of those content areas in the teacher pool. A number of teachers in these content areas emailed the research team to explain that the survey was "not relevant" or "not applicable" to their positions and thus they declined to participate. This could be seen as a finding in its own right, because standardized tests impact entire school systems but are only seen as relevant to a subset of teachers.

Teacher survey responses were analyzed using built-in SurveyMonkey analysis tools as well as *SPSS* version 21.0 to compute proportions of respondents, and mean responses on perceptions of usefulness, for the various subgroups of teachers. These results were then contrasted to similar items from district testing coordinators to explore similarities across both groups.

Survey Findings

Participation in Testing

State Assessments

District Assessment Coordinators were asked to indicate which state-mandated tests they currently administer from a provided list of tests. While these assessments are not optional, the specific tests offered in any given district may vary because of different grade level configurations (i.e. some districts operate schools at either the elementary or high school level, but not both).

All respondents reported administration of at least one state-mandated assessment. The most common state-mandated assessments in which districts participate are eMPowerME (91.1%), the MDOE Science test for 5th and 8th grade (88.6% and 74.0% respectively), and the SAT in Math and Language Arts (76.4%) (see Table 1).

Table 1. State-mandated tests administered by school districts (District coordinator survey)

(N=123 districts)	Response Percentage (% districts responding)
eMPowerME	91.1
Science (5th and 8th grade)	88.6
SAT (Math/Language Arts)	76.4
Science (11th grade)	74.0

Commercially-Developed Assessments

Of the 123 districts that completed the assessment coordinator survey, 122 (99.2%) indicated they also administer at least one type of commercially-developed assessment. Based on the survey responses, NWEA Reading and Math (65.9% and 65% respectively), the Kindergarten Screening Inventory (KSI) (56.9%), and PSAT (56.9%) are the most commonly used commercial tests by districts (see Table 2).

Table 2. Commercial tests administered by school districts (District coordinator survey)

(N= 123 districts)	Percent of Districts Using
NWEA (Reading)	65.9
NWEA (Math)	65.0
Kindergarten Screening Inventory (KSI)	56.9
PSAT	56.9
Fountas & Pinnell	40.7
NWEA (Writing)	37.4
First Grade Observation Survey for Reading Recovery	35.8
Developmental Reading Assessment (DRA)	35.0
NWEA (Science)	20.3
AIMSweb (Reading)	19.5
AIMSweb (Math)	19.5
Dynamic Reading Assessment (DIBELS)	17.9
STAR (Math)	17.9
STAR (Reading)	17.1
TS-Gold	9.8
AIMSweb (Writing)	5.7
iReady (Reading)	2.4
iReady (Math)	2.4

Teachers were also asked about the types of commercially-developed standardized tests their students take for various purposes. The proportion of teachers overall reporting that their students participate in commercially-developed tests is compatible with the combined use across all grades that was reported by district testing administrators. Table 3 provides another look at the tests most commonly used, but with added detail from the teacher survey to break down test participation by grade level. The types of tests used and frequency of use vary by grade level, and the tests are not mutually exclusive. Some

assessment packages could be used for more than one purpose and schools could administer more than one type of commercially-developed test to serve different needs. Thus the total participation rates in the last row of Table 3 are not the sum of all tests offered, but rather the proportion of regular classroom teacher respondents that indicated their students participate in one *or more* commercial tests.

Table 3. Commercially-Developed Standardized Test Usage, By Grade Span:
Percent of Core Subject Classroom Teachers Reporting Student Test Participation

	pK-2	Grades 3-8	Grades 9-12
NWEA (One or more subtests)	53%	74%	46%
Other Assessments	Fountas & Pinnell: 41% DRA: 28% Teacher's College: 28% DIBELS: 20% Clay/Rdg Rcvry: 13% STAR: 10%	STAR: 20% Teacher's College: 14% AIMSWeb: 9% DIBELS: 5% Other: 9%	PSAT: 71% Accuplacer: 43% STAR: 24%
Total Using One or more Tests	98%	93%	80%

Teachers were asked to indicate the tests in use in any grade for their subject areas, even if only some of their students participated. High school teachers reported that NWEA and STAR assessments were used most often with 9th and 10th graders, administration of PSAT was most often with 9th, 10th and 11th graders, and ACCUPLACER tests were targeted primarily at 11th and 12th graders. Within a given grade span, a test could be used with all students or just selected students depending on its purpose. Thus, while Table 3 provides a realistic depiction of the prevalence of commercial tests at different grade spans, the amount of time each individual student spends in testing can and does vary substantially.

Use of District-Developed Tests

The district survey asked testing coordinators if they conduct a district-wide writing prompt as one type of district-developed assessment and, if so, to indicate which grades are included in this assessment. Roughly half of all responding school districts (50.4%) indicated they use district-developed tests to administer a writing prompt. Districts were also provided the option to write in "other" types of district-developed tests and to indicate

the grade levels included. About 30% of all school districts indicated they use district-developed tests to administer common assessments, benchmark assessments, standards-based assessments, etc. (see Table 4). Nearly a third of districts answering this survey question (30.1%) indicated they do not administer any standardized, district-developed assessments.

Table 4. District-developed tests administered by school districts
(District coordinator survey)

(N=123 districts)	Response Percentage
Writing prompt	50.4%
Other (common assessments, benchmark assessments, standards-based assessments, etc.)	30.1%
No district developed assessments	30.1%

Self-Reported Test Preparation and Practice Time

District coordinators were also asked to indicate (yes/no) if their district encourages teachers to provide time for students to prep or practice for tests, and this question was repeated for each of the three categories of testing (state, commercial, and district). School districts (N=118) identified state-mandated tests as the category of assessment for which they are most likely to encourage their teachers to provide time for students to prepare or practice, with 83.1% of all responding district assessment coordinators stating their districts encourage this practice, compared to 62.2% who indicated their districts encourage this practice for district tests and 41.7% who stated they encourage this practice for commercially-developed tests.

Teachers were asked to provide more detail on the amount of classroom time spent preparing or practicing for tests. They corroborated the trend reported by district coordinators, reporting using more classroom time to prepare for state assessments than for other commercially-developed standardized tests. Tables 5 and 6 summarize the teachers' responses about the classroom time dedicated to practicing or preparing for the

two most common test purposes: state required tests and commercially-developed standardized tests used in universal screening.

Table 5. Teachers' Estimates of Classroom Time Spent on Test Preparation for State Assessments

	Teachers of Tested Subjects and Grades			Teachers of Tested Subjects in Untested Grades		Teachers of Untested Subjects
	Grades 3-8 Math & ELA	Grades 5 or 8 Science	Grade 11	Grades 3, 4, 6, 7 Science	Grades 9, 10, or 12	
N	*204*	*44*	*74*	*69*	*36*	*131*
None	7%	14%	42%	**67%**	39%	**54%**
< One hour	13%	5%	**10%**	12%	**19%**	11%
1 to 3 hours	**31%**	30%	19%	14%	28%	16%
3.5 to 6 hours	21%	**30%**	11%	3%	6%	6%
6.5 to 10 hours	13%	2%	8%	1%	3%	4%
> 10 hours	16%	20%	9%	3%	6%	10%

*Cells in **bold font** represent time spent by the median student on test preparation. Teachers in grades pK-2 were not asked about class time spent preparing for state tests.*

Table 6. Teachers' Estimates of Classroom Time Spent on Test Preparation for Universal Screening Assessments

	Grades pK-2*	Grades 3-8	High School
N	*75*	*198*	*106*
None	20%	30%	43%
Less than one hour	23%	20%	**16%**
1 to 3 hours	**25%**	**25%**	13%
3.5 to 6 hours	15%	10%	6%
6.5 to 10 hours	8%	4%	8%
More than 10 hours	9%	13%	13%

*Cells in **bold font** represent time spent by the median student on test prep.*

Combining these estimates of practice or preparation with information about the length of the standardized assessments commonly used in Maine, researchers developed a rough estimate of the total time a typical student in each grade spends preparing for and participating in standardized assessments. These time estimates, which are summarized in Table 7, are based on a combination of documentation and assumptions:

- Time spent in test preparation is the median time range, in hours, reported by teachers in each grade band on the teacher surveys.
- State assessment testing time is the actual median student time spent testing, as detailed in annual technical reports for each state assessment. The median times for each subtest (math, reading, writing and language, and science in selected grades) were combined and rounded to the nearest ten minutes.
- Estimates of universal screening testing time are based on an assumption of test administration using a typical assessment scheme commonly reported by teachers in each grade band, rounded to the nearest half hour:
    - NWEA for K-2 in two subjects plus Fountas and Pinnell three times per year in grades pK-2
    - NWEA administered three times per year in two subjects for grades 3-8
    - PSAT taken once or NWEA administered twice in grades 9, 10, and 12
    - ACCUPLACER or NWEA administered in grade 11 (averaged)

The result is the typical testing times represented in Table 7.

Table 7. Estimated Total Classroom Time Spent on Standardized Testing for the Median/Typical Student in 2016-17, in HOURS per year, by Grade Band

	State Assessment Test Prep / Practice	State Assessment Median Testing Time	Other Standardized Test prep / practice	Other Standardized Assessment Test Time (Total Hours)	**Estimated Typical Hours Spent on Testing per Year**
Grades pK-2	--	0	1-3	4h	**5-7**
Grades 3, 4	1-3	5h 10m	1-3	4.5h	**12-16**
Grade 5	4-6	7h 40m	1-3	4.5h	**17-21**
Grades 6, 7	1-3	4h 50m	1-3	4.5h	**12-16**
Grade 8	4-6	7h 20m	1-3	4.5h	**17-21**
Grades 9, 10, 12	<1	0	<1	3h	**4**
Grade 11	<1	5h 50m	<1	1h	**8**

As noted previously, the actual time each individual student spends in testing will depend on the number and specific type of tests used by their district for their grade level. Using a traditional schedule of 175 school days dedicated to instruction at 6 hours of school per day, the time a typical student spends on these standardized tests varies from about 0.7% of the school year for primary grades and 2% for grades 5 and 8. There is no available evidence base to suggest whether or not this is an appropriate proportion of time.

Also, most of the standardized assessments are untimed, and using the median time spent will underestimate testing time for some students and overestimate for others. Students who participate in intervention programs may take standardized progress monitoring tests in addition to the state assessments and universal screening tests, as will students tested for gifted and talented program eligibility and AP coursework. The estimate also does not include time spent preparing for or taking standardized assessments that are closely associated with district-developed common assessments because pilot conversations with practitioners discerned that teachers considered all of their classroom instruction activities to be "preparation" for district-developed assessments, and thus found the question difficult to answer reliably.

Perceptions of Usefulness

State-Mandated Tests: District perceptions

Broadly speaking, districts perceive state-mandated tests to be the least useful compared to commercial and district-developed assessments. Table 8 highlights the perspectives of school districts regarding the usefulness of state-mandated tests, specifically to inform district-level, school-level, and classroom-level decisions. Based on a six-point Likert scale, the mean responses were all less than three, indicating districts view the state assessments to be less useful to decision making at all levels. Recording a rating of four or greater on a six-point Likert scale, roughly one third of all school districts (33.5%) believe state-mandated tests are at least somewhat useful to inform district-level decisions about instruction. Given the responses of the participating school districts, the usefulness of state-mandated tests to inform instructional decisions is lower at the school level (32.7% at least somewhat useful) and much lower at the classroom level (16.1% at least somewhat useful).

