

Assessment Task Force Meeting Materials

July 13, 2015

9:00 a.m. via WebEx

In this meeting, we will be focused on design issues in summative assessment to prepare for the in-person work in Laramie later this month. In preparation for the webinar, please find attached three documents for pre-reading. For the last meeting, we asked you to do some background reading to prepare you for the discussion on summative assessment, but those readings featured only somewhat in the discussion. This time, our presentation will be focused closely on the three readings below, and will expand on them.

1. Intuitive Test Theory, page 2

This article compares understanding of educational assessment to understanding of physics. It starts with an explanation of intuitive versus scientific physics, showing how most people employ an intuitive theory of physics that generally explains everyday phenomena, while experts employ a very different theory to understand complex phenomena. Test theory is similar. Intuitive test theory works in general for everyday educational situations, but the more complex and the more high stakes testing becomes, that intuitive theory breaks down with considerable consequences.

2. Assessment Triangle, page 11

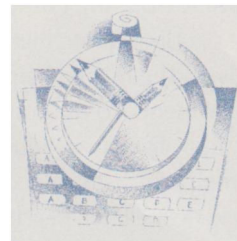
This excerpt from Pellegrino's addresses a (slightly exaggerated) need to move away from 19th century psychology in the way we design, implement, and interpret the results of assessment in terms of effectively tying together how we think about student learning, how we make observations to support conclusions about student learning, and how we interpret those observations to make those conclusions.

3. Michigan Common Core Assessment Options Report, page 16

Last time we asked you to review this document to become familiar with potential formats for evaluating existing assessments and to become familiar with some of the available options. This time, we would like you to reread just the text of the reports (don't worry about the results) with the goal of becoming familiar with the kind of design considerations that are important. This is not a complete list of important design considerations, but it will give you an introduction to why design considerations are important.

Intuitive Test Theory

Many of us have an intuitive understanding of physics that works surprisingly well to guide everyday action, but we would not attempt to send a rocket to the moon with it. Unfortunately, Mr. Braun and Mr. Mislevy argue, our policy makers are not as cautious when it comes to basing our school accountability system on intuitive test theory.



BY HENRY I. BRAUN AND ROBERT MISLEVY

A LONG WITH making sure that our bodily needs are met, one of our first tasks upon entering this world is to try to make sense of it. We do so by continuous observation and generalization, as well as by absorbing the norms of the culture in which we find ourselves. Our understandings typically take the form of stories — narratives, as the psychologist Jerome Bruner has called them. These stories are attempts to identify why people do what they do — their beliefs, motives, and plans.

This mode of developing and retaining understanding carries over to the physical world, whether natural or human-made. We hear thunder and see lightning, see objects being thrown and falling to the ground, observe cars and computers working (or not), and we construct stories

HENRY I. BRAUN is Distinguished Presidential Appointee at the Educational Testing Service (ETS), Princeton, N.J. ROBERT MISLEVY is a professor in the Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park. They wish to acknowledge Neal Dorans, Paul Holland, and Howard Wainer for stimulating conversations on the topic of this article during its preparation. The research reported here was underwritten by ETS and grants from the Office of Educational Research and Improvement, U.S. Department of Education (No. R305B60002), and from the National Center for Research on Evaluation, Standards, and Student Testing, UCLA. However, the opinions expressed are solely those of the authors.

about causes, patterns, and linkages. Now, we make up these stories whether or not we truly understand what is going on. Adults are driven, in exactly the same way as 5-year-olds are, to express their understanding of what is happening around them in terms of narratives.

As Howard Gardner has pointed out, stories can differ, often substantially.

In most domains of knowledge, we develop very powerful theories when we are very young. . . . No one has to tell a kid that heavy objects fall more quickly than light objects. It's totally intuitive. It happens to be wrong. Galileo showed that it was wrong. Newton explained why it was wrong. But, like others with a robust 5-year-old mind, I still believe heavier objects fall more quickly than lighter objects.

The only people on whom these engravings change are experts. Experts are people who actually think about the world in more sophisticated and different kinds of ways. . . . In your area of expertise, you don't think about what you do as you would when you were five years of age. But I venture to say that if I get to questioning you about something that you are not an expert in, the answers you give will be the answers you would have given before you had gone to school.¹

Richard Feynman's story for what happens when we throw a rock might be based on the principle of the path

of least action and admit to a rigorous rendering in differential calculus, whereas little Jimmy's story is that the rock wanted to get back down to the ground where it belongs. The point is that people construct plausible stories for actions and events based on what they've experienced themselves and on what they've picked up, however loosely or informally, from the culture around them.

The Gardner quote highlights two other aspects of these narratives. The first is their tendency to persist, even in the face of evidence to the contrary or confrontation with methods of analysis that are much more powerful. Bruner makes the same point with respect to what he calls "folk psychology." He defines folk psychology as a system by which people organize their experience in, knowledge about, and transactions with the social world. We learn our culture's folk psychology along with its language and norms of social behavior. Bruner asserts, "Folk psychology changes but is not displaced by scientific psychology."² It is the persistence of these narratives (say, in physics) that can be so frustrating for teachers.

The second aspect of these stories is that expertise is often very narrowly focused. That is, outside one's area of specialized training, it is uncommon to do much better than a 5-year-old. Indeed, the situation may be even more dire. In a now classic study, the psychologists Amos Tversky and Daniel Kahneman questioned a large number of research psychologists on various aspects of probability and statistics (the design of experiments and the interpretation of the results) that would ordinarily be relevant to their work. Surprisingly, a majority of the respondents harbored naive (and incorrect) beliefs that, presumably, influenced how they conducted their research.³

What is true of psychology or physics is true of just about every discipline you can think of. It is also true, we will argue, in educational assessment. Before we begin to explore this, our own field, we will examine briefly how people who are not experts in physics think about physical phenomena. This "intuitive physics" is a set of basic premises about how the world works. It consists of story elements or subplots, as it were, called phenomenological primitives (or p-prims, for short), a term coined by psychologist Andrea diSessa. These p-prims are primitive notions in the sense that they "stand without significant explanatory substructure or explanation."⁴ And just as the idea of p-prims can help explain most people's understanding of the physical world, so too can p-prims help us explain the "intuitive test theory" that nonexperts use to explain the world of assessment.

Perhaps it is not surprising that such p-prims — and the narratives in which they are embedded — work well enough for most situations in our everyday lives. After all, they are

grounded in the experiences of many people over many, many years. They can lead to trouble, though, when employed in situations that lie outside their range, in which case expert models are indispensable. Unfortunately, unlike prescription drugs, p-prims (in physics or other disciplines) are usually not accompanied by warning labels with contraindications for use. In a fast-changing world, it is increasingly likely that we will find ourselves relying on p-prims that are not up to the task.

INTUITIVE PHYSICS

One consequence of the "cognitive revolution" in psychology that began in the 1960s was a closer look at how people develop expertise in real-life activities as varied as radiology, writing, chess, and volleyball. A significant finding across domains is that experts don't simply know more facts than novices — although they usually do — but that they also organize what they know around deeper principles and relationships. The knowledge novices have is more fragmented and is related to particular situations or organized around surface features of problems.

For example, Micki Chi, Paul Feltovich, and Robert Glaser asked expert physicists and novices to sort a number of problems into groups. The novices produced piles of spring problems, pulley problems, and inclined-plane problems. The experts produced piles associated with equilibrium, Newton's third law, and the conservation of energy, each containing some spring problems, some pulley problems, and some inclined-plane problems. The experts' categorization leads directly to solution strategies for the problems.⁵

When diSessa introduced the term "p-prims" in 1983, it was expressly to explain nonexperts' ways of reasoning about physics. Familiar examples of such p-prims are "Heavy objects fall faster than light objects," "Things bounce because they are 'springy,'" and "Continuing force is needed for continuing motion." These physical p-prims are based on our everyday experience. A box moves when we push it, and it stops moving when we stop pushing. Cannon balls really do fall faster than feathers. Physicists know this, of course, but, when necessary, they can appeal to a deeper level of explanation, to the more sophisticated primitives of scientific physics. The distinguishing feature of intuitive physics (or intuitive reasoning in any field) is that the p-prims are the bottom line. For nonexperts, they are the final explanation. In other words, sometimes we just have to say, "Well, that's just the way it is."

Some of the p-prims of intuitive physics use such words as force, energy, and momentum, a legacy of the general culture or of a physics class taken long ago. But the terms

are not employed in the same way that experts use them. Nonexperts don't sort concepts in the same ways as experts or embed them in the same web of qualitative and quantitative relationships. A set of p-prims is not a coherent system, and a person's set of p-prims can easily contain some that contradict others. They are employed to reason about physical situations, and a model of sorts is assembled to address a given situation. The surface features of a situation tend to elicit some p-prims but not others, so a person's intuitive models can be quite different for two situations that are formally equivalent.

The surprising thing is how well they work for guiding everyday action. You can think you are imparting a substance called "impetus" to the tennis ball when you throw it for your dog. The ball flies until the impetus wears off. You estimate how much of this substance you want to impart to the ball and gauge your throw accordingly — and, by golly, the ball goes where you want it to. Your impetus theory is wrong, but neither you, nor the dog, nor the ball knows this, and the job gets done just fine.