Table 8. District assessment coordinator perceptions about the usefulness of state test results to inform decisions (N=119 districts)

	Mean	Std. Dev.	Not at all useful (1)	(2)	(3)	(4)	(5)	Extremely useful (6)
Inform district level decisions	2.83	1.3	17%	28%	22%	24%	9%	1%
Inform school level decisions	2.87	1.3	16%	24%	27%	23%	9%	1%
Inform classroom level decisions	2.25	1.2	34%	29%	21%	11%	5%	0%

A closer analysis of the perceived usefulness of results from state-mandated tests based on district enrollment size are reported in Table 9. Overall, none of the responses for this question had means above 3.4 on the scale of 1-6, indicating fairly negative views about the usefulness of results from state-mandated tests based on enrollment size. Additionally,

Table 9. District perceptions about the usefulness of state test results (mean scores and standard deviations) by district enrollment size (N=115)

	100 or less	101 – 199	200 – 585	586 – 1699	1700 +	Total
	N = 11	*N = 16*	*N = 20*	*N = 38*	*N = 30*	*N=115*
Inform district level decisions	2.73 (1.4)	2.25 (1.0)	3.10 (1.4)	2.74 (1.1)	3.07 (1.5)	2.82 (1.3)
Inform school level decisions	2.91 (1.6)	2.31 (1.0)	3.40 (1.4)	2.63 (1.0)	3.10 (1.3)	2.87 (1.3)
Inform classroom level decisions	2.73 (1.7)	2.06 (0.8)	***2.84*** *(1.2)*	***1.87*** *(1.1)*	2.23 (1.1)	2.24 (1.2)

Scale: 1-Not at all useful to 6-Extremely useful; $p < 0.05$

districts considered state-mandated tests to be more useful for informing school-level decisions (*M*=2.87), than for district-level decisions (M=2.82), or classroom decisions (*M*=2.24). A Oneway ANOVA revealed a statistically significant difference between school districts enrolling 200-585 (M=2.84) and school districts enrolling 586-1699 students (M=1.87) when considering the usefulness of state-mandated tests to inform instructional decisions at the classroom level (p = .033). In other words, the 38 school districts with a larger district enrollment of 585-1699 students were statistically significantly less positive about the use of state-mandated tests to inform instructional decisions at the classroom

level than the 20 school districts with a smaller enrollment of 200-585 students. There were no other statistically significant differences between district enrollment groups. Additionally, data were analyzed to determine if there was a statistically significant difference between: 1) Perceptions of state-mandated tests based on districts that include secondary grades and those that do not include secondary grades (independent *t*-test); 2) Perceptions of commercially-developed tests based on district enrollment size (ANOVA); 3) Perceptions of commercially-developed tests based on districts that include secondary grades and those that do not include secondary grades (independent *t*-test); 4) Perceptions of district-developed tests based on district enrollment size (ANOVA); and 5) Perceptions of district-developed tests based on districts that include secondary grades and those that do not include secondary grades (independent *t*-test). None of these statistical tests revealed a significant different between groups. Additional tables detailing these analyses can be found in Appendix C.

Open-ended items on the survey provided respondents with additional opportunities to share their perceptions through written comments about each category of testing. The qualitative data generated through these comments provide important and detailed ideas and information for policymakers, researchers, and practitioners to consider. Written comments about state-mandated tests were analyzed to identify broad, thematic ideas. Overall, the views expressed in the written comments were largely consistent with views indicated in response to the other survey items on state-mandated tests. Five main themes emerged from the qualitative coding of the written comments regarding the usefulness of state-mandated tests, which are described below:

Lack of timely feedback to drive instructional decisions. First, 52% of the district comments about state-mandated tests (36 out of 69) provided feedback that results from state-mandated tests were received too late to drive instructional decisions. Many participants noted that results are often received after teachers have been assigned a new group of students in the fall, making the use of data for planning purposes difficult. Many respondents stated the results would be more useful if they could be provided prior to the end of the school year. Below are several examples of responses that convey this theme:

- *"Test results come back too late. By the time we get them we have already started the school year or teachers have left for summer vacation. We do not have money*

to pay our teachers to come during the summer and most teachers work a second job and make more money at that job than contracts pay."

- *"Not very useful due to the lag time in receiving results. We need the results in late July to effect any changes in teaching strategies or curriculum for the next school year. Some use for the 2nd year out from the tests. Current MEA has not been around long enough to see long-term trends."*
- *"Because we get the information so late, we cannot really use the information to make informative decisions, which is why we supplement with alternative tests. The downside to this is that there ends up being too much testing and students get burnt out."*

Results are not useful to inform instructional decisions. Second, 33% of the comments indicated that districts believe state-mandated test results are not useful to inform instructional decisions. Some responding districts indicated testing results were less useful for instructional decisions because results are released by the state so many months after testing. Other districts indicated that state-mandated tests are less useful for informing classroom instruction because they are given only once each school year and don't clearly measure growth in individual student's learning. Below are some examples of responses that help to explain the perception that state-mandated test results are less useful for informing instructional decisions:

- *"The delay in releasing results make it hard for them to be instructionally useful at the classroom level, since by the time results are returned the teacher has an entirely different group of students."*
- *"Not useful for classroom decisions and we have other ways to measure growth and achievement. They take a lot of organizational time and disrupt instruction. It is more than the testing time - it is also the schedule disruptions and make ups. Sampling would be a good option instead of every student every year."*
- *"While the assessment results may be useful to see district/school wide trends (when testing platforms are consistent) the lapse time makes the data almost useless for individual student instruction."*

Changing state tests prevents analysis of achievement trends. Third, 25% of the written comments indicated that the adoption of different tests and revision of existing state tests over the last several years have prevented the ability to compare achievement trends over time. Participants noted the desire to be able to use state-level testing to make comparisons, but changes in the assessments have impeded the ability to use this

information to inform changes in instruction. Below are several examples of responses indicating this theme:

- *"It is frustrating to continually have changes in the State testing. It is terribly difficult to compare and see trends/patterns over time."*
- *"Because our state test assessment has changed so many times over the past few years the data is virtually useless in making any district-level decisions."*
- *"With the changes in the assessment, the timeliness of getting the results, the data system changing over the years, and not getting item analysis, it is very difficult to use the data to inform instruction."*

Lack of ability to disaggregate data. Fourth, 22% of the comments described districts' struggles to disaggregate data in order to inform instructional decisions. Again, the comments indicated a desire to use assessment results to inform instruction but shared that state-mandated test results don't currently provide in-depth information about student performance that could be used to drive instructional improvements. Further, testing data is not provided in a manner that would allow districts to easily conduct their own analysis by aggregation or disaggregation to examine student performance for groups of students or by types of testing items. Below are several representative responses:

- *"It would be helpful to have a better means to perform data analysis as well as a clearer understanding of alignment to standards."*
- *"State mandated testing data is very disconnected to the data we use in our schools and classrooms to inform programming and instruction. First, we receive the data far too late (6 months or so after the tests are taken) to impact current student programming. Second, at this point the data is at such a global level we cannot drill down far enough to look at trends and achievement on specific items for individual students."*
- *"Empower results have not been useful - results received too late, cannot aggregate in ways useful to us as an AOS, no released items."*

State-mandated tests may be useful over time. Fifth, 17% of the written comments (12 out of 69) indicated a belief that state-mandated tests may be more useful in the future. There are several caveats to this hopeful prediction of the usefulness of state-mandated tests, which include: maintaining consistency in tests administered, returning results earlier to inform immediate intervention, and closer alignment of assessments with curricula. Below are examples of responses with this theme:

- *"While the assessment results may be useful to see district/school wide trends (when testing platforms are consistent) the [time] lapse time makes the data almost useless for [informing] individual student instruction."*
- *"Now that we have 2 years of data, this will help. Keeping the type of assessments used consistent over a period of time will be most helpful."*
- *"By the time we get the results of the state-mandated tests, students have already moved on to the next grade. It's too late for any immediate intervention. It may be somewhat useful in looking at overall programming after several years, looking at trends. However, at this point, there is no state mandated test which has been around long enough to compare with our curriculum, which also changes based on student needs."*

State Assessments: Teacher Perceptions

As explained in the Methods section, the teacher survey was conducted after reviewing the feedback from the district assessment coordinators. Using the same 1 to 6 survey scale, they were asked to rate the usefulness of state-mandated assessments in several specific areas. Overall, teacher perspectives about usefulness followed similar patterns to district level testing coordinators. Teachers, like district coordinators, reported that state tests were not helpful for informing classroom-level decisions (Table 10).

Table 10: Teacher Perceptions of State Assessment Usefulness, By Type of Teaching Position (Mean Ratings and Standard Deviations)

	Grades 3-8 (N=216)	Grades 9-12 (N=105)	Untested grades / subjects (N=113)	All teachers (N=434)
Feedback to students and parents about student learning	2.39 (1.4)	2.80 (1.4)	2.78 (1.5)	2.59 (1.4)
Guiding my instructional practices	2.27 (1.4)	2.67 (1.4)	2.29 (1.4)	2.37 (1.4)
Informing changes in school or district curricula based on information about areas where students have struggled in the past	2.77 (1.4)	2.94 (1.3)	3.19 (1.5)	2.93 (1.4)
Informing communities about schools' academic performance	2.52 (1.4)	2.65 (1.3)	2.71 (1.5)	2.60 (1.4)

Scale: 1=Not at all useful to 6=Extremely useful

On all items across all grade levels, average usefulness of state assessments was below the midpoint of 3.5 on the scale of 1 to 6. Teachers found state tests slightly more useful for informing school-level curricula than for evaluating individual student learning, or guiding instruction. These are analogous to district coordinator responses related to "classroom level decisions."

Teacher survey perception items were worded with more specificity than district survey items; teachers were asked for targeted feedback on the tests' usefulness for providing feedback to students and parents, and for informing communities about school performance (i.e. school accountability). Their responses are consistent with district coordinators responses to the more generically-worded items related to informing "school level" and "district" level decisions, which could include school accountability. Notably, teachers did not find the state assessments useful even for "informing communities about schools' academic performance," which is a key purpose for which the state tests are intended (mean rating 2.60).

In optional open-ended survey items, teachers had the opportunity to comment and provide additional feedback about state standardized testing. The themes present in their comments were highly aligned to the themes brought up by district testing coordinators. The most consistent comments, by far, were complaints about the delay in availability of state assessment results. Similar to district assessment coordinators, teachers expressed frustration with their limited ability to use the results. This matched the district coordinators' theme of **Lack of timely feedback to drive instructional decisions**. Representative comments included:

- *"Taking a test in March then not receiving results till the following fall is not useful at all to a classroom teacher. I feel these tests are just for the state to see if they feel schools are doing their job. They are a waste of my instructional time."*
- *"We don't get access to our test results until the following year. So teachers can't plan too well based on that, and parents are getting feedback about how the previous grade went after it's already in the rear-view mirror."*

Less frequently, teachers expressed other reservations with the state assessment system with comments that echoed the same themes expressed by district coordinators:

- "*Scores are delivered in categories but specific errors are not shared. I do not know WHY my students do better with informative comprehension than narrative"*

(matching district coordinator theme: **Results are not useful to inform instructional decisions**)

- "*Too many changes, can't compare individual growth in specific areas for students over a period of time. Do not have access to results presented in a complete and clear manner. How well students are doing is judged differently each year by the state."* (matching district coordinator theme: **Changing state tests prevents analysis of achievement trends**)
- "*The data that we get does not tell us what standards that the students are missing. It just tells us the students individual scores and the breakdown for the entire grade. However, I can't change my classroom practices, and the school can't change their practices if we don't know what standards students are missing."* (matching district coordinator theme: **Lack of ability to disaggregate data**)

A few teachers did have positive comments about state assessments that also lined up with the district assessment coordinators (matching district coordinator theme: **State-mandated tests may be useful over time**)

- "*MEA results are useful to individual teachers who are interested in raising the overall performance of all students in a given content area. One gets a general sense of the growth or lack of growth in a particular group of students."*
- *"Used to help determine advanced placement in math and/or GT identification*"

Unlike district coordinators, some teachers expressed outright skepticism about the validity or potential usefulness of state assessments (<u>new</u> theme: **lack of validity**):

- "*The only piece of data that is useful from standardized testing is how many students are in your school. There are so many factors that can impact a student score, and an assessment only tracks one to five days of performance. Education is supposed to be about the whole person, and how they have grown in various aspects, not just academically. Social growth and emotional growth, for example, are equally important, yet can't be measured from a standardized assessment*"
- *"The questions are unfair and my kids that do well on their other tests don't on these* :("
- "*These are not helpful. Too many variables - student ability, nervousness, amount of sleep, general attitude, amount of effort, the order of presentation of content, the emphasis by different teachers, the amount of test prep time by different teachers - make this one-shot testing a waste of instruction time.*
- *"It does not factor in where the student was at the beginning of the year in terms of background knowledge and how far they have come since then. Instead, these scores keep deflating students who have difficulty learning and may never meet the benchmarks their peers do. Special education students are an example. No matter*

how much they grow throughout the year, their standardized test score will always show they never meet "the standards" set for students who are their age."

Taken together, the results of district assessment coordinator and teacher feedback reveal the limitations of Maine's statewide summative assessment system for informing purposes other than their intended purpose of state accountability.

Commercially-Developed Tests: District perceptions

Districts consistently indicated that they perceive commercially-developed tests to be most useful for district and school-level decisions. Yet, districts also said they were least likely to encourage teachers to provide time to prepare for these tests. Table 11 highlights the perspectives of districts. Based on a six-point Likert scale, all of the mean scores were well above 4. Two of the mean responses were slightly less than 5, indicating that districts view commercial assessment results as being highly useful for school and classroom-level decisions. Over four-fifths of all responding districts (82.3%) believe commercially-developed tests are at least somewhat useful to inform district-level decisions about instruction. Additionally, participating school districts responded that the usefulness of commercially-developed tests to inform instructional decisions is somewhat higher at the school level (86% at least somewhat useful) and at the classroom level (84.2% at least somewhat useful). Only district-developed assessments were considered more valuable for informing classroom-level decisions.

Table 11. District Testing Coordinator Perceptions about the Usefulness of Commercial Test Results to Inform Decisions (N=114 Districts)

	Mean	SD	Not at all useful (1)	(2)	(3)	(4)	(5)	Extremely useful (6)
Inform district level decisions	4.59	1.12	1%	1.8%	15.0%	25.7%	**32.7%**	23.9%
Inform school level decisions	4.81	1.09	0%	2.6%	11.4%	20.2%	**34.2%**	31.6%
Inform classroom level decisions	4.82	1.16	0%	3.5%	12.3%	20.2%	**27.2%**	36.8%

Cells in **bold font** represent the median respondent rating

Many districts also provided written comments to share their perceptions of commercially-developed tests. Written comments were consistent with responses to other survey questions about commercially-developed tests. Three main themes emerged from the coding of the written comments regarding the usefulness of commercially-developed tests, which are highlighted below:

Ability to analyze growth in achievement over time (and specifically to be able to **differentiate instruction**). First, 68% of the written comments about commercially-developed tests (30 out of 44) emphasized the idea that commercial tests tend to analyze test results in a variety of useful ways which provides information about different aspects of student performance. This includes the ability to analyze trends in achievement or growth over time, a focus and emphasis on student growth, and the ability to use results to provide differentiated instructional opportunities for students. Some districts also described the ability of some commercial tests to adapt to individual student performance, which is an affordance that standardized tests do not provide. Below are several examples of comments with this theme:

- *"The tests adapt to each student so instructional groupings can be made within the classroom. Students needing additional services are known right away. Trends can be seen across classrooms and across the curriculum showing what is taught well and where changes need to be made."*
- *"We use immediately available test results from NWEA to guide RtI programming and to help plan differentiated instruction. It follows the common core, which is perfectly compatible with our standards-based assessment approach. It is easy to administer, low stress (which results in more accurate data), and quick."*
- *"The immediacy of NWEA results, the normed data and individualized growth targets, and the learning continuum and student reports all are very useful in informing personalized instruction and setting school and district goals."*

Ability to disaggregate data to inform instructional decisions. Second, 52% (23 out of 44) of the comments indicated the ability of commercially-developed tests to disaggregate data for the purpose of targeting instructional decisions. By targeting specific skills and standards, the results can be used to target students for interventions, remediation, and enrichment opportunities, as well as closing achievement gaps. Below are several examples of comments with this theme:

- *"This data is extremely valuable to us. The data comes frequently throughout the year, and it comes QUICKLY. We can drill down into performance for each student to see gaps in learning, as well as use it to look globally at school and district improvement goals. We also have this data longitudinally over time (10+ years), so we can use it to look at trends in cohort growth over time (we can't do that with state data because we have been bounced around from test to test)."*
- *"Teachers use the test results to drive instruction, individually and within content teams. Students develop goals based on the information we get and use to measure progress within our RTI System."*
- *"The smaller tests are targeted towards specific standards and skills and are a better way to influence teacher decisions in instruction. They are generally shorter in length and time requirements."*

Immediate and timely results. Third, 43% of the comments (19 out of 44) stated the importance of being able to have immediate results. The use of these assessments as *formative* data (i.e., data to inform instruction) instead of *summative* data (i.e., data to measure students' achievement) to immediately inform and adjust instructional practices is important to highlight. Thus, practitioners using test results seemed to be focused on using assessment data to improve achievement and not simply measure achievement. Below are several examples of responses with this theme:

- *"The most useful aspect of these tests is the immediacy of the results. Some are more useful than others in informing teachers' instructional decisions. For example, the detailed information provided by Fountas [and] Pinnell and DIBELS is extremely useful to teachers. NWEA is somewhat useful but less so because teachers don't see the test questions."*
- *"The immediacy of results, the resources, the longitudinal student data, and the professional development opportunities all make these tests useful."*
- *"Immediate results; effectively informs instruction. Students test at their level and the emphasis is on growth."*

Commercially-Developed Tests: Teacher perceptions

Using the same scale as for prior items, teachers who reported that their students participated in standardized universal screening were asked to rate the usefulness of the tests. Results are summarized in Table 12.

Table 12: Teacher Perceptions of Usefulness of Screening Assessments (Mean Ratings and Standard Deviations; N=454)

	Grades 3-8	Grades 9-12	Grades pK-2	Untested grades / subjects	Total, All Grades
Initial identification of students who may need supplemental support/instruction	4.88 (1.3)	3.85 (1.5)	4.44 (1.3)	--	4.54 (1.4)
Feedback to students and parents about student learning	4.39 (1.4)	3.26 (1.3)	3.63 (1.4)	3.44 (1.7)	3.84 (1.5)
Guiding my instructional practices	4.36 (1.4)	3.07 (1.4)	3.88 (1.3)	2.93 (1.8)	3.71 (1.6)
Informing changes in school or district curricula based on information about areas where students have struggled in the past	4.05 (1.4)	3.08 (1.3)	3.25 (1.4)	3.42 (1.6)	3.60 (1.5)
Determining whether a student is proficient in specific learning standards	4.18 (1.5)	2.88 (1.3)	3.73 (1.3)	--	3.78 (1.5)

In general, teachers found universal screening tests to be more useful than state assessments for all purposes, and especially for their intended purpose of identifying students who may need RtI supports. Unsurprisingly, teachers of untested grades and subjects found benchmark assessments less helpful than other classroom teachers. Among teachers of tested subjects (math, literacy / ELA, or science), grades 3-8 teachers generally found the results most useful, followed by primary grade (pK-2) teachers, with high school respondents having the lowest ratings.

As with district coordinators, teachers had more positive things to say in their comments related to commercially-developed standardized assessments. Representative comments include:

Matching district coordinator themes **Ability to analyze growth in achievement over time, Ability to disaggregate data and inform instructional decisions:**

- *"The NWEA offers many beneficial features. One example is the NWEA breaks down of each test into sub-categories so students know their own strengths and weaknesses. NWEA offers teachers quadrant charts that place student based on achievement and growth. This helps me pinpoint students that need more help and*

students who are progressing. NWEA also offers goal setting sheets for students to be a part of the growth mindset, take ownership of their work, and feel proud of achievements. I can also print class data sheet that finds data points so I can reflect as a teacher what sub-categories my class does well in as a whole and what areas I need to improve."

- "*I don't always enjoy the NWEA. Sometimes I think the RIT goal at the end of the year is a difficult stretch for some kids to make. But I do try to push them there. It's improved my teaching."*

Matching district coordinator theme: **Immediate and timely results**:

- *"The speed at which we get these results helps direct my instruction for my current class right in September when we do the NWEA";*
- *"It helps to have the results immediately as opposed to have to wait months to see the data";*
- *"I am not a fan of standardized testing. But, NWEA tests do give immediate feedback which is what is most useful for my instructional purposes.";*
- *"I find the NWEAs to be useful because the information is reliable and we get the information quickly which helps to make curriculum decisions for remediation or enrichment of students in 'real-time'*."