Intuitive physics works well enough for playing catch with your dog or for building a birdhouse. But it doesn't work for constructing a bridge or shooting a rocket to the moon. One aspect of becoming an expert in physics is learning more sophisticated ways of thinking, but another is knowing when you need to use them, and yet another is recognizing when they fail. (Science is also about telling stories, but they are stories that submit to reality checks.) In scientific physics, concepts and relationships that may be non-intuitive, or even counterintuitive, can be brought to bear on familiar and unfamiliar situations alike. Individuals facing challenges that lie outside everyday experience ignore scientific physics at their peril.

SCIENTIFIC TEST THEORY

To Americans who go to school or hold jobs in the 21st century, taking tests is an experience nearly as familiar as pushing boxes or watching things fall. So we need to tell stories about tests — their purposes, their construction, our performances on them — and we need concepts to do so. Below, we will briefly sketch how experts in assessment think about these aspects of tests. But unless you are an expert in assessment, it is probably not the way you think about them. Indeed, some of the ideas may be quite foreign to you.

A scientific approach to assessment recognizes that, fundamentally, assessment isn't about items and scores. These are more like the springs and pulleys of testing. Rather, assessment is a special kind of evidentiary argument. Assess-

ment is about reasoning from a handful of particular things students say, do, or make, to more broadly cast inferences about what they know, have accomplished, or are apt to do in the future.⁶

The starting point for an application of scientific test theory is a clear understanding of the purpose of the assessment and a perspective on the nature of the knowledge or skills that are the focus of attention. Next is the link between this view of knowledge and skills, which you can't see, to things that you can see — right and wrong answers, problem-solving steps, justifications for building designs, or comparisons of characters in two novels in terms of transaction theory, to cite just a few examples. This analysis resolves into making a case, in light of the purpose of the assessment, for what is meaningful in a student's performance and why. A rationale is also required for the kinds of assignments or challenges that will elicit the evidence to support the intended inferences about students. Conceptual links connect tasks to student performances to judgments about what they know and can do. These are the testing counterparts of Newton's laws.

Now, Newton's laws of motion are deterministic. That is, given a complete description of an object (e.g., its mass, current position, and velocity), we can calculate exactly the effect on its motion of an application of a particular force. In test theory, we can formulate a student model that describes one or more aspects of a student's knowledge or skills. Since the components of the student model cannot be observed directly, we have to use probability theory to express our beliefs about the likely values of these components. As we accumulate more data about the student, we can employ the calculus of probabilities to update our beliefs.

The use of probability-based models to describe what we know, and what we don't know, about a student is a key tool in scientific assessment. It provides a quantitative basis for planning test configurations, calculating the accuracy and reliability of the measurement process, figuring out how many tasks or raters we need to be sufficiently sure about the appropriateness of decisions based on test scores, or monitoring the quality of large-scale assessment systems. We can also apply the tools of probability to new kinds of testing processes, such as ones that select discrete tasks to present to individual students in light of how well they are doing or their instructional backgrounds, or computer-based tests of problem solving in which the problem itself evolves in response to the student's actions. These probability models and their essential role in reasoning are all but unknown to the nonexpert.

It is worth pointing out that the use of probability models

to manage information doesn't restrict the kinds of knowledge and skills we can model. While psychometrics arose around 1900 with the goal of measuring traits such as intelligence, the same modeling approach can be applied with all kinds of psychological perspectives and all kinds of data. The variables in the student model can be many or few; they can be measures or categories; they can concern knowledge, procedures, strategies, or attunement to social situations; they can be as coarse as "verbal reasoning" or as fine-grained as "being able to describe playground situations in terms of Newton's laws."

What is observed and how it is modeled and evaluated will depend partly on a psychological perspective and partly on the job at hand. Designing an assessment is like building a bridge. The evidentiary arguments and the probability models are like Newton's laws in that you have to get them right or the entire structure will collapse. But they aren't sufficient to determine the project. In architecture and engineering, decisions about location, materials, and various features of the design are strongly influenced by the resources available, by the situational constraints, and by the needs of the clients. Similar processes are at work in measurement.⁷

The typical classroom teacher brings to bear little if any of this machinery in constructing, analyzing, and drawing inferences from Friday's math quiz. Usually, this is perfectly fine and appropriate to the purpose and the context. Assessment practices have evolved into familiar forms of testing that often work well enough in common situations. The principles that account for why they work in the situations

for which they evolved are there — invisible but built into the pieces that we can see. Popular conceptions of how and why familiar tests work hold the same ontological status as impetus theory — dead wrong in the main, but close enough to guide everyday work in familiar settings. It is when we move beyond the familiar that these notions can betray us.

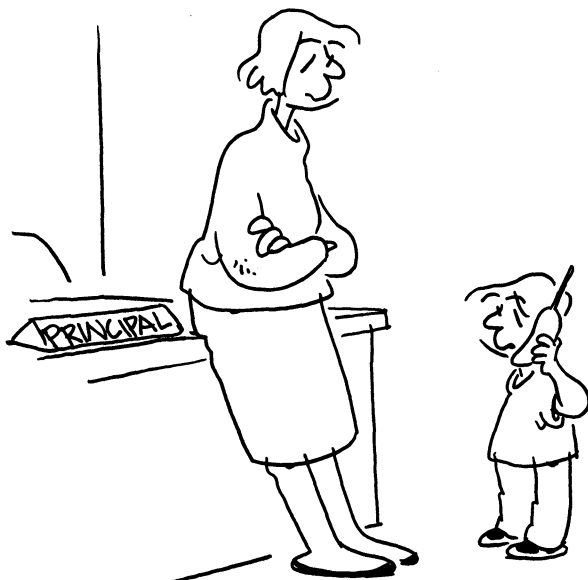
P-PRIMS UNDER SCRUTINY

Let us now consider a number of p-prims of test theory. Just as in intuitive physics, these are the underpinnings of the view of testing held by many nonexperts. Our goal is to use the insights of scientific test theory to begin to understand how these beliefs might have arisen and in which situations they can break down. In what follows, we sometimes use the phrase "drop-from-the-sky" to describe a test — by which we mean a test that is developed outside the school context. The term is meant to connote the remoteness of the test from the day-to-day experiences of the students.

A test measures what it says at the top of the page. It is natural to assume that a name carries meaning. Thus we expect that a test called a history test will measure a student's accomplishments or proficiency in history. However, a student's score on such a test can be determined less by how well a student can analyze or interpret historical materials than by a host of other factors that also influence performance and on which individuals can differ substantially. Such factors include, for example, a student's familiarity with the testing situation, the kind of test and mode of administration, and even what the grader of the test is looking for.

A common manifestation of this p-prim is making inferences from test scores that extend well beyond what can be reasonably supported. Perhaps the most notorious example is the overinterpretation of the results of standardized intelligence tests. Performance on a particular drop-from-the-sky intelligence test does typically indicate a capability to do productive reasoning in certain circumstances. But there are many kinds of intelligent behavior, some of which are predicted pretty well by scores on intelligence tests and others that are not.⁸ For example, people are good chess players not because they are intelligent in a general sense but because — through study, practice, and reflection on their performance in many, many games — they have learned a great deal about the patterns and successful strategies in the domain of chess.⁹

A test is a test is a test. This p-prim is a corollary of the preceding one. Some tests that are called fourth-grade mathematics tests, for example, focus more on concepts, others



Campbell.

"I can't talk now. I'm in a meeting."

focus on computations, and still others focus on using math in real-world situations. They reflect different aspects of what students know about and can do with math. Furthermore, a classroom teacher can build her quiz assuming that students are familiar with her notation, item types, and evaluation standards. This is more difficult for a drop-from-the-sky test. Moreover, assessments in the form of projects requiring extended work in math can be done over time as part of a program of instruction, but they aren't well suited for a drop-from-the-sky test that occurs on a single day.

Each assessment can be described in terms of the skills and knowledge it can tell you about, how much information it provides, its implications for learning, how closely it corresponds to students' background and instruction, and its demands on such resources as equipment, money, and student and teacher time. The trick is to match a test — with all its many characteristics — with the purpose of the testing and the context in which it will be used. Getting the proper match can be a delicate balancing act. For any number of reasons, the same test can be exactly right for one purpose and situation but quite useless for another. Good test developers know this, and they design different assessments for different purposes in light of the characteristics of the students, the available resources, and the constraints of the setting.

A particularly dangerous fallacy follows from this p-prim: you can take a drop-from-the-sky test constructed to gauge knowledge in a broad content area, give it to students about whom you know little else, and, by coming up with a different way of scoring it, obtain diagnostic information that will be useful for individual, small-scale instructional decisions. This generally doesn't work, and the problem isn't with the items or the scoring rules. It is that effective information about what to do next requires assessment that takes into account what a teacher already knows about a student and provides information in terms of instructional options — not necessarily better items or more items, just the right items for the right student at the right time. Good diagnostic information results from good match-ups, not from good one-size-fits-all tests.

A score is a score is a score. With all the criticism that testing attracts, it is remarkable how much credence is typically attached to a single test score. After all, the reasoning goes, how could there be a "truer" score than the score a student actually gets? This p-prim is reinforced by the familiar practice of making decisions on the basis of a single test score without considering what the scores might have been in hypothetical administrations of alternative measures. Measurement experts recognize that different data could have arisen from testing on other occasions; from using more, fewer, or different test items; or from employing more, fewer,

or different raters. (Perhaps the best way to bring home the concept of "noise" in test scores is to administer multiple tests and let people see for themselves the surprisingly large differences that result.)