Matching district coordinator theme: **Ability to analyze growth in achievement over time**:

- *"The data provided by NWEA is very detailed and useful to teachers. The breakdown in content and skills performance provided with the results allows teachers to pinpoint individual student strengths and weaknesses."*;
- *"Because these tests measure a student's personal growth over time, they can be helpful in pinpointing an area of need."*

In addition to the themes expressed by district testing coordinators, teachers' comments also covered new themes, both positive and negative. On the positive side, a new theme emerged about NWEA being **easier for students, parents, and teachers;** this theme was implicit in district coordinator comments but emerged with a more explicit emphasis in teacher comments such as the following:

- *"The feedback we get from the test is easy for parents to understand."*
- *"The MEA data [tools] do not offer anything compared to what the NWEA does."*
- *"This assessment is relatively quick for the students. When students take this assessment they do not get as stressed or as exhausted as the state assessment. Also, this assessment adjusts to their levels, ensuring that students are able to participate."*

In another new theme that did not appear in district coordinator comments, teachers expressed a lack of credibility in commercially-developed assessments (new theme: **inadequate accuracy, reliability, or validity for the purpose**):

- *"It's a 1 hour test that does not give me specific information to guide instruction. We use the scores as one part of determining needs."*
- *"I believe our district overuses the assessment by having students sit for it 3 times a year.*
- *"Results vary depending on many variables. For example: issues going on at home, bullying, hunger, sleep deprivation."; "They are not a good measure for those students who struggle with anxiety."*
- *"I find that these assessments are not always accurate. You are testing the student on that day and that particular time. They may do better or worse depending on what kind of day they are having."*

Some teachers also expressed frustration not with the tests themselves but with capacity to make use of them (new theme: **lack of capacity**):

- *"Our district does not provide any time to analyze and apply the information gathered from the tests to improve instruction or curriculum.*
- *"It would be a more useful test if we were provided with some professional development on A) how to interpret the data in a meaningful way and B) how to alter our classroom instruction based on that data."*

Overall, teachers preferred commercially-developed tests to state-mandated assessments, and felt that they were useful for several purposes. However, some teachers still expressed doubts about the value of these standardized assessments.

District-Developed Common Assessments: District perceptions

As stated previously, results from the survey show district-developed assessments are valued by school districts in Maine. Table 13 shows the results from the district assessment coordinator survey item about perceptions of district assessments. Based on a six-point Likert scale, the mean responses were all more than 4 for each level of decision making. The highest mean, just under 5, indicates that districts leaders view district-made assessments as highly useful for classroom-level decisions. Almost three-quarters of all school districts (74.9%) believe district-developed tests are at least somewhat useful to inform district-level decisions about instruction. Additionally, participating school districts

responded that the usefulness of district-developed tests to inform instructional decisions is higher at the school level (82.5% at least somewhat useful) and highest at the classroom level (94.2% at least somewhat useful). Regarding the ability to inform instructional decisions and improve instruction, the use of district-developed assessments was the highest (94.2%) compared to commercially-developed tests (84.2%) and state-mandated tests (16.1%).

Table 13: District perceptions about the usefulness of district test results to inform decisions (N=64 districts)

	Mean	SD	Not at all useful (1)	(2)	(3)	(4)	(5)	Extremely useful (6)
Inform district level decisions	4.18	1.2	2.9%	4.4%	17.6%	**39.7%**	17.6%	17.6%
Inform school level decisions	4.46	1.1	1.4%	1.4%	14.5%	**33.3%**	30.4%	18.8%
Inform classroom level decisions	4.93	1.0	1.4%	0%	4.3%	24.6%	**37.7%**	31.9%

Cells in **bold font** represent the median respondent rating

Written comments regarding perceptions of district-developed tests also provide important information. Overall, the views expressed by participants were consistent with other survey questions about district-developed tests. Two main themes emerged from the coding of the open-ended items regarding the usefulness of district-developed tests, which are highlighted below:

Useful data to improve instruction. First, 41% of comments about district-developed tests (7 out of 17) indicated that districts feel these assessments are useful data to improve instruction. Many participants noted that district assessments are useful in helping teachers develop a common understanding of proficiency. Below are several examples of responses for this theme:

- *"Common development and common scoring have increased teacher knowledge of writing curriculum and instruction."*
- *"We use writing prompts to help distinguish proficiency in written language standards."*

- *"We have just started using this assessment; I expect it to become more important in making district, school, and teacher instructional decisions as we become more familiar with it."*

Assessments align with district standards and instructional practices. Second, 35% (6 out of 17) of the written comments indicated that common district assessments are useful in teaching explicitly towards a set of standards. Participants highlighted the importance of district-developed assessments being imbedded in instructional practices and being used to improve student performance. Below are several examples of responses for this theme:

- *"The preparation/practice for the district-developed assessments is embedded into our instructional practice."*
- *"Again, a very useful data point that is specific to student performance in our classrooms on specific skills."*
- *"Timely and directly related to what's taught in the classroom."*
- *"We are working to improve our data collection and the fidelity of implementation of the writing curriculum and prompts. Our hope is that these efforts will yield reliable data to inform school and district goals."*

As described above, teachers were not asked about usefulness of district tests.

Context Related to Student Intervention Systems

In preliminary conversations during survey pilot testing, it was suggested that teachers' perceived usefulness of screening assessments might depend on whether their school had an adequately functioning assessment system (i.e. timely provision of data reports and capacity to use the results to inform decisions). In addition, teachers in schools without robust intervention systems might see less value in screening systems because the testing results alone do not provide additional learning resources. Thus, a short series of items was added to the teacher survey to collect information on the extent to which teachers felt their schools had strong systems in place to support student learning. This series of items was asked only of teachers of tested grades and subjects, as those categories were deemed most engaged with RTI systems. Items in **bold** and shaded gray are areas where the median teacher agreed (6) or strongly agreed (7) with the statement. Note that the "agreement" scale of 1 to 7 (1 = strongly disagree, 4 = neither agree nor disagree, 7 = strongly agree) is different from the "usefulness" scale of 1 to 6 used in prior items.

Table 14. Teacher Perceptions Related to Student Screening and Support Systems

		Teachers of Tested Subjects, by Grade Level			**Total, All Tchrs**
		pK-2	**3-8**	**9-12**	
	Number of respondents	*76*	*194*	*102*	*400*
1	My school has an adequate screening process, including testing two or more times per year, to identify students who are at-risk academically.	**5.8**	**5.7**	4.2	**5.3**
2	My school has adequate programs or supports for students who have been identified as at-risk academically (i.e., RtI Tier 2 and Tier 3 interventions).	5.1	5.0	4.5	4.9
3	My school monitors the progress of students in interventions and uses these data to determine whether interventions are working.	**5.5**	**5.3**	4.5	**5.1**
4	I have adequate time to meet in grade-level or content-area teams to discuss student progress.	3.6	4.1	3.5	3.8
5	I have adequate knowledge to know how to use test results to inform my teaching.	**5.6**	**5.5**	4.6	**5.2**
6	I have adequate resources, including appropriate data reports, to be able to use test results to inform my teaching.	5.0	5.0	4.0	4.7
7	Information that I get from standardized tests influences how or what I teach.	4.5	4.4	3.6	4.1

Overall, teachers reported agreement with statements about the adequacy of assessment processes for identification of students who may need additional academic supports, and for monitoring the progress of those receiving interventions. They also felt knowledgeable about using test results for their intended purpose. They had more moderate agreement that their schools had adequate intervention programs available to serve students, and that they had access to the data reports and tools they needed to use test results. They were neutral or disagreed that the results of testing influenced their teaching, and disagreed that they had adequate time to meet with peers to discuss student progress.

As with perceptions of usefulness, average teacher responses about MTSS systems differed by the grade level taught. High school teachers were significantly less likely to report adequate supports in place for a functioning RtI / MTSS system. Differences between pK-2 and grade 3-8 teachers were not significant. Only one item, "I have adequate time to meet in grade-level or content-area teams to discuss student progress," elicited more

disagreement than agreement. High school teachers also disagreed with the statement "Information that I get from standardized tests influences how or what I teach," and other grade levels were more neutral than in agreement.

Despite these generally positive results, it is noteworthy that there was a non-negligible respondent pool of grades pK-8 core classroom teachers that reported inadequate infrastructure in place to make use of universal screening test results. Less than 9% of these classroom teachers disagreed with the statement that the screening system of assessment at their school was adequate. However, the proportion in disagreement jumped to 23% when asked if there were adequate resources to support students who had been identified. A similar group (19%) disagreed that they had adequate resources to be able to use test results to inform teaching. While it is somewhat encouraging that a majority of teachers do feel supported in this area, the fact that about 1 in 5 do not is worthy of further attention. Given the importance of early academic interventions, and the evidence that schools are dedicating substantial resources to put identification systems in place using standardized screening tools, it is critical that teachers have the ability to act on assessment information with provision of adequate supports.

It is also important to note that teacher perceptions about MTSS/RtI programs represent one perspective. Anecdotal reports from professional organizations representing special education directors and school psychologists have surfaced concerns about the adequacy of program implementation and supports. A more complete study may be warranted to construct a more complete picture of the status of MTSS programs.

Summary of Findings

Test Use: Types of Assessments and Time Spent

MEPRI's statewide survey of school district assessment coordinators stimulated a strong response from districts with nearly a 70% response rate (68.7 %), indicating a keen interest at the district level in the topic of student testing and testing results. Teacher survey responses generally aligned well with the perceptions from district assessment coordinators. Response rates on the teacher survey were lower (31%) but were deemed acceptable by the research team.

All Maine school districts contain at least one grade level that is subject to mandatory state assessments (i.e. grades 3 to 8 and 11). The vast majority of districts include a full grade span and thus have experience with administering all of the state exams: the eMPowerME tests of math and literacy (grades 3-8), the state science assessments (grades 5, 8, and 11), and the 11th grade SAT exams in mathematics and English language arts. Some districts have a limited grade span, such as pK-8 or 9-12, and only administer the assessments that are applicable to the grades they serve.

The overwhelming majority of school districts—over 99%—also opt to administer additional commercially-developed standardized assessments to supplement the state test results. There were only a few assessment products that had high rates of use. The NWEA MAP Growth Reading and Math assessments were by far the most frequently administered assessments among responding districts (over 65%).

Input from teachers provided a more fine-grained depiction of the tests used most frequently at different grade levels. Classroom teachers in primary grades (pK-2), which do not participate in mandatory state assessments, were the most likely to use commercially-developed standardized tests; 98% reported use of one or more tests. They administered a wide variety of early literacy assessment tools (such as Fountas and Pinnell, Teacher's College Assessments, Developmental Reading Assessment, DIBELS, Clay's Observation Survey, STAR, and AIMSweb) as well as the more common NWEA MAP Growth tests. Intermediate and middle level teachers (grades 3-8) were only slightly less likely to use supplemental assessments, with 93% of classroom teachers reporting their use. They also administered NWEA MAP Growth most often, with a smaller array of other options including STAR, Teacher's College assessments, and AIMSweb. About 80% of core high school teachers administered commercial assessment tests to their students, though these were more often used in grades 9, 10, or 12 (i.e. years that students did not participate in SAT exams). The most commonly used high school tests were NWEA MAP Growth, PSAT, ACCUPLACER, and STAR.

At the district level, 50% of the responding districts indicated they also give a common writing prompt to all students in certain grade levels, and 30% of the responding districts gave other kinds of district-wide diagnostics. However, 30% of responding districts indicated they do not administer any district-made standardized assessments.

District leaders also indicated they are more likely to encourage teachers to provide time for students to prep for or practice the state-mandated assessments, than for other commercially-developed or district-made tests. This finding is interesting in contrast to the prevailing views about which tests are more useful for informing district, school, or classroom-level decisions. That is, districts said they encourage more time on preparing and practicing for state tests than other types of tests, yet there is a strong perception that the state assessment results were less useful or informative than commercial or district assessment results at all levels of decision making.

Classroom teachers in tested grades corroborated that they spend more time practicing or preparing for state assessments than for other types of tests. In grades 3 to 8, 80% of teachers reported dedicating one or more hours to practicing for state assessments while only 50% reported the same for optional tests. In high school, 48% of teachers of 11th graders reported spending one hour or more on state test prep and 40% of all teachers said the same for other tests. Using teacher responses to develop rough estimates, a typical Maine student spends between 0.5% and 2% of their school year preparing for and participating in standardized tests, depending on grade level.

Perceptions of Test Usefulness

District leaders report that commercial assessments were more useful for district and school-level decisions, and that district-made assessments were more useful for classroom-level or instructional decisions. Table 15 below presents mean score responses that illustrate these contrasting views about the usefulness of different assessments for informing decisions at each of the three levels.

Table 15. District Testing Coordinator Perceptions about the Usefulness of Test Results (mean scores and standard deviations)

	State-mandated tests	Commercially-developed tests	District-developed tests
	Mean (SD)	Mean (SD)	Mean (SD)
Inform district level decisions	2.83 (1.3)	4.59 (1.1)	4.18 (1.2)
Inform school level decisions	2.87 (1.3)	4.81 (1.1)	4.46 (1.1)
Inform classroom level decisions	2.25 (1.2)	4.82 (1.2)	4.93 (1.0)

Scale: 1-Not at all useful to 6-Extremely useful

As with district coordinators, teachers found state assessments to be less useful for their needs than other types of standardized tests. The usefulness items posed in the teacher survey were different for each test category in accordance with the varying purposes of each test. Three of the items (a to c) focused on both state assessments and universal screening tests, and three items (d to f) were posed for only one type of test or the other, as seen in Table 17.

Table 17: Comparison of Teacher Perceptions of Usefulness of Types of Standardized Assessments (All teacher types, mean responses and standard deviations)

	State Assessments	Universal Screening Assessments
Number of respondents	N=434	N=454
a. Feedback to students and parents about student learning	2.59 (1.4)	3.84 (1.5)
b. Guiding my instructional practices	2.37 (1.4)	3.71 (1.6)
c. Informing changes in school or district curricula based on information about areas where students have struggled in the past	2.93 (1.4)	3.60 (1.5)
d. Informing communities about schools' academic performance	2.60 (1.4)	--
e. Initial identification of students who may need supplemental support/instruction	--	4.54 (1.4)
f. Determining whether a student is proficient in specific learning standards	--	3.78 (1.5)

Scale: 1=Not at all useful to 6=Extremely useful

In open-ended comments, district assessment coordinators and teachers provided explanations for their perceptions that state assessment results are not informative for decision-making, particularly decisions about instruction at the classroom level. Three themes were shared by both groups. First, state test results are returned several months after administration, when students have moved on to the next grade. Second, state assessments are given at only one point in time each year, making them less useful as a measure of growth in student learning—particularly for schools with high student mobility. Third, state assessment results are not sufficiently fine-grained to allow for diagnostic information about student learning.

In addition to these problems, district assessment leaders cited the frequent overhaul in the state assessment program in recent years, which prevents analysis of change in test results over time. Districts indicated their desire for a stable assessment program, but also one that provides results in a more timely and informative manner. They

also indicated the desire to obtain test results in a form that they could disaggregate more easily on their own to examine differences among groups of students or tested content.

Teachers also expressed reservations about the accuracy and validity of standardized test results, a view that was not voiced by testing coordinators. Some teachers believed that the results did not triangulate with their own knowledge of student learning, and concluded that either the testing process or the instruments themselves were not as good as their own classroom assessment practices for understanding what students know and are able to do.

By contrast, practitioners viewed commercial and district-developed assessments to be more useful for instructional decisions for all the reasons that state tests were not useful. That is, the tests could be administered more frequently to provide information about growth in student learning. Results from commercial and district assessments were immediate or available in a timely manner. Testing results could be disaggregated to investigate different questions about student performance. All of these factors contributed to the view that commercial and district assessments could be more quickly and easily used to inform instruction, allowing classroom teachers to adjust their teaching or differentiate instruction. Finally, district assessment was valued for its close alignment with district learning standards. However, a small proportion of teachers had the same reservations about test quality and validity as had been voiced for state assessments. Nonetheless, teachers overall reported that commercially-developed tests were useful in informing their instruction and supporting student learning.

Lastly, teachers shared feedback on the extent to which they believed their schools had adequate assessment systems in place for identifying students in need of additional academic supports and providing them with supplemental instruction. Most teachers generally felt that they had adequate MTSS system components in place, though they wanted more time to meet with colleagues to discuss student progress. Some teachers reported inadequate processes in place for assessing and supporting student learning in an MTSS framework, despite state policy that mandates these systems.

Conclusions and Policy Implications

The study of standardized testing use in Maine identified several findings with potential implications for policymakers, researchers, and practitioners regarding district testing practices and perceptions. Broadly, we draw several conclusions.

First, school districts in Maine are employing a variety of standardized tests to meet different needs, and nearly all units employ additional tests beyond those required by the state. Because tests are constructed for specific purposes, the opportunities to use one test to serve multiple needs may be limited based on the assessment products currently available on the market. In particular, the exams currently in use for state testing (eMPowerME and SAT) cannot also be used for universal screening, and the most common tool for universal screening (NWEA MAP Growth) does not meet specifications as a state accountability test. As long as schools need to administer assessments for both state accountability and universal screening, the only way to materially reduce testing time would be to develop or identify a system where the assessments can serve both purposes. This is not a straightforward task, as the different purposes demand different test specifications to meet minimum thresholds of validity and reliability.

Secondly, educators at the practitioner level see an overall lack of value regarding the data provided by state-mandated tests. This is somewhat expected because the primary purpose of state assessments is for school accountability purposes, not for informing short-term classroom instructional decisions that are of most concern to practitioners. However, teachers' perceived lack of usefulness extended to "informing communities about schools' academic performance" which is a primary function of state accountability assessments. While several factors contribute to the perception of un-usefulness of state tests, the most cited challenge was the time lapse between testing and availability of results. Practitioners do report value in standardized tests other than state-required assessments for informing their instructional and policy needs, so their perceptions cannot be attributed to a general lack of appreciation for the potential of assessment tools.

Some perspectives are decidedly absent from the scope of this study – i.e. those of district school boards, superintendents, the State Board of Education, and the State

Department of Education. These stakeholders have different needs than practitioners, and are the ones for whom having stable and reliable assessment results that can be compared across schools and districts is most important. Their viewpoints would be critically important to add if the scope were expanded to have a complete and balanced view of the perceived value of state summative assessments.

Finally, to improve instruction, districts need the ability to disaggregate data from assessments to target specific areas for improvement. Districts report difficulty in disaggregating results of state assessments. The ability to differentiate instruction is crucial to improving student achievement outcomes. Moreover, there is a lack of alignment with the philosophy of proficiency-based education (currently required for diplomas issued in Maine) and the quality of data provided from state-mandated tests. Greater alignment between Proficiency-Based Education (PBE) and data provided by assessments that focus on growth could be helpful. Testing results need to be fine-grained enough to be informative for all of these instructional purposes, and the current state assessment results were not perceived as sufficient.

Based on these conclusions, researchers identified policy implications and potential next steps in three areas: increasing perceived validity of tests, reducing testing time, and issues for further study.

To increase the perceived value of state-mandated tests, finding ways to markedly increase the timeliness of feedback would be important. While the assessments may never be optimal for informing day-to-day instructional decisions as that is not their intended purpose, the delay in availability of results rules out other potential uses such as informing curriculum revisions and school-level support needs. Stability in the assessment program is also important, as it allows comparison of achievement over time.

Next, if policymakers judge that the current amount of instructional time spent on testing is not acceptable, there are limited options using existing assessment products for decreasing testing. No current products are readily configured to meet both federal accountability requirements and Maine's MTSS/RtI expectations, and a new assessment system would need to be developed. The following options could be explored for feasibility; all would require an investment of time, expertise, and funding to develop and implement:

- A growth-based universal screening system such as NWEA MAP Growth could be modified so that the final test administration in the sequence can also meet federal testing requirements. This would require a longer test period for the final test so that there is sufficient reliability to use the results for accountability purposes—i.e. the final test would need to be similar in length to current state assessments. It is likely that students that are substantially below or above grade-level expectations would not see much of a decrease in testing time, because in order to complete the end-of-year growth trajectory they would still need to answer the appropriate computer-adaptive test items to hone in on their learning at a different grade level.
- A status-based universal screening system such as eMPowerME Benchmarks could be similarly modified to use the full assessment as the final screening assessment. This option is likely to be quite feasible without a large outlay of development funds, but would still require changes in school practices. It is unclear whether districts currently using NWEA MAP Growth would be willing to give up a growth-focused measure in exchange for some cost savings and a modest reduction in testing time.
- Without a large investment in new test development, it may be possible to develop and use "concordance tables" based on NWEA MAP Growth assessment results to estimate student-level proficiency in state grade level expectations. Such concordance tables have been developed for other states for alignment purposes. The results would only be usable in the ESSA accountability system if the U.S. Department of Education were to accept the results as a valid approach to meeting federal statutory requirements; there is no current precedent for such a method. If approved, schools could switch entirely to NWEA and discontinue the current state assessment system that they find unhelpful.
- Another option that would require federal policy approval would be to use NWEA MAP Growth as the basis for the individual growth measure in Maine's ESSA accountability system, and to use a sampling system similar to NAEP for measurement of overall school-level proficiency. This may require a liberal interpretation of federal statute.

- Within existing federal statute, states may develop assessment systems that are administered at multiple points in time rather than all at once at the end of the year. Thus the grade-level content could be assessed through items included on universal screening tests. Ideally, such a system would be constructed to robustly measure growth and improvement of student learning.
- As a different approach to reducing total testing time, instead of combining *universal screening* with state accountability assessments Maine could combine *course assessment* with state accountability. In other words, standardized end-of-course exams that serve as summative measures of student proficiency and course learning can also be used as accountability assessments if the content tested aligns with grade-level expectations, as is required in federal statute. This would only have a net reduction in testing time if schools used the tests to decrease classroom assessments that measure the same knowledge areas—i.e. substitute standardized end-of-course exams for teacher-developed final exams. This might also have added value for informing proficiency-based diploma systems in Maine's current context.