Once we decide what we want to make inferences about from the data available, we can use scientific test theory to gauge how much evidence we have and compare it with what might have occurred in a variety of hypothetical alternative situations. This concept, roughly that of measurement error, is not a natural part of everyday reasoning about test scores (with the major exception that occurs when someone's score is lower than he or she expected). Assessment data are not perfect. Relying on a single score without regard to the uncertainty attached to it may be good enough for typical, low-stakes applications, but it is problematic for more consequential ones. Without scientific test theory, we could neither quantify that uncertainty nor evaluate the validity of the use of a particular test score in a particular setting.

Any two tests that measure the same thing can be made interchangeable with a little "equating" magic. This is intuitive test theory's equivalent of the perpetual motion machine. Why do people believe it? First, it seems to happen all the time. Almost everyone knows that large-scale testing programs like the SAT I and the Iowa Tests of Basic Skills (ITBS) regularly generate new test forms and that psychometricians routinely equate scores on the new forms to scores on the old ones. Second, it seems to make sense, because it follows from the preceding p-prim. If you think that tests measure what they say they measure and that all tests that measure it are essentially the same, and if you don't concern yourself with measurement error, then there is no apparent reason not to treat evidence from different tests as more or less equivalent.

But the strength of the correspondence between the evidence from one test and that from another, superficially similar, test is determined by the different aspects of knowledge and skills that the two tests tap, by the amount and quality of the information they provide, and by how well they each match the students' instructional experiences. The SAT I and ITBS testing programs can do this not so much because of the equating procedures they use but because they expend considerable effort in creating test forms with very similar combinations of questions (item types, content areas, mix of difficulties), in order to tap the same sets of skills in the same ways. When tests are not designed to be "parallel" in this way, quantifying in what ways information from one test can be used as if it came from another requires expert-level (scientific) test theory. Some inferences across tests will work well, and others will fail.

With legislation mandating the measurement of student progress and the establishment of common standards for achievement, policy makers have expressed considerable interest in linking tests from different states or different test publishers to the National Assessment of Educational Progress (NAEP). There is a long and definitive line of scientific publications pointing out the very real limitations of linking and equating different tests with the same name.¹⁰ Unfortunately, the notion that disparate tests can somehow be made equivalent by applying equating magic will not die, because life would be much easier if it were true. And by the reasoning of intuitive test theory, there is no reason why it can't be done.

You score a test by adding up scores for items. Almost all classroom quizzes and tests are graded in this way, and it works just fine for their purposes. Consequently, one can hardly be blamed for holding this p-prim. But it presumes that the target of inference is a student's overall proficiency in some domain and that the tasks on the test are relatively independent positive indicators of that proficiency. Indeed, this is the simplest (and most familiar) case of a relationship between targets of inference and bits of evidence about them. When interest focuses on dependencies among more complex forms of evidence and multifaceted models of knowledge and skill, however, this "natural" approach to scoring is severely deficient.

This approach fails for large integrated performances such as the videotaped lesson plans and teaching sessions of the National Board for Professional Teaching Standards, because multiple, interconnected judgments across many parts of the work are required. It fails for interactive problem-solving simulations (e.g., troubleshooting or patient management), because each action taken changes the situation and constrains or facilitates the next action. It fails for collections of tasks that tap a variety of skills and knowledge in different mixes, such as language tests that assess not only vocabulary and grammar but also how to conduct meaningful conversations, use cultural information, and accomplish real-world aims such as bargaining. Patterns of what is done well and where performance is inadequate are required, with the added complication that people trade off their strengths against their weaknesses when they use language in real life.

This approach also fails for assessments that aim to distinguish conceptions and misconceptions (as opposed to correctness). That is, it fails when the goal isn't to count how many problems a student can solve, but rather to develop a useful description of her thinking — so that we can better decide what she might work on next to improve her understanding.

In all of these cases, simple scoring rules don't make the "grade" because they extract only a part of the evidence contained in students' responses — sometimes completely missing the patterns that are most important — and therefore can't support the nuanced inferences that are desired. Scientific test theory, extended and elaborated as needed to deal with new kinds of data and new kinds of inferences about students, is the best foundation for both effectively designing these more complex assessments and for making sense of the data they produce.

An A is 93%, a B is 85%, a C is 78%, and 70% is passing. This p-prim follows from the previous one, with the additional assumption that the tasks that make up a test have been written so that these percentages line up nicely with the traditional percent-correct metric of satisfaction for how well students have done on tests of materials that were specifically matched to their instruction. It presumes that somehow, for all tests and all uses and all students, the same percentage of correct answers corresponds to the same level of performance.

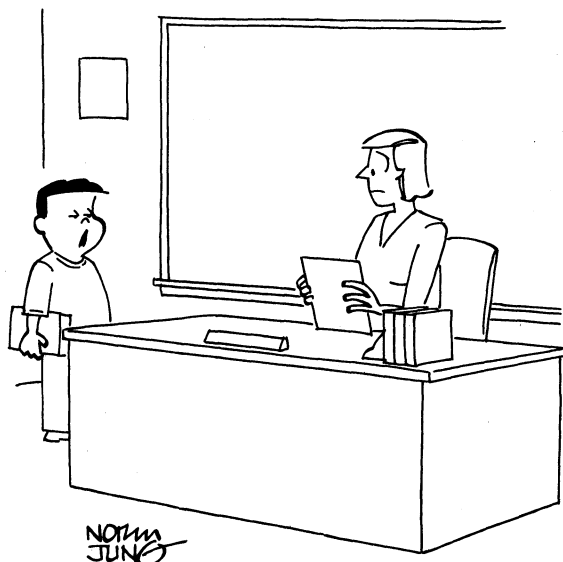
A colleague who works on certification and licensing tests tells the story of a state legislature that passed a law mandating that "the passing score on the plumber's licensing exam will be 70." Following good test-design practices, our colleague worked with plumbers to determine the kinds of knowledge and skills needed to be a competent plumber, one who is able to ply the craft ably and with due regard to safety. The committee then created a collection of tasks to probe the targeted knowledge and skills and pilot-tested them with groups of competent plumbers and with apprentices who were judged to be not yet ready to practice on their own. A passing score was selected that best differentiated the two groups. This is a sound foundation for creating a valid licensing assessment and setting a defensible level of performance for a high-stakes decision. When they got that number, it shouldn't have mattered what its numerical value was. Within the constraints of the testing program, it had been constructed to be a valid cut point for the purpose of obtaining a license. As a final step, however, the test developers had to add (or subtract) a "fudge factor" to make the passing score exactly 70.

This p-prim is plausible because for many of the tests we took in school, this grading scheme is not a bad choice. But this didn't happen by accident. Good teachers who wanted to use this grading scheme thought carefully about what they wanted students to learn and about the conditions under which students could exhibit that learning. They set up tasks and evaluated them to get data. Then they looked hard at the numbers. If the scores they saw from their students didn't jibe with their expectations, they went back to

the drawing board to figure out why. Were the items unreasonable or unclear? If so, then revise or replace them. Were the students just not learning what was intended? If so, then check whether the students have the background they need, verify that they are really working, improve the pedagogy, and so on.

The difficulties encountered in applying this p-prim and the previous one in more complex settings have led to advances in measurement theory. Indeed, it is possible to construct both easy and hard tests from the same collection of items, and the same level of knowledge will produce a higher score on the easy test than on the hard one. Psychometric models based on item response theory originated in the 1960s to characterize items in terms of their difficulty and other features, so that students can be given different sets of items and still be compared on the same scale — harder ones for fifth-graders and easier ones for third-graders, for example, or computer-administered tests that are customized to each examinee on the basis of his or her performance as it unfolds.¹¹ So what now is an A, a B, or a C? You can't decide just by calculating the percentage of correct answers; you should decide on the basis of the pattern of correct and incorrect answers, taking into account the relative difficulty of the items presented.

Under some circumstances, the results may be reasonably well approximated by a simple sum. But the underlying principles provide a deeper understanding of why the standard procedures work in familiar situations, as well as the machinery for creating new procedures for novel situations — very different arrangements of springs and pulleys, but undergirded by the same Newtonian laws.



"Of course you've seen that essay done before. I'm repeating the fifth grade."

Multiple-choice questions measure only recall. This p-prim is often stated as an epithet, as part of a comparison to open-ended questions. Certainly most of the multiple-choice questions that people encounter in school test only recall, and it is surely true for multiple-choice questions written by someone who believes the p-prim. But while factual recall items may be the easiest kinds of multiple-choice items to write, other types are certainly possible. For example, a multiple-choice test of subtraction can be written so that patterns of right answers and wrong answers will reflect particular misconceptions and tell us more about a student's understanding than would overall performance on a test made up of only open-ended items.

Similarly, research in physics education sparked by work like diSessa's has led to the development of multiple-choice tests that reveal which p-prim students are using. Rather than the usual open-ended computation and modeling items, the items on the Force and Motion Conceptual Evaluation present descriptions of everyday situations and ask students to choose explanations of what is happening or predict what will happen next.¹² Some alternatives reflect Newton's laws, but others reflect p-prim that are more consistent with Galileo's thinking, medieval impetus theory, Aristotle's beliefs, or wholly nonscientific reasoning. The situations vary in ways that research suggests will bring particular p-prim to light.

For example, Newton's third law says that for every action (or force) there is an equal and opposite reaction. If object number 1 exerts a force on object number 2, then object number 2 exerts an equal and opposite force on object number 1. When a car and a small truck of the same weight moving at the same speed collide head-on, most students chose the response that says, "The truck exerts the same amount of force on the car as the car exerts on the truck." That's okay so far, but this is a canonical example for the third law — easy to give the answer Newton would without understanding the underlying principle. When the small pickup truck is replaced with a huge semi traveling only half as fast, more students choose "The truck exerts a larger force on the car" because the truck is larger. Or they choose "The car exerts a larger force on the truck" because the car is going faster. These responses reflect alternative — and in this case, conflicting — p-prim.

In and of itself, the format of a task — be it multiple-choice, open-ended, simulation-based, or hands-on performance — doesn't fully determine the kind of thinking it will elicit from a student. What's more, the same task can give rise to different kinds of thinking in different students, depending on how it fits with their background and experiences. To a high school algebra student, figuring out

the sum of the numbers from 1 to 100 is a simple application of a familiar formula. But rather different cognitive processes were at play when the 7-year-old Karl Friedrich Gauss derived the formula as an original insight.

Multiple-choice items can be used to test recall of facts, and most of them are used in this way. But if one has clearly in mind the concepts and relationships one wants to probe, as well as the kinds of discriminations that an understanding of them entails, then it is possible to write multiple-choice items that go far beyond recall. The principles for creating such items aren't obvious and, unfortunately, aren't a part of most people's theory of tests.

You can tell if an item is good by looking at it. Like most of the others, this p-prim rests on the assumption that items and tests are really simple objects whose essence can be grasped by their surface characteristics. However, for an item to serve a given purpose, there has to be a reasonable coherence between its particular purpose, what the item provides and what it requires, the student's understanding of the context of the item and the scoring rules, and what else the assessor knows about what the student knows. A bad mismatch at any point, and the item may fail to generate the evidence needed, no matter how "good" it looks.

For example, consider an open-ended item devised by a teacher for her Advanced Placement calculus class that uses her notation, will be scored with the rubric her students have become used to, and calls for applying what they've been studying for the last month to a real-world situation that is similar to one discussed in class. This is an ideal probe to elicit their understanding of an important learning objective. However, it would be a poor item to include in the grade-12 NAEP, which presents tasks to a random sample of students across the country — many of whom would not be familiar with the notation or the grading rubric. Ten minutes of valuable testing time would be wasted for almost everyone who confronted the question. (The converse of this p-prim is more nearly true: You can often tell an item is bad just by looking at it. Logical flaws and confusing instructions, for example, will keep an item from providing useful information for almost any purpose.)

That the appropriateness of an item depends on "more than meets the eye" implies that writing good items is more difficult than most people would imagine. In addition to having a coherent conceptual framework and a strong evidentiary perspective, item writers must also work under constraints of time and money as they build tasks and assemble tests. It is not a vocation for the faint of heart or the novice, as recent missteps in many high-stakes state tests attest. Ironically, the more one knows about writing test items, the more challenging it is to write good ones.



Multiple-choice tests equal standardized tests equal high-stakes tests. Many of the highly visible tests used today for college admissions, for licensure and certification, and for state accountability for public schools are alike in three important ways: they have meaningful consequences for students or schools, they are presented under standard conditions, and they use multiple-choice items. This configuration occurs often enough that these three distinct properties are conflated in the public eye so that the adjectives "multiple-choice," "standardized," and "high-stakes" are thought to be synonymous — all ways of describing the same familiar package.

But high-stakes tests can be less standardized and require performances, as is the case with doctoral dissertations and solo flights for pilot certification. Multiple-choice items are found as often in low-stakes classroom quizzes as they are in high-stakes assessments. Finally, standardization is not an all-or-nothing quality. For each aspect of an assessment, there are options about how similar to make the experience for different examinees. And, as always, seeking to standardize involves tradeoffs. Greater similarity across examinees in some facets tends to support comparisons and facilitate communication of results across time and distance. More individualization allows the tests to be better targeted to individuals' circumstances, although the interpretation of results is more tightly bound to those circumstances.

DISCUSSION

While intuitive test theory is sufficient for classroom testing and for the quizzes in *Seventeen* magazine, it gets you

into trouble when you want to evaluate performance on simulation-based activities, run a high-stakes testing program, or measure change in populations using an achievement survey like NAEP. There is a strong similarity — and an important difference — between intuitive physics and intuitive test theory that has implications for assessment use and policy. As one's understanding and expertise in physics become more profound, the concepts and tools depart from everyday physics. The same is true with assessment design and analysis at the frontiers.

It is generally accepted that this is the case in physics and, moreover, that the complexity must be confronted if one is embarking on a serious undertaking. Consider the paradigmatic example of launching a rocket to the moon. In fact, in 1961, when President Kennedy made his famous promise that by the end of that decade the U.S. would send a man to the moon and return him safely back to Earth, his staff had already consulted with experts about the feasibility of such an endeavor. Two points are noteworthy. First, everyone expected that all the options that would be considered would be in accord with Newton's laws of motion, not Aristotle's. Second, President Kennedy did not assert that, on its flight to the moon, the rocket would have to meet specific milestones that he and his advisors deemed appropriate.

In most issues that involve technical considerations, experts are consulted, and their perspectives become part of the policy debate. They don't make the decisions, and they shouldn't. In any social setting, there are more considerations than purely technical ones. But policy options should be restricted to those that are in accord with basic principles and broadly held standards of practice — the analogs of Newton's laws of motion.

Unfortunately, this is often not the case in assessment, as a review of the testing policies in many states and the legislative history of the No Child Left Behind Act demonstrate. As assessment-based accountability becomes a more prominent feature of education policy, those standing on the technical side of assessment must confront the reality that critical decisions are made and regulations are drafted on the basis of intuitive test theory, with untoward consequences a likely result. The advent of technology-based assessment may, in many ways, exacerbate the problem. No doubt voluminous data will be produced, but insight will still be in short supply. In fact, a disciplined application of the principles of evidentiary reasoning to design, development, and analysis will be all the more necessary if the investment in technology is to yield meaningful returns.

We remain, then, with the problem that p-prims are both widely held and persistent. What, then, should those of us

in educational measurement do? There are at least three lines of attack, one negative and two positive. First, we should not shy away from critiquing policies and programs that are based on intuitive test theory. This involves telling lots of people (some of them very important) that what they want to do won't work and that doing something right is harder or takes longer than they might like.

A second approach is to use scientific test theory, in conjunction with developments in psychology and technology, to achieve goals that could not have been accomplished otherwise — certainly not by relying on intuitive test theory. These existence proofs are the most compelling argument for test theory as a scientific discipline and for its utility in the setting of education policy.

Finally, we need to do a much better job of communicating to a variety of audiences the basics of testing and the dangers we court when we ignore the principles and methods of educational measurement. Communication is a form of teaching, and we should take the challenge of this kind of teaching more seriously than ever before. Perhaps we should consider using narratives as a framework for this effort. We have an obligation to be as creative in this effort as we pride ourselves on being in our technical research.

1. Howard Gardner, *Educating the Unschooled Mind* (Washington, D.C.: Federation of Behavioral, Psychological, and Cognitive Sciences, 1993), p. 5.

2. Jerome Bruner, *Acts of Meaning* (Cambridge, Mass.: Harvard University Press, 1990), p. 14.

3. Amos Tversky and Daniel Kahneman, "Belief in the Law of Small Numbers," *Psychological Bulletin*, vol. 76, 1971, pp. 105-10.

4. Andrea diSessa, "Phenomenology and the Evolution of Intuition," in Dedre Gentner and Albert L. Stevens, eds., *Mental Models* (Hillsdale, N.J.: Erlbaum, 1983), p. 15.

5. Micki T. H. Chi, Paul Feltovich, and Robert Glaser, "Categorization and Representation of Physics Problems by Experts and Novices," *Cognitive Science*, vol. 5, 1981, pp. 121-52.

6. Robert J. Misyevy, "Substance and Structure in Assessment Arguments," *Law, Probability, and Risk*, December 2003, pp. 237-58.


7. Henry I. Braun, "A Postmodern View of the Problem of Language Assessment," in Antony J. Kunnan, ed., *Studies in Language 9: Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium* (Cambridge: Cambridge University Press, 2000), pp. 263-72.

8. Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences* (New York: Basic Books, 1983); and Robert J. Sternberg, *The Triarchic Mind: A New Theory of Human Intelligence* (New York: Viking-Penguin, 1988).

9. Adrianus de Groot, *Thought and Choice in Chess* (The Hague: Mouton, 1965).

10. See, for example, Michael J. Feuer et al., eds., *Uncommon Measures: Equivalence and Linkage Among Educational Tests* (Washington, D.C.: National Academies Press, 1999).

11. Howard Wainer et al., *Computerized Adaptive Testing: A Primer*, 2nd ed. (Hillsdale, N.J.: Erlbaum, 2000).

12. Ronald K. Thornton and David R. Sokoloff, "Assessing Student Learning of Newton's Laws: The Force and Motion Conceptual Evaluation," *American Journal of Physics*, vol. 66, 1998, pp. 228-351. 