Lastly, there were several areas identified over the course of the research that would benefit from further study and/or data collection, including:

- Investigation of emerging alternative accountability testing options. Testing companies have research and development segments that are constantly working to improve the products they sell, and states have been working to develop new testing systems – such as a district-level performance-based assessment system being piloted in New Hampshire – that may present an opportunity for replication in Maine.
- Additional data collection from Maine stakeholders, such as feedback from policymakers about the criteria that would make a state summative assessment more useful and whether the estimates of time spent on testing are appropriate. It may also be valuable to collect empirical data about the status of implementation of MTSS / RtI systems to include perceptions from additional stakeholder groups. It is unknown whether teachers' perceptions of what constitutes an adequate MTSS

system are the same as those that are held by specialists that function primarily at the higher tiers of student interventions. A more complete picture may emerge with additional perspectives.

In summary, Maine's current array of standardized testing practices appears to be meeting both the state's need for a valid and reliable assessment system that complies with federal requirements, and districts' need for more timely and detailed information about individual student learning. Researchers did not identify any readily available options for decreasing testing time; achieving additional efficiencies would require development of new assessment products. Additional study might help to maintain continued attention to the issue of testing time so that improvements can be identified and implemented when feasible.

References

2016-2017 eMPowerME ELA/Literacy and Mathematics Technical Report. Retrieved from http://www.maine.gov/doe/assessment/math-ela/documents/2016-17%20MeCAS%20ELA%20and%20Math%20Tech%20Report.pdf

Black, P. & William, D. (2009). Developing the theory of formative assessment. *Education, Assessment, Evaluation, and Accountability, 21,* 5-31.

Elliott, S. N., Witt, J. C., Kratochwill, T. R., & Stoiber, K. C. (2002). Selecting and evaluating classroom interventions. In M. R. Shinn, H. M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavioral problems II: Preventive and remedial approaches* (pp. 243–294). Bethesda, MD: National Association of School Psychologists

FairTest, The National Center for Fair and Open Testing (2017). "Test Reform Victories Surge in 2017, What's Behind the Winning Strategies?" Retrieved from http://www.fairtest.org/fairtest-report-test-reform-victories-surge-in-2017

Fuchs, D., Compton, D. L., Fuchs, L. S., & Bryant, J. (2008). Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal reading study at the National Research Center on Learning Disabilities. *Reading and Writing: An Interdisciplinary Journal, 21*, 413–436.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to responsiveness-to-intervention: What, why, and how valid is it? *Reading Research Quarterly, 4*, 93–99.

Harlen, W. & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principle, Policy, and Practice, 4*, 365-379.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*, 582–600.

Johnson, E. S., Pool, J., & Carter, D. R. Screening for Reading Problems in Grades 1 Through 3: An Overview of Select Measures. Retrieved from http://www.rtinetwork.org/essential/assessment/screening/screening-for-reading-problems-in-grades-1-through-3

Kazdin, A. E. (1977). Assessing the clinical or applied significance of behavior change through social validation. *Behavior Modification, 1*, 427–452.

Koretz, D. (2017). The Testing Charade: Pretending to Make Schools Better. University of Chicago Press.

MEPRI (2017). Analysis of Essential Programs and Services Components: Student Assessment.

NWEA's Measures of Academic Progress (MAP). (January, 2013). Retrieved from https://www.edweek.org/media/nweamyths-blog.pdf

NWEA MAP Concordance to Ohio State Assessments https://www.nwea.org/content/uploads/2017/01/OH-MAP-Growth-Linking-Study-AUG2016.pdf

NWEA MAP Suite. (2018). Retrieved from https://www.nwea.org/the-map-suite/

Shaw, L. (2013). Educators debate the validity of MAP testing. *Seattle Times.* Retrieved from https://www.seattletimes.com/seattle-news/educators-debate-validity-of-map-testing/

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP Norms for Student and School Achievement Status and Growth*. Retrieved from http://www.sowashco.org/files/department/rea/2015NormsReport_Reading.pdf

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum

(i) The State may average data from the school year for which the determination is made with data from one or two school years immediately preceding that school year.

(ii) Until the assessments described in paragraph (3) are administered in such manner and time to allow for the implementation of the uniform procedure for averaging data described in clause (i), the State may use the academic assessments that were required under paragraph (3) as that paragraph was in effect on the day preceding the date of enactment of the No Child Left Behind Act of 2001, provided that nothing in this clause shall be construed to undermine or delay the determination of adequate yearly progress, the requirements of section 1116, or the implementation of assessments under this section.

(iii) The State may use data across grades in a school.

(K) ACCOUNTABILITY FOR CHARTER SCHOOLS.—The accountability provisions under this Act shall be overseen for charter schools in accordance with State charter school law.

[Note: Effective on August 2, 2016, paragraph (2), as amended by section 1005 of Public Law 114–95, is amended to read as follows:]

(2) ACADEMIC ASSESSMENTS.—

(A) IN GENERAL.—Each State plan shall demonstrate that the State educational agency, in consultation with local educational agencies, has implemented a set of high-quality student academic assessments in mathematics, reading or language arts, and science. The State retains the right to implement such assessments in any other subject chosen by the State.

(B) REQUIREMENTS.—The assessments under subparagraph (A) shall—

(i) except as provided in subparagraph (D), be—

(I) the same academic assessments used to measure the achievement of all public elementary school and secondary school students in the State; and

(II) administered to all public elementary school and secondary school students in the State;

(ii) be aligned with the challenging State academic standards, and provide coherent and timely information about student attainment of such standards and whether the student is performing at the student's grade level;

(iii) be used for purposes for which such assessments are valid and reliable, consistent with relevant, nationally recognized professional and technical testing standards, objectively measure academic achievement, knowledge, and skills, and be tests that do not evaluate

or assess personal or family beliefs and attitudes, or publicly disclose personally identifiable information;

(iv) be of adequate technical quality for each purpose required under this Act and consistent with the requirements of this section, the evidence of which shall be made public, including on the website of the State educational agency;

(v)(I) in the case of mathematics and reading or language arts, be administered—

(aa) in each of grades 3 through 8; and

(bb) at least once in grades 9 through 12;

(II) in the case of science, be administered not less than one time during—

(aa) grades 3 through 5;

(bb) grades 6 through 9; and

(cc) grades 10 through 12; and

(III) in the case of any other subject chosen by the State, be administered at the discretion of the State;

(vi) involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding, which may include measures of student academic growth and may be partially delivered in the form of portfolios, projects, or extended performance tasks;

(vii) provide for—

(I) the participation in such assessments of all students;

(II) the appropriate accommodations, such as interoperability with, and ability to use, assistive technology, for children with disabilities (as defined in section 602(3) of the Individuals with Disabilities Education Act (20 U.S.C. 1401(3))), including students with the most significant cognitive disabilities, and students with a disability who are provided accommodations under an Act other than the Individuals with Disabilities Education Act (20 U.S.C. 1400 et seq.), necessary to measure the academic achievement of such children relative to the challenging State academic standards or alternate academic achievement standards described in paragraph (1)(E); and

(III) the inclusion of English learners, who shall be assessed in a valid and reliable manner and provided appropriate accommodations on assessments administered to such students under this paragraph, including, to the extent practicable, assessments in the language and form most likely to yield accurate data on what such students know and can do in academic content areas, until such students have achieved English language proficiency, as determined under subparagraph (G);

(viii) at the State's discretion—

(I) be administered through a single summative assessment; or

(II) be administered through multiple statewide interim assessments during the course of the academic year that result in a single summative score that provides valid, reliable, and transparent information on student achievement or growth;

(ix) notwithstanding clause (vii)(III), provide for assessments (using tests in English) of reading or language arts of any student who has attended school in the United States (not including the Commonwealth of Puerto Rico) for 3 or more consecutive school years, except that if the local educational agency determines, on a case-by-case individual basis, that academic assessments in another language or form would likely yield more accurate and reliable information on what such student knows and can do, the local educational agency may make a determination to assess such student in the appropriate language other than English for a period that does not exceed 2 additional consecutive years, provided that such student has not yet reached a level of English language proficiency sufficient to yield valid and reliable information on what such student knows and can do on tests (written in English) of reading or language arts;

(x) produce individual student interpretive, descriptive, and diagnostic reports, consistent with clause (iii), regarding achievement on such assessments that allow parents, teachers, principals, and other school leaders to understand and address the specific academic needs of students, and that are provided to parents, teachers, and school leaders, as soon as is practicable after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand;

(xi) enable results to be disaggregated within each State, local educational agency, and school by—

(I) each major racial and ethnic group;

(II) economically disadvantaged students as compared to students who are not economically disadvantaged;

(III) children with disabilities as compared to children without disabilities;

(IV) English proficiency status;

(V) gender; and

(VI) migrant status,

except that such disaggregation shall not be required in the case of a State, local educational agency, or a school in which the number of students in a subgroup is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student;

(xii) enable itemized score analyses to be produced and reported, consistent with clause (iii), to local edu-

cational agencies and schools, so that parents, teachers, principals, other school leaders, and administrators can interpret and address the specific academic needs of students as indicated by the students' achievement on assessment items; and

(xiii) be developed, to the extent practicable, using the principles of universal design for learning.

(C) EXCEPTION FOR ADVANCED MATHEMATICS IN MIDDLE SCHOOL.—A State may exempt any 8th grade student from the assessment in mathematics described in subparagraph (B)(v)(I)(aa) if—

(i) such student takes the end-of-course assessment the State typically administers to meet the requirements of subparagraph (B)(v)(I)(bb) in mathematics;

(ii) such student's achievement on such end-of-course assessment is used for purposes of subsection (c)(4)(B)(i), in lieu of such student's achievement on the mathematics assessment required under subparagraph (B)(v)(I)(aa), and such student is counted as participating in the assessment for purposes of subsection (c)(4)(B)(vi); and

(iii) in high school, such student takes a mathematics assessment pursuant to subparagraph (B)(v)(I)(bb) that—

(I) is any end-of-course assessment or other assessment that is more advanced than the assessment taken by such student under clause (i) of this subparagraph; and

(II) shall be used to measure such student's academic achievement for purposes of subsection (c)(4)(B)(i).

(D) ALTERNATE ASSESSMENTS FOR STUDENTS WITH THE MOST SIGNIFICANT COGNITIVE DISABILITIES.—

(i) ALTERNATE ASSESSMENTS ALIGNED WITH ALTERNATE ACADEMIC ACHIEVEMENT STANDARDS.—A State may provide for alternate assessments aligned with the challenging State academic standards and alternate academic achievement standards described in paragraph (1)(E) for students with the most significant cognitive disabilities, if the State—

(I) consistent with clause (ii), ensures that, for each subject, the total number of students assessed in such subject using the alternate assessments does not exceed 1 percent of the total number of all students in the State who are assessed in such subject;

(II) ensures that the parents of such students are clearly informed, as part of the process for developing the individualized education program (as defined in section 614(d)(1)(A) of the Individuals with Disabilities Education Act (20 U.S.C. 1414(d)(1)(A)))—

(aa) that their child's academic achievement will be measured based on such alternate standards; and

(bb) how participation in such assessments may delay or otherwise affect the student from completing the requirements for a regular high school diploma;

(III) promotes, consistent with the Individuals with Disabilities Education Act (20 U.S.C. 1400 et seq.), the involvement and progress of students with the most significant cognitive disabilities in the general education curriculum;

(IV) describes in the State plan the steps the State has taken to incorporate universal design for learning, to the extent feasible, in alternate assessments;

(V) describes in the State plan that general and special education teachers, and other appropriate staff—

(aa) know how to administer the alternate assessments; and

(bb) make appropriate use of accommodations for students with disabilities on all assessments required under this paragraph;

(VI) develops, disseminates information on, and promotes the use of appropriate accommodations to increase the number of students with significant cognitive disabilities—

(aa) participating in academic instruction and assessments for the grade level in which the student is enrolled; and

(bb) who are tested based on challenging State academic standards for the grade level in which the student is enrolled; and

(VII) does not preclude a student with the most significant cognitive disabilities who takes an alternate assessment based on alternate academic achievement standards from attempting to complete the requirements for a regular high school diploma.