Excerpted from Pellegrino & Chudowsky (2003). *The Foundations of Assessment. Measurement: Interdisciplinary Research and Perspectives*, 1(2), 103-148.

The Need to Rethink the Foundations of Assessment

In this paper we address educational assessments used for three broad purposes: to assist learning (also referred to as *formative* assessment), to measure individual attainment (also referred to as *summative* assessment), and to evaluate programs. Every assessment, whether used in the classroom or large-scale context, is based on a set of scientific principles and philosophical assumptions, or *foundations* as they are termed here. The central problem addressed in this paper is that most widely used assessments of school achievement are based on highly restrictive beliefs not fully in keeping with current scientific understanding about human cognition and learning, and how they can be measured.

Impact of Prior Theories of Learning and Measurement

Current assessment practices are the cumulative product of theories of learning and models of measurement that were developed to fulfill the social and educational needs of a different time. As Mislevy (1993, p. 19) has noted, "It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology." Although the core concepts of prior theories and models are still useful for certain purposes, they need to be augmented or supplanted to deal with newer assessment needs.

Some aspects of current assessment systems are still linked to earlier trait theories of learning that assumed individuals have basically fixed dispositions to behave in certain ways across diverse situations. According to such a view, school achievement is perceived as a set of general proficiencies (e.g., mathematics ability) that remain relatively stable over situations and time. Current assessments are also derived from early theories that characterize learning as a step-by-step accumulation of facts, procedures, definitions, and other discrete bits of knowledge and skill. Thus, assessments tend to include items of factual and procedural knowledge that are relatively circumscribed in content and format and can be responded to in a short amount of time. These test items are typically treated as independent, discrete entities sampled from a larger universe of equally good questions. It is further assumed that these independent items can be added together in various ways to produce overall scores.

Assessment Based on Contemporary Foundations

Several decades of research in the cognitive sciences has advanced the knowledge base about how children develop understanding, how people reason and build structures of knowledge, which thinking processes are associated with competent performance, and how knowledge is shaped by social context (NRC, 1999c). These findings, summarized in Part II, suggest directions for revamping assessment to provide better information about students' levels of understanding, their thinking strategies, and the nature of their misunderstandings. During this same

period, there have been significant developments in measurement methods and theory. A wide array of statistical measurement methods are currently available to support the kinds of inferences that cognitive research suggests are important to assess when measuring student achievement; these are also presented in Part II.

In this paper we describe some initial and promising attempts to capitalize on these advances (a much more extensive presentation of examples is provided in the full NRC report). However, these efforts have been limited in scale and have not yet coalesced around a set of guiding principles. In addition to discerning those principles, more research and development is needed to move the most promising ideas and prototypes into the varied and unpredictable learning environments found in diverse classrooms embedded within complex educational systems.

In pursuing new forms of assessment, it is important to remember that assessment functions within a larger system of curriculum, instruction, and assessment. Radically changing one of these elements and not the others runs the risk of producing an incoherent system. All of the elements and how they interrelate must be considered together. Moreover, while new forms of assessment could address some of the limitations described above and give teachers, administrators, and policy makers tools to help them improve schooling, it is important to acknowledge that tests, by themselves, do not improve teaching and learning, regardless of how effective they are at providing information about student competencies.

Issues of fairness and equity must be also central concerns in any efforts to develop new forms of assessment. To improve the fairness of assessment, it must be recognized that cultural practices equip students differently to participate in the discourse structures that are often unique to testing contexts. It is all too easy to conclude that some cultural groups are deficient in academic competence, when the differences can instead be attributable to cultural differences in the ways that students interpret the meaning, information demands, and activity of taking tests (e.g., Steele, 1997). These sorts of differences need to be studied and taken into account when designing and interpreting the results of assessments. If well-designed and used, new models of assessment could not only measure student achievement more fairly, but also promote more equitable opportunity to learn by earlier identification of individual students' learning needs.

The Assessment Triangle

The committee developed a framework for thinking about the foundations of assessment, referred to as the *assessment triangle*, which is based on the idea of assessment as a process of reasoning from evidence (Mislevy, 1996). The assessment triangle is useful for analyzing current assessments or designing new ones.

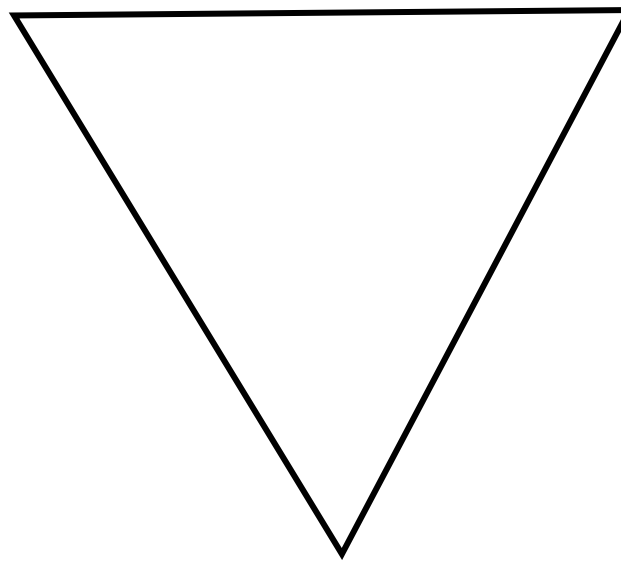
Every assessment, regardless of its purpose or the context in which it is used, rests on three pillars: 1) a model of how students represent knowledge and develop competence in the subject domain, 2) tasks or situations that allow one to observe students' performance, and 3) interpretation methods for drawing inferences from the performance evidence thus obtained. These three foundational elements—

cognition, observation, and interpretation—influence all aspects of an assessment’s design and use, including content, format, scoring, reporting, and use of the results. Even though these elements are sometimes more implicit than explicit, they are still influential. In fact, it is often the tacit nature of the foundations and the failure to question basic assumptions about one or more of the three elements and their interconnection that creates conflicts about the meaning and value of assessment results.

The three elements, each described further below, are represented as corners of a triangle because each is connected to and dependent on the other two (see Figure 1). A central tenet of this report is that for an assessment to be effective, the three elements must be in synchrony.

Observation

Interpretation



Cognition

Cognition

The *cognition* corner of the triangle refers to a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain. The theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain. These findings should derive from cognitive and educational research about how people learn, as well as the experience of expert teachers. As scientific understanding of learning evolves, the cognitive underpinnings of assessment should change accordingly. Our use of the term “cognition” is not meant to imply that the theory must necessarily come from a single cognitive research perspective. As discussed later, theories of student learning and understanding can take different forms and encompass several levels and types of knowledge representation that include social and contextual components.

It would be unrealistic to expect that assessment design will take into account every subtlety and complexity about learning in a domain that has been uncovered by research. Instead, what is being proposed is that assessment design be based on a representation or approximation of cognition that is consistent with a richer psychological perspective, at a level of detail that is sufficient to get the job of assessment done. Any model of learning underlying an assessment will necessarily be a simplification of what is going on in the head of the examinee and in the social situation within which the assessment takes place.

Observation

The *observation* corner of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. The observation model describes the stimuli presented to examinees and the products, such as written or oral responses, or the answers students have to choose among for multiple choice items. In assessment, one has the opportunity to structure some small corner of the world to make observations. The assessment designer can use this capability to maximize the value of the data collected, as seen through the lens of the underlying beliefs about how students learn in the domain.

The tasks selected for observation should be developed with the purpose of the assessment in mind. The same rich and demanding performance task that provides invaluable information to a teacher about his tenth grade class—because he knows they have been studying transmission genetics for the past six weeks—could prove impenetrable and worthless for assessing the knowledge of the vast majority of students across the nation.

Interpretation

Finally, every assessment is based on certain assumptions and models for interpreting the evidence collected from observations. The *interpretation* corner of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. It includes the rules used for scoring or evaluating students' responses. In the context of large-scale assessment, the interpretation method also usually includes a statistical model, which is a characterization or summarization of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher, and is usually based on an intuitive or qualitative model rather than a formal statistical one.

Connections among the vertices

To have an effective assessment, all three vertices of the triangle must work together in synchrony. For instance, a cognitive theory about how people develop competence in a domain provides clues about the types of situations that will elicit evidence about that competence. It also provides clues about the types of interpretation methods that are appropriate for transforming the data collected about students' performance into assessment results. And knowing the possibilities and

limitations of various interpretation models helps in designing a set of observations that is at once effective and efficient for the task at hand. Sophisticated interpretation techniques used with assessment tasks based on impoverished models of learning will produce limited information about student competence. Likewise, assessments based on a contemporary, detailed understanding of how students learn will not yield all the information they otherwise might if the statistical tools used to interpret the data, or the data themselves, are not sufficient for the task.

**Report on
Options for Assessments
Aligned with the
Common Core
State Standards**



*Submitted
to the
Michigan Legislature*

*December 1,
2013*

ACKNOWLEDGEMENTS

The completion of this report would not have been possible without the dedicated efforts of the Michigan Department of Education (MDE) staff listed below in alphabetical order.

Brandy Archer	Content Area Literacy Consultant, Office of Education Improvement and Innovation
Matt Ayotte	Online Assessment Specialist, Office of Systems, Psychometrics, and Measurement Research
Stephen Best	Assistant Director, Office of Education Improvement and Innovation
Phil Chase	Composition and Professional Development Manager, Office of Standards and Assessment
Deb Clemmons	Executive Director, School Reform Office
Doug Collier	Financial Manager, Office of Assessment Business Operations
Vince Dean	Director, Office of Standards and Assessment
Gregg Dionne	Supervisor, Curriculum and Instruction, Office of Education Improvement and Innovation
Jan Ellis	Communications and Process Specialist, Division of Accountability Services
Sue Fransted	Department Analyst, Office of Standards and Assessment
Linda Forward	Director, Office of Education Improvement and Innovation
Jill Griffin	Consultant, Urban Education, Grant Requirements, and Credit Recovery Office of Education Improvement and Innovation
Jim Griffiths	Test Administration Manager, Office of Standards and Assessment
Amy Henry	Consultant, Statewide System of Support, Office of Education Improvement and Innovation
Ruth Anne Hodges	Mathematics Consultant, Office of Education Improvement and Innovation
Linda Howley	Accessibility Specialist, Office of Standards and Assessment
Dave Judd	Director, Office of Systems, Psychometrics, and Measurement Research
Venessa Keesler	Deputy Superintendent, Division of Education Services
Pat King	Senior Project Management Specialist, Office of Systems, Psychometrics, and Measurement Research
Tom Korkoske	Financial Analyst, Office of Assessment Business Operations
Joseph Martineau	Deputy Superintendent, Division of Accountability Services
Kim Mathiot	Composition Editor, Office of Standards and Assessment
Andy Middlestead	Test Development Manager, Office of Standards and Assessment
Kim Mull	Graphic Designer, Office of Standards and Assessment
Jen Paul	English Learner Assessment Consultant, Office of Standards and Assessment
Karen Ruple	Statewide System of Support Manager Office of Education Improvement and Innovation
Paul Stemmer	National Assessment of Educational Progress State Coordinator Office of Standards and Assessment
Steve Viger	Measurement Research and Psychometrics Manager Office of Systems, Psychometrics, and Measurement Research
Shannon Vlassis	Graphic Designer, Office of Standards and Assessment
Kim Young	Interim Assessment Consultant, Secondary Level, Office of Standards and Assessment

In addition, MDE would like to thank Brian Gong, Executive Director of the Center for Assessment at the National Center for the Improvement of Educational Assessment, for providing an independent review of a draft of this report for the purpose of ensuring its clarity and usefulness.

Finally, MDE would like to express its sincere appreciation to all the service providers that responded to the abbreviated Request for Information (RFI) represented by the survey that forms the basis for this report's content.

TABLE OF CONTENTS

Introduction	4 - 5
Summative Assessment	6 - 21
Content and Item Type Alignment	6 - 7
Transparency and Governance	8 - 9
Overall Design and Availability	10 - 11
Test Security	12 - 13
Scoring and Reporting	14 - 15
Cost - Standard Product	16 - 17
Constructed Response – Standard Product Cost Implications	18 - 19
Interim Assessment	20 - 33
Content and Item Type Alignment	20 - 21
Transparency and Governance	22 - 23
Overall Design and Availability	24 - 25
Test Security	26 - 27
Scoring and Reporting	28 - 29
Cost - Standard Product	30 - 31
Constructed Response – Standard Product Cost Implications	32 - 33
Accessibility	34 - 35
Technical Requirements	36 - 37
Formative Assessment Resources	38 - 39
Local Implications	40 - 41
Summary Conclusions and Recommendations	42 - 43
References	43
Appendices	
Appendix A - Survey	
Appendix B - Survey Responses	
Appendix C - Survey Questions Cross Reference Table	

COMMON CORE Assessment Options Report

INTRODUCTION

Michigan students and educators need a rich, next-generation assessment system that is suitable for the numerous, high-stakes purposes toward which it will be applied. The solutions described in this report must be considered in light of how the test results will be used, and the fact that every school, educator and community will feel real consequences of their use, both intended and possibly unintended. Michigan's transition to new, online assessments that include multiple measures designed to capture student achievement and growth, is a powerful opportunity to improve the strength of our entire education system. This report represents an important source of information about the various options available to the state.

The Legislative Resolution

Both the House Concurrent Resolution 11 passed by the Michigan House of Representatives on September 26, 2013 and the substitute for House Concurrent Resolution 11 passed by the Senate on October 24, 2013 (subsequently adopted by the House on October 29, 2013) included a requirement for the State Board of Education and MDE to develop and submit a report on options for assessments fully aligned with the Common Core State Standards (CCSS). The report was to be completed and submitted to both chambers of the legislature by December 1, 2013 and be factual and unbiased. In addition, the final resolution expressed a preference for state assessments that are computer adaptive, provide real-time results, are able to be given twice per year, and assist in the evaluation of individual teachers. Other requirements included availability by the 2014-15 school year in grades 3 through 11.

In order to comply with the final resolution, the primary requirement for assessment solutions described in this

report is that they be adequately aligned with Michigan's college- and career-ready standards, in this case the CCSS. Some aspects of alignment (e.g., coverage of the mathematics and reading standards) are relatively straightforward. Other facets are more challenging to capture and have far-reaching implications for categories such as cost. An example of this is the use of Constructed-Response (CR) items; test questions that require students to develop a short or long written response. These are often significantly better than other types of items for measuring complex skills such as research, problem solving or communicating reasoning, that are found in the CCSS. However, these types of items are often time-consuming for students to answer and are the most expensive and complicated to score. Because CR items have significant implications for a variety of categories presented in this report, references will be made to them in appropriate sections, and overall implications will be described in the summary conclusions and recommendations section.

MDE Request for Information Process

In order to complete this project by December 1, 2013, the decision was made to develop a survey covering the primary topics of concern and permit any vendor registered to do business in Michigan through the state's Buy4Michigan web system to respond. Development of the survey commenced immediately following the approval of the Senate resolution on October 24, when it was apparent that the final resolution was highly likely to require this report. Through the Buy4Michigan website, 185 entities are registered under the category of educational examination and testing services. The survey was posted to the site; each registered entity received a notification email indicating that an opportunity was available for them on October 30, and indicated that all replies were due in two weeks. Twelve service providers submitted responses, all of which were included in the report.

The survey questions were separated into three distinct categories, to capture information on the three primary types of assessment solutions that are essential elements of a balanced assessment system needed to support educational improvement. The goal of this was to learn what solutions were available or being developed for:

- **Summative purposes** (e.g., test like MEAP for high-stakes accountability)
- **Interim purposes** (e.g., tests administered multiple times throughout the year to measure student growth with relative frequency), and
- **Formative purposes** (e.g., resources to support real-time measurement of student learning).

It was also important to ask about these different classes to ensure that no service provider was excluded for having a viable solution for one product in any category, versus requiring that each vendor have something for all three categories. MDE is open to the idea that the strongest overall solution in the end may involve selecting the 'best in class' for each type, although this concept introduces substantial risk on aspects such as the comparability of test scores.

Responding to the extensive survey in two weeks was undoubtedly a challenging task for service providers, as the questions were detailed and covered a wide range of topics. MDE is appreciative that so many qualified teams made the time to submit complete responses. One service provider, the Northwest Evaluation Association (NWEA), which currently has products deployed in a number of Michigan schools, chose not to complete the survey and instead submitted a letter explaining some aspects of their assessments and why they elected to not complete the survey. Since they did not submit a full response, NWEA is not included in the report. However, as they were the only vendor to submit such a letter, and many Michigan stakeholders are familiar with what they have to offer, MDE felt it was appropriate to include their letter in Appendix B with the completed survey information from the other entities. The table below provides a summary of the report development schedule.

Report Development Milestones	
Senate passes resolution and survey development begins	October 24, 2013
Survey posted to Buy4Michigan website	October 30, 2013
Responses due from service providers	November 13, 2013
Report submitted to Legislature and State Board of Education	December 1, 2013

Organization and Composition of the Report

Once responses were received, MDE staff members needed to review them all, compile them by category, and assign ratings. In order to complete this task, teams of staff with relevant subject matter expertise were assigned to each category with explicit instructions on how to view the responses and assign ratings. It was determined that a 'Consumer Reports' type of display would be the most user-friendly. The tables displayed in the body of the report provide a snapshot of how each service provider completed the questions germane to each category. There are three important caveats about the ratings assigned:

- Due to the timeline, it was not possible to thoroughly evaluate the quality of evidence provided by each service provider. The highest rating is based on complete responses that included some evidence indicating they were likely to meet all requirements, the middle rating indicating unclear or partial meeting of requirements, etc. Therefore, development and rigorous vetting of scoring criteria could not be accommodated. Additionally, the decision was made to limit the number of rating categories to three, to help ensure that even if a longer timeline had been available and a more rigorous, fine-grained (e.g., 5 or 7 categories) scoring system developed, only minor changes in scoring would have likely resulted.

- Responses from service providers were not compared against each other, only against the content of the survey questions. Comparing responses across multiple survey questions related to each category would have required substantially more time in order to evaluate the quality of the response and accompanying evidence.
- It is important to remember that many of the solutions described in this report are under construction, so a true evaluation of their qualities will not be possible until after the first year of operational test administration.

Based on these caveats, it is **essential** to recognize that this report alone is not sufficient to determine which assessments would truly be viable with regard to measuring the full breadth of career- and college-ready standards, interfacing with state systems, not adding additional burdens to local districts and schools, and cost effectiveness.