(ii) SPECIAL RULES.—

(I) RESPONSIBILITY UNDER IDEA.—Subject to the authority and requirements for the individualized education program team for a child with a disability under section 614(d)(1)(A)(i)(VI)(bb) of the Individuals with Disabilities Education Act (20 U.S.C. 1414(d)(1)(A)(i)(VI)(bb)), such team, consistent with the guidelines established by the State and required under section 612(a)(16)(C) of such Act (20 U.S.C. 1412(c)(16)(C)) and clause (i)(II) of this subparagraph, shall determine when a child with a significant cognitive disability shall participate in an alternate assessment aligned

with the alternate academic achievement standards.

(II) PROHIBITION ON LOCAL CAP.—Nothing in this subparagraph shall be construed to permit the Secretary or a State educational agency to impose on any local educational agency a cap on the percentage of students administered an alternate assessment under this subparagraph, except that a local educational agency exceeding the cap applied to the State under clause (i)(I) shall submit information to the State educational agency justifying the need to exceed such cap.

(III) STATE SUPPORT.—A State shall provide appropriate oversight, as determined by the State, of any local educational agency that is required to submit information to the State under subclause (II).

(IV) WAIVER AUTHORITY.—This subparagraph shall be subject to the waiver authority under section 8401.

(E) STATE AUTHORITY.—If a State educational agency provides evidence, which is satisfactory to the Secretary, that neither the State educational agency nor any other State government official, agency, or entity has sufficient authority, under State law, to adopt challenging State academic standards, and academic assessments aligned with such standards, which will be applicable to all students enrolled in the State's public elementary schools and secondary schools, then the State educational agency may meet the requirements of this subsection by—

(i) adopting academic standards and academic assessments that meet the requirements of this subsection, on a statewide basis, and limiting their applicability to students served under this part; or

(ii) adopting and implementing policies that ensure that each local educational agency in the State that receives grants under this part will adopt academic content and student academic achievement standards, and academic assessments aligned with such standards, which—

(I) meet all of the criteria in this subsection and any regulations regarding such standards and assessments that the Secretary may publish; and

(II) are applicable to all students served by each such local educational agency.

(F) LANGUAGE ASSESSMENTS.—

(i) IN GENERAL.—Each State plan shall identify the languages other than English that are present to a significant extent in the participating student population of the State and indicate the languages for which annual student academic assessments are not available and are needed.

(ii) SECRETARIAL ASSISTANCE.—The State shall make every effort to develop such assessments and may

request assistance from the Secretary if linguistically accessible academic assessment measures are needed. Upon request, the Secretary shall assist with the identification of appropriate academic assessment measures in the needed languages, but shall not mandate a specific academic assessment or mode of instruction.

(G) ASSESSMENTS OF ENGLISH LANGUAGE PROFICIENCY.—

(i) IN GENERAL.—Each State plan shall demonstrate that local educational agencies in the State will provide for an annual assessment of English proficiency of all English learners in the schools served by the State educational agency.

(ii) ALIGNMENT.—The assessments described in clause (i) shall be aligned with the State's English language proficiency standards described in paragraph (1)(F).

(H) LOCALLY-SELECTED ASSESSMENT.—

(i) IN GENERAL.—Nothing in this paragraph shall be construed to prohibit a local educational agency from administering a locally-selected assessment in lieu of the State-designed academic assessment under subclause (I)(bb) and subclause (II)(cc) of subparagraph (B)(v), if the local educational agency selects a nationally-recognized high school academic assessment that has been approved for use by the State as described in clause (iii) or (iv) of this subparagraph.

(ii) STATE TECHNICAL CRITERIA.—To allow for State approval of nationally-recognized high school academic assessments that are available for local selection under clause (i), a State educational agency shall establish technical criteria to determine if any such assessment meets the requirements of clause (v).

(iii) STATE APPROVAL.—If a State educational agency chooses to make a nationally-recognized high school assessment available for selection by a local educational agency under clause (i), which has not already been approved under this clause, such State educational agency shall—

(I) conduct a review of the assessment to determine if such assessment meets or exceeds the technical criteria established by the State educational agency under clause (ii);

(II) submit evidence in accordance with subsection (a)(4) that demonstrates such assessment meets the requirements of clause (v); and

(III) after fulfilling the requirements of subclauses (I) and (II), approve such assessment for selection and use by any local educational agency that requests to use such assessment under clause (i).

(iv) LOCAL EDUCATIONAL AGENCY OPTION.—

(I) LOCAL EDUCATIONAL AGENCY.—If a local educational agency chooses to submit a nationally-

recognized high school academic assessment to the State educational agency, subject to the approval process described in subclause (I) and subclause (II) of clause (iii) to determine if such assessment fulfills the requirements of clause (v), the State educational agency may approve the use of such assessment consistent with clause (i).

(II) STATE EDUCATIONAL AGENCY.—Upon such approval, the State educational agency shall approve the use of such assessment in any other local educational agency in the State that subsequently requests to use such assessment without repeating the process described in subclauses (I) and (II) of clause (iii).

(v) REQUIREMENTS.—To receive approval from the State educational agency under clause (iii), a locally-selected assessment shall—

(I) be aligned to the State's academic content standards under paragraph (1), address the depth and breadth of such standards, and be equivalent in its content coverage, difficulty, and quality to the State-designed assessments under this paragraph (and may be more rigorous in its content coverage and difficulty than such State-designed assessments);

(II) provide comparable, valid, and reliable data on academic achievement, as compared to the State-designed assessments, for all students and for each subgroup of students defined in subsection (c)(2), with results expressed in terms consistent with the State's academic achievement standards under paragraph (1), among all local educational agencies within the State;

(III) meet the requirements for the assessments under subparagraph (B) of this paragraph, including technical criteria, except the requirement under clause (i) of such subparagraph; and

(IV) provide unbiased, rational, and consistent differentiation between schools within the State to meet the requirements of subsection (c).

(vi) PARENTAL NOTIFICATION.—A local educational agency shall notify the parents of high school students served by the local educational agency—

(I) of its request to the State educational agency for approval to administer a locally-selected assessment; and

(II) upon approval, and at the beginning of each subsequent school year during which the locally selected assessment will be administered, that the local educational agency will be administering a different assessment than the State-designed assessments under subclause (I)(bb) and subclause (II)(cc) of subparagraph (B)(v).

(I) DEFERRAL.—A State may defer the commencement, or suspend the administration, but not cease the development, of the assessments described in this paragraph, for 1 year for each year for which the amount appropriated for grants under part B is less than $369,100,000.

(J) ADAPTIVE ASSESSMENTS.—

(i) IN GENERAL.—Subject to clause (ii), a State retains the right to develop and administer computer adaptive assessments as the assessments described in this paragraph, provided the computer adaptive assessments meet the requirements of this paragraph, except that—

(I) subparagraph (B)(i) shall not be interpreted to require that all students taking the computer adaptive assessment be administered the same assessment items; and

(II) such assessment—

(aa) shall measure, at a minimum, each student's academic proficiency based on the challenging State academic standards for the student's grade level and growth toward such standards; and

(bb) may measure the student's level of academic proficiency and growth using items above or below the student's grade level, including for use as part of a State's accountability system under subsection (c).

(ii) STUDENTS WITH THE MOST SIGNIFICANT COGNITIVE DISABILITIES AND ENGLISH LEARNERS.—In developing and administering computer adaptive assessments—

(I) as the assessments allowed under subparagraph (D), a State shall ensure that such computer adaptive assessments—

(aa) meet the requirements of this paragraph, including subparagraph (D), except such assessments shall not be required to meet the requirements of clause (i)(II); and

(bb) assess the student's academic achievement to measure, in the subject being assessed, whether the student is performing at the student's grade level; and

(II) as the assessments required under subparagraph (G), a State shall ensure that such computer adaptive assessments—

(aa) meet the requirements of this paragraph, including subparagraph (G), except such assessment shall not be required to meet the requirements of clause (i)(II); and

(bb) assess the student's language proficiency, which may include growth towards such proficiency, in order to measure the student's acquisition of English.

(K) RULE OF CONSTRUCTION ON PARENT RIGHTS.—Nothing in this paragraph shall be construed as preempting a State or local law regarding the decision of a parent to not have the parent's child participate in the academic assessments under this paragraph.

(L) LIMITATION ON ASSESSMENT TIME.—Subject to Federal or State requirements related to assessments, evaluations, and accommodations, each State may, at the sole discretion of such State, set a target limit on the aggregate amount of time devoted to the administration of assessments for each grade, expressed as a percentage of annual instructional hours.

---

(3) EXCEPTION FOR RECENTLY ARRIVED ENGLISH LEARNERS.—

(A) ASSESSMENTS.—With respect to recently arrived English learners who have been enrolled in a school in one of the 50 States in the United States or the District of Columbia for less than 12 months, a State may choose to—

(i) exclude—

(I) such an English learner from one administration of the reading or language arts assessment required under paragraph (2); and

(II) such an English learner's results on any of the assessments required under paragraph (2)(B)(v)(I) or (2)(G) for the first year of the English learner's enrollment in such a school for the purposes of the State-determined accountability system under subsection (c); or

(ii)(I) assess, and report the performance of, such an English learner on the reading or language arts and mathematics assessments required under paragraph (2)(B)(v)(I) in each year of the student's enrollment in such a school; and

(II) for the purposes of the State-determined accountability system—

(aa) for the first year of the student's enrollment in such a school, exclude the results on the assessments described in subclause (I);

(bb) include a measure of student growth on the assessments described in subclause (I) in the second year of the student's enrollment in such a school; and

(cc) include proficiency on the assessments described in subclause (I) in the third year of the student's enrollment in such a school, and each succeeding year of such enrollment.

(B) ENGLISH LEARNER SUBGROUP.—With respect to a student previously identified as an English learner and for not more than 4 years after the student ceases to be identified as an English learner, a State may include the results of the student's assessments under paragraph (2)(B)(v)(I) within the English learner subgroup of the subgroups of students (as defined in subsection (c)(2)(D)) for

Appendix B: Standardized Testing Costs

In initial planning, the scope of the present study included an analysis of school district expenditures for standardized testing. However, in early stages of analysis it became apparent that districts appear to be using different methods for coding and reporting their assessment expenditures (MEPRI, 2017). This reduces confidence that the available data are an accurate representation of actual district spending. In addition, the expenditure data are not detailed enough to isolate funds specifically spent on standardized tests as opposed to other costs, such as development of new assessments to prepare for proficiency-based diploma systems. Thus, the initial attempt to analyze existing data to capture costs was discontinued.