A conscious decision was made not to consolidate the ratings for each category into an overall Executive Summary. This process would have diluted the responses provided by each service provider by not properly accounting for the many areas where solutions are partially available or in various stages of development. Based on this, each category should be reviewed on its own merits and given equal weight. Additionally, in a number of cases, the survey responses required a 'yes' or 'no' response, but the opportunity to provide comments for the purpose of further clarification was made available. This introduced nuances, or possible opportunities for negotiation in areas such as control over data or opportunities to have Michigan educators involved with test question development, that could not be captured equitably in each section's table or narrative. The survey responses from each service provider are included in their entirety, unaltered, in Appendix B, if any readers of this report are interested in exploring the comments that accompanied some responses.

In addition to the specific items listed in the final resolution, four key documents guided the development of the survey questions and helped shape the lenses through which the responses were viewed by Department staff. Two of the documents, the *Standards for Educational and Psychological Testing* and the *Standards and Assessments Peer Review Guidance* have been important sources of requirements for technical quality and ensuring that all state standards and assessment systems meet criteria specified under the federal Elementary and Secondary Education Act. Recently, two other documents have been produced to guide the development of high quality, next-generation assessments and thoroughly define the requirements and responsibilities of clients (e.g., states) and service providers in all aspects of bringing large-scale assessment programs to operational status. Respectively, these are the *CCSSO Assessment Quality Principles* and the *Operational Best Practices for Statewide Large-Scale Assessment Programs-2013 Edition*.

Content & Item Type Alignment

SUMMATIVE

INTRODUCTION

The Common Core State Standards are organized into five content areas: Mathematics, Reading, Writing, Listening and Speaking. They provide goals and benchmarks to ensure that students are achieving certain skills and knowledge by the end of each year. They were carefully written so that students leave high school with a deep understanding of the content and skills they need to be career- and college-ready. It is important, then, that the summative assessments accurately reflect the intended content emphasis and important understandings of each grade level, 3–8, and high school.

Multiple-choice and technology-enhanced item types are critical components of an assessment, and in order to truly assess the rigor and higher-order thinking skills required from the CCSS and career and college readiness, an assessment solution must offer a substantial number of constructed response items as well. Constructed response test questions are essential as they are the only item type that can truly measure certain areas of the CCSS such as writing, research, and problem solving skills. Please note a detailed report

Service Provider	Content Alignment				Item Types		
	<i>Content aligned to the CCSS</i>	<i>Solution addresses all 5 content areas</i>	<i>Solution addresses all grade levels (G3-G11)</i>	<i>Qualifications for educators involved in alignment for content, diversity and special populations</i>	<i>Standard item types (multiple choice and constructed response) will be available</i>	<i>Diverse set of technology-enhanced item types will be available</i>	<i>Performance tasks/assessments will be available</i>
ACT Aspire	●	○	●	○	●	●	○
Amplify Education, Inc.	NR	NR	NR	NR	NR	NR	NR
College Board	○	○	○	○	●	○	○
CTB/McGraw-Hill	●	●	●	○	●	●	●
Curriculum Associates LLC	●	○	●	○	●	●	○
Discovery Education Assessment	NR	NR	NR	NR	NR	NR	NR
Houghton Mifflin Harcourt/Riverside	○	○	●	○	●	●	●
Measured Progress	●	●	○	●	NR	NR	●
PARCC	●	●	●	●	●	●	●
Scantron	NR	NR	NR	NR	NR	NR	NR
Smarter Balanced	●	●	●	●	●	●	●
Triumph Learning	NR	NR	NR	NR	NR	NR	NR

on constructed response items in the cost section of this report. The quantity of constructed response items will also be covered in both the cost and scoring and reporting sections of this report. Performance tasks provide insights into students' depth of knowledge on important content because they require students to engage in authentic problem solving and to persevere through the multiple steps of the task.

CONCLUSION

Of the 12 respondents, two of them, Measured Progress and PARCC, indicated that their solutions included all five subject areas for summative assessments and were able to demonstrate sufficient evidence that their solution is aligned with the CCSS. Smarter Balanced has all but speaking as part of their current solution.

In terms of item types, CTB McGraw-Hill, PARCC, and Smarter Balanced were able to demonstrate item types that included standard item types, technology enhanced item types, and performance tasks for all grade levels and content areas.

- KEY:**
- — Appears to fully meet requirements based on responses provided
 - — Unclear if meets or appears to partially meet requirements based on responses provided
 - — Does not appear to meet requirements based on responses provided
 - NR — No response or did not indicate having a summative product



INTRODUCTION

It is essential that Michigan’s educators and students have an assessment system that meets the unique needs of the state while providing opportunities for valid comparison with other states and large-scale assessment systems. This means that a balance must be found between customizability and compromise, with service providers (e.g., with off-the-shelf products) and other states (e.g., with multi-state consortia), in order to find the best solution for these two competing goals. Michigan is one of a few states that have significant experience with this challenge. Our current assessment programs include tests that are state-specific and completely customized (e.g., MEAP) and tests that are not customizable (e.g., the ACT, which is part of the Michigan Merit Examination) as they are administered across many states and therefore must be static for important reasons such as test security and comparability. Over the course of several months of testimony and debate around implementation of the Common Core State Standards, it was apparent that Michigan’s ability to retain control over elements such as personally-identifiable student data was crucial. This section of the report includes ratings for responses to survey questions documenting opportunities for Michigan educators and MDE staff to have a direct and substantive influence in the development and operational implementation of the assessments.

The ratings in the table above were made in light of the high-stakes purposes that summative tests are designed to inform. These types of tests are the ones

Service Provider	Clear opportunities for Michigan educators to participate in...			Clear opportunities for MDE involvement in...			Clear evidence the State of Michigan retains sole and exclusive ownership of all student data
	Test question development processes	Bias/sensitivity and accessibility reviews of test questions	Test question scoring processes	Test design	Test question scoring administration and reporting processes	Technical quality processes	Retains sole and exclusive ownership of all student data
ACT Aspire	○	●	○	○	◐	○	○
Amplify Education, Inc.	NR	NR	NR	NR	NR	NR	NR
College Board	●	●	●	◐	◐	○	○
CTB/McGraw-Hill	◐	●	●	◐	◐	●	●
Curriculum Associates LLC	○	○	○	○	◐	○	○
Discovery Education Assessment	NR	NR	NR	NR	NR	NR	NR
Houghton Mifflin Harcourt/Riverside	●	●	●	●	●	●	●
Measured Progress	●	●	●	●	○	○	●
PARCC	◐	●	●	●	●	●	●
Scantron	NR	NR	NR	NR	NR	NR	NR
Smarter Balanced	●	●	●	●	●	●	●
Triumph Learning	NR	NR	NR	NR	NR	NR	NR

where comparability across schools, districts and/or states is paramount, and historically, data from them have formed the foundation for accountability systems. In light of this, it is essential that opportunities exist for Michigan educators to provide input in the development of the products used in the accountability system. This is very important with regard to demonstrating validity, especially when these instruments will be used for accountability metrics and evaluations. As administrators of these systems, it is also critical that MDE staff have a formal governance role to assure the results of the assessments are defensible.

If readers are interested in reviewing the specific survey questions and a particular service provider’s response, Appendix B contains the necessary information.

CONCLUSION

As documented in the table above, it is evident that there is a limited number of options where the opportunity to strike a balance between a customizable solution for Michigan and a purely off-the-shelf product exists for summative assessments. Based on the responses to the survey questions on this topic, only the College Board, CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside, Measured Progress, PARCC and Smarter Balanced appear to be developing solutions that permit robust opportunities for Michigan educator involvement in developing test questions. MDE input into essential areas of governance and design is only apparent with this same group of service providers,

- KEY:** ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having a summative product

albeit with much more opportunity in the case of Houghton Mifflin Harcourt/Riverside, PARCC, and Smarter Balanced. The other service providers clearly indicated that significantly fewer opportunities existed for MDE input in these two areas.

It is also important to note that clear differences exist with regard to Michigan control of student data for these high-stakes summative tests. In light of that critical factor, only CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside, PARCC, and Smarter Balanced would be recommended for further consideration based on the responses to this survey.

Overall Design & Availability

SUMMATIVE

INTRODUCTION

Michigan is committed to adopting online assessments, as well as providing paper-and-pencil test options while schools and districts continue to acquire and implement the technology required to administer online assessments. MDE believes strongly that the nature of computer-adaptive assessment, where each student receives a customized test event based on his or her performance, is the best solution for improving how student achievement and growth is measured. This is particularly true in the case of high-achieving students, and students that may be much lower than average due to factors such as disability, learning English, etc.

When reviewing the survey responses, MDE asked each respondent to note if their solution offered a computer-adaptive or computer-based test, as well as a paper-and-pencil option for administration. One key difference between computer-adaptive and computer-based assessments is that a computer-adaptive assessment scientifically selects items based on estimated student ability level, therefore a unique test is crafted for each student. A computer-based test is a fixed-form test (similar to paper-and-pencil solutions) where every student will see the same items. MDE also believes that offering a re-take opportunity for the summative assessments is a key component of our desired assessment system.