Instead, data on the costs of assessments were researched to provide some context in the spirit of the initial plan. These data points are not a complete depiction of all the costs of standardized testing as they only cover the price of test purchase, administration, and reporting. The full costs would also include expenses for employee salaries, stipends, and benefits for time spent developing or managing the tests, professional development, consulting services, and technology costs for testing equipment (including computers or rentals).

State Assessment Costs

The Maine Department of Education is engaged in a one-year contract with Measured Progress from 11/1/17 to 10/31/18 to administer the state assessment system, including eMPowerME, SAT, science tests, alternative assessments for students with significant cognitive impairments, and ACCESS tests for English Learners. The total cost of the system is $ 4,562,412, with $1,500,000 paid from general fund and the remainder from federal funds. Based on Maine's prior year (AY2017) total K-12 student enrollment of 174,481, this equates to approximately $26 total per Maine student and $9 in general funds per student for the annual state accountability system assessments. About 94,000 of the total students were enrolled in tested grades and eligible for assessment participation.

Other Standardized Assessment Costs

Researchers also investigated the currently advertised prices for some of the commercially-developed standardized assessments used most commonly in Maine public schools. As above, the results do not include the full costs of the program including teacher time and equipment, just the cost of the exam product. At least one assessment company indicated during a follow-up conversation that prices are negotiated with each participating district, and the actual price paid per student may be discounted in practice.

Table B1. Recent Price Per Student for Selected Standardized Assessments

Assessment	Price Per Student
NWEA MAP Growth (110 student minimum)	$13.50
PSAT	$16.00
ACCUPLACER	~ $2
AIMSweb (Math and Reading)	$13.00
Fountas & Pinnell: Flat fee of $425 for assessment materials	
STAR: software purchase of $1600 per subject area (math, reading, early literacy), then $.99 per student per subject in subsequent years ($200 minimum)	

Appendix C1

District Testing Practices in Maine - October 2017
MEPRI Study to Audit School District Testing Practices in Maine
District Survey

Introduction: The Maine State Legislature has asked the Maine Education Policy Research Institute (MEPRI) to conduct an audit of standardized testing (tests administered with a common format and items to students) that is conducted by school districts in Maine. This survey will collect some general information about what tests are given in various grade levels and feedback from a district perspective about the usefulness of testing results for informing decisions. The survey is confidential. No individuals or their school districts will be identified in any reports from this survey. Only aggregate results will be shared with the Maine State Legislature and the Maine Dept. of Education. The survey should be completed by the designated district assessment coordinator by **November 22**. Completing the survey should take approximately 10-15 minutes.

If you are a school administrative unit operating without schools (tuition only) you do not need to complete this survey

Guidance on Completing the Survey: This survey asks for information about academic tests given to regular education students (not alternative or diagnostic tests) from any of the following sources:

1. **State mandated tests** for regular education students grades 3-11 (e.g., eMPowerME for grades 3-8, science grades 5 and 8, science grade 11, SAT for math and ELA grade 11)
2. **Commercially developed tests** that are standardized and given to a grade level or several grades district-wide (e.g., AIMSWEB, NWEA, etc.)
3. **District developed tests** conducted in the same format for an entire grade level or several grades district-wide (e.g., a writing prompt conducted district wide)

We realize there are many other types of formal testing or assessments that your district may conduct or participate in, such as: early literacy diagnostic testing, alternative and diagnostic testing for students with IEPs; national assessments (NAEP, Middle Grades Longitudinal Study); international assessments (TIMSS—Trends in International Mathematics and Science Study); Gifted and Talented screening tests; Health risk assessments (Gallup Poll, MYIHS-Maine Youth Integrated Health Survey; PEAR Holistic Assessment); classroom-based assessments (e.g., end of unit or end of course assessments including Advanced Placement tests); or college placement tests (e.g., ACCUPLACER). **This survey will not ask about those types of assessments.**

Part 1: State Mandated Tests

Q1. Please indicate which **statewide tests** your district currently administers:

- ☐ eMPowerME, grades 3 - 8
- ☐ Science, grade 5 and 8
- ☐ Science, grade 11
- ☐ Math SAT/ ELA SAT, grade 11
- ☐ Other, please specify test name and grades: _____
- ☐ Other, please specify test name and grades: _____
- ☐ My district DOES NOT administer any state mandated tests

Skip To: Section 2 if Q3 = My district does not administer any state mandated tests

Q2. Are teachers encouraged to provide time for students to prep or practice for any of the **state mandated** tests in your district?

- ○ Yes
- ○ No

Q3. From a district perspective, please rate how useful results are from **state mandated** tests for decisions about educational programs or instruction, using a scale of 1 (*not at all useful*) to 6 (*extremely useful*).

	1 - not at all useful	2	3	4	5	6 - extremely useful
Informing district-level decisions	○	○	○	○	○	○
Informing school-level decisions	○	○	○	○	○	○
Informing teachers' instructional decisions in the classroom	○	○	○	○	○	○

Q4. Comments (optional) about state mandated tests:

Part 2: Commercially Developed Tests

Q1. Please indicate which **standardized, commercially developed academic tests** are administered to a majority of your students district-wide.

Please use check boxes to indicate any test administered in your district, and specify for which grades the test is administered.

- ☐ Kindergarten Screening Inventory; list grades: _____
- ☐ Fountas & Pinnell; list grades: _____
- ☐ TS-Gold; list grades: _____
- ☐ iReady (Reading); list grades: _____
- ☐ iReady (Math); list grades: _____
- ☐ DRA-Developmental Reading Assessment; list grades: _____
- ☐ DIBELS-Dynamic Reading Assessment; list grades: _____
- ☐ First Grade Observation Survey for Reading Recovery; list grades: _____
- ☐ NWEA-Northwest Evaluation Association (Reading); list grades: _____
- ☐ NWEA-Northwest Evaluation Association (Language Arts); list grades: _____
- ☐ NWEA-Northwest Evaluation Association (Math); list grades: _____
- ☐ NWEA-Northwest Evaluation Association (Science); list grades: _____
- ☐ AIMSweb (Reading); list grades: _____
- ☐ AIMSweb (Writing); list grades: _____
- ☐ AIMSweb (Math); list grades: _____
- ☐ PSAT; list grades: _____
- ☐ STAR (Reading); list grades: _____
- ☐ STAR (Math); list grades: _____
- ☐ TerraNova; list grades: _____
- ☐ Other, please specify test name and grades: _____
- ☐ Other, please specify test name and grades _____
- ☐ My district DOES NOT administer any standardized, commercially-developed academic tests (skips to Section 3)

Q2. Does your district encourage teachers to provide time for students to prep or practice for any of the **commercially developed tests**?

- ○ Yes
- ○ No

Q3. From a district perspective, please rate how useful results are from **standardized, commercially developed tests** for decisions about educational programs or instruction, using a scale of 1 (*not at all useful*) to 6 (*extremely useful*).

	1 - Not at all useful	2	3	4	5	6 - Extremely useful
Informing district-level decisions	○	○	○	○	○	○
Informing school-level decisions	○	○	○	○	○	○
Informing teachers' instructional decisions in the classroom	○	○	○	○	○	○

Q4. Comments (optional) about commercially developed tests:

Part 3: District Developed Tests

Q1. Please indicate which **standardized, district-developed tests or assessments** are administered to a majority of your students district-wide for one or more grade levels. Do not include course-based assessments.

- ☐ Writing prompt, specify grades _____
- ☐ Other, please list test name and specify grades _____
- ☐ Other, please list test name and specify grades _____
- ☐ Other, please list test name and specify grades _____
- ☐ Other, please list test name and specify grades _____
- ☐ Other, please list test name and specify grades _____
- ☐ My district DOES NOT administer any standardized, district-developed tests or assessments

Skip To: Demographics If Q1 = My district does not administer any standardized, district-developed tests or assessments

Q2. Does your district encourage teachers to provide time for students to prep or practice for any of the **district developed** tests?

- ○ Yes
- ○ No

Q3. From a district perspective, please rate how useful the test results from **district developed tests** are for decisions about educational programs or instruction, using a scale of 1 (not at all useful) to 6 (extremely useful)

	1 - Not at all useful	2	3	4	5	6 - Extremely useful
Informing district-level decisions	○	○	○	○	○	○
Informing school-level decisions	○	○	○	○	○	○
Informing teachers' instructional decisions in the classroom	○	○	○	○	○	○

Q4. Comments (optional) about district developed tests:

Demographics

Q1. Please indicate which enrollment band your district falls into for total district enrollment.

- ○ 100 or less
- ○ 101 - 199
- ○ 200 - 585
- ○ 586 - 1699
- ○ 1700 or greater

Q2. Does your district include secondary grades?

- ○ Yes
- ○ No

Q3. District county location:

- ○ Androscoggin
- ○ Aroostook
- ○ Cumberland
- ○ Franklin
- ○ Hancock
- ○ Kennebec
- ○ Knox
- ○ Lincoln
- ○ Oxford
- ○ Penobscot
- ○ Piscataquis
- ○ Sagadahoc
- ○ Somerset
- ○ Waldo
- ○ Washington
- ○ York

Thank you for completing this survey!

About the Authors

Janet C. Fairman (janet.fairman@maine.edu) is an Associate Professor in the College of Education and Human Development, University of Maine, and co-Director of MEPRI. Dr. Fairman holds a doctorate degree in education policy and has expertise in the areas of education policy analysis, program evaluation, and qualitative research methodology. Her research includes a focus on STEM education, innovative and reform practices in education, and teacher leadership.

Amy F. Johnson (amyj@maine.edu) is Co-Director of the Maine Education Policy Research Institute at the University of Southern Maine. Her areas of interest include equitable school funding models, teacher preparation program accountability, STEM education, and college readiness.

Sharon LaBrie, Research Associate in the College of Education and Human Development, University of Maine, assisted with the district survey development and administration. Ms. LaBrie has a master's degree in human development and has expertise in project management, program evaluation, survey administration, and data management and analysis.

Ian M. Mette is an Assistant Professor in Educational Leadership in the College of Education and Human Development at the University of Maine. His research interests include school reform policy, teacher supervision and evaluation, and bridging the gap between research and practice to inform and support school improvement efforts. Specifically, his work targets how educators, researchers, and policymakers can better inform one other to drive school improvement and reform policy.

Garry Wickerd is an Assistant Professor of Educational Psychology and School Psychology at the University of Southern Maine. Dr. Wickerd started his career as a public school teacher teaching Latin in Florida and Georgia for six years before returning to school for a Ph.D. in school psychology from the University of South Dakota. He has research interests in the measurement of behavior, multicultural/bilingual school psychology, social skills interventions for individuals with autism, the history of psychology, and adaptive behavior assessment.