Service Provider	Overall Test Design				Availability <i>(Solution will be fully available (including all item types) for the 2014–2015 school year)</i>
	<i>Solution will be available in a computer-adaptive modality</i>	<i>Solution will be available in a computer-based modality</i>	<i>Solution will include a comparable paper-pencil option</i>	<i>Solution will offer a re-test option</i>	
ACT Aspire	○	●	●	○	●
Amplify Education, Inc.	○	●	●	NR	○
College Board	○	○	●	NR	○
CTB/McGraw-Hill	●	●	●	○	●
Curriculum Associates LLC	●	○	○	●	◐
Discovery Education Assessment	○	●	●	NR	○
Houghton Mifflin Harcourt/Riverside	○	●	●	●	●
Measured Progress	○	●	●	●	○
PARCC	○	●	●	●	●
Scantron	◐	●	●	NR	●
Smarter Balanced	●	○	●	●	●
Triumph Learning	◐	●	●	NR	◐

All but one of the solutions presented clearly offered an online administration option; although only three, Curriculum Associates, CTB McGraw-Hill, and Smarter Balanced, offered an online computer-adaptive delivery of their assessment. Two of the three, CTB McGraw-Hill and Smarter Balanced, clearly offered a comparable paper-pencil alternative to their proposed computer-adaptive assessment. Two of the solutions presented were not clear in their offerings as they may have only had a computer-adaptive solution in certain grades or content areas.

Of the solutions presented, ACT Aspire and CTB McGraw-Hill did not offer a re-take option for their summative assessment.

Based on the information provided in the survey, many of the solutions would be fully available by the 2014-2015 school year as desired. ACT Aspire, CTB McGraw-Hill, Houghton-Mifflin Harcourt Riverside, PARCC, Scantron, and Smarter Balanced indicated they would have solutions ready within that timeframe.

CONCLUSION

Given the information provided, if Michigan desires to continue in the direction of adopting an online computer-adaptive assessment system with a comparable paper-pencil alternative and a re-take option, it appears that the Smarter Balanced solution would be the only one prepared to meet those requirements. If Michigan decides that a computer-adaptive solution is not indeed a requirement then the solutions from Houghton-Mifflin Harcourt Riverside and PARCC would also be suitable options.

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having a summative product



INTRODUCTION

School accountability (including designation as priority or focus schools) and educator evaluation are high-stakes uses of Michigan’s next generation assessment. Test results from the next generation assessment must therefore be valid for such uses. A key to maintaining validity is the assurance that student performance reflects the learning that students have experienced rather than advance familiarity with test questions or receiving inappropriate assistance in obtaining a high score.

Critical to assuring that student performance reflects student learning are two issues:

- Keeping test questions secure.
- Timely monitoring for security breaches and an ability to respond appropriately.

This section focuses on survey questions providing evidence regarding how well each solution is able to address these two issues.

The number of test forms available for administration to students is critical to keeping test questions secure. A minimum standard is having at least one additional test form to administer to students in the event of a security breach. Even better is to have many forms available such that multiple forms can be administered in the same classroom. In the optimal situation, each student would receive a unique test form, as is the case for Computer Adaptive Testing (CAT). Providers were identified as meeting this criterion if they meet the minimum standard of having at least one additional test form available in the event of a security breach.

Timely provision of security-related data to MDE is critical in being able to monitor for security breaches and respond appropriately. MDE will need to be provided with timely access to security-related data in order to perform forensic analyses on the data for potential security breaches. Timely analysis is needed to initiate and conduct investigations, and (if possible) provide for re-testing, before the testing window closes in the case of a security breach. Providers were asked whether MDE would be provided with timely access to security-related data for analysis.

CONCLUSION

Because maintaining test security is so integral to appropriate high-stakes use of Michigan’s next generation assessments, MDE qualified only those that clearly indicated that at least one test form is available for use in the event of a breach of test security and clearly indicated that security-related data would be provided to MDE in a timely manner. The only service providers meeting this criteria based on the responses provided are CTB/McGraw-Hill, PARCC, and Smarter Balanced.

Service Provider	Assessment Integrity and Security	
	<i>Multiple forms are used in operational testing with others available for emergency or misadministration.</i>	<i>MDE will be provided timely and adequate information needed to monitor and investigate test administration, including student level data and psychometric data to perform forensic and security analyses</i>
ACT Aspire	○	○
Amplify Education, Inc.	NR	NR
College Board	◐	●
CTB/McGraw-Hill	●	●
Curriculum Associates LLC	●	◐
Discovery Education Assessment	NR	NR
Houghton Mifflin Harcourt/Riverside	○	●
Measured Progress	NR	NR
PARCC	●	●
Scantron	NR	NR
Smarter Balanced	●	●
Triumph Learning	NR	NR

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having a summative product



INTRODUCTION

Future scoring and reporting functions of state testing programs need to provide (1) faster, virtually instant, results back to the classroom; (2) more definition as to the depth of knowledge demonstrated by students on the content and standards being assessed; and (3) flexible testing reports that offer students, parents, teachers, and administrators the ability to access data specifically customized according to their individual learning, teaching, and/or evaluation needs.

To do this, we need systems designed to take the most efficient advantage of the data available.

Those who responded to the MDE request for information regarding their summative assessment offerings related to the Common Core State Standards were presented with a series of questions in two major areas regarding Scoring and Reporting. The two areas are Data Analysis Capabilities and Scoring, and Assessment Reporting.

In the area of Data Analysis Capabilities and Scoring, the focus was on vendor-provided products and data that would allow the MDE to run analyses verifying that vendor results were sufficient and accurate measures, as well as provide the MDE with additional opportunities for research and evaluation using the supplied data.

There was also emphasis on the amount of input the State would have into the design of student-level and aggregate data sets, statistical procedures, and scoring protocols. Having opportunities at the design level make it possible to assure the service provider is implementing

Service Provider	Data Analysis Capabilities and Scoring		Assessment Reporting			
	<i>MDE will have sufficient information for verification and analysis done in-house, using vendor-provided products and data.</i>	<i>MDE will have direct influence on student and aggregate level data structures, psychometric procedures, and scoring procedures and protocols.</i>	<i>Reporting will be at a level sufficient to provide necessary information to educators, MDE, and to satisfy federal and state requirements.</i>	<i>Reporting of assessment results will be timely (i.e., significantly improved over results from current, paper-pencil tests).</i>	<i>MDE and schools/districts will be provided with all data underlying the reports and will have the capability to perform further analysis if desired.</i>	<i>Students who test with State-approved accommodations will receive the same menus and types of score reports provided to students in the general population.</i>
ACT Aspire	○	○	○	◐	○	●
Amplify Education, Inc.	NR	NR	NR	NR	NR	NR
College Board	○	○	●	◐	○	●
CTB/McGraw-Hill	●	●	●	●	●	●
Curriculum Associates LLC	○	○	○	●	○	●
Discovery Education Assessment	NR	NR	NR	NR	NR	NR
Houghton Mifflin Harcourt/Riverside	●	○	●	◐	○	●
Measured Progress	NR	NR	NR	NR	NR	NR
PARCC	●	◐	●	NR	●	●
Scantron	NR	NR	NR	NR	NR	NR
Smarter Balanced	●	●	●	●	●	●
Triumph Learning	NR	NR	NR	NR	NR	NR

processes that are the most current and efficient, with an aim to obtaining the highest degree of reliability.

In Assessment Reporting, the areas examined include vendor provisions for:

- reporting at a level sufficient to provide necessary information to educators, MDE, and satisfy federal and state requirements.
- reporting of assessment results that will be timely (i.e., significantly improved over results from current, paper-pencil tests). The immediacy with which reports can be obtained following testing is of constant concern to our stakeholders at all levels. It is critical that new systems take advantage of the opportunities made available by computer-delivered testing.

- assurance that MDE and schools/districts will be provided with all data underlying the reports and will have the capability to perform further analyses if desired. Many schools want and need the capability to examine data in ways that serve their unique populations. This also assures that data will be available as needed to those involved in efforts where improvement is a critical priority.
- parity for students who test with State-approved accommodations to the extent they will receive the same menus and types of score reports provided to students in the general population.

CONCLUSION

The symbols displayed in the table on these pages provide a visual representation of how service providers offering a summative assessment product appear to meet the requirements for scoring and reporting. Based on responses provided, CTB/McGraw-Hill and the Smarter Balanced Assessment Consortium appear to fully meet requirements in all scoring and reporting categories.

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having a summative product



INTRODUCTION

This table displays the average, per-student cost for the standard summative products offered by each service provider. While most offered thorough solutions for most of the desired grade span (3-11) indicated in the resolution, there was some degree of variability. Average cost was generated by taking the mean price for each modality (i.e., computer-based assessment/computer-adaptive assessment (CBA/CAT) or paper/pencil) across all grades. This table is provided as an informational snapshot, to which MDE staff did not attempt to assign ratings; therefore no conclusions are provided for this section. While these proposed costs give some idea as to which products are likely to be more or less expensive in a general sense, the information gathered by the survey is insufficient to determine an accurate cost model.

As noted in the introduction to the report, that level of detailed information can only be produced by going through the full, formal state procurement process. The Grade Levels column of this section’s table indicates that service providers reported having items of each type for only those grades. Additional notes about this are included in the Exceptions Column.

Summative Assessment Per Student Cost (Standard Product)								
Service Provider	Average per student cost		Types of Test Questions Included					Exceptions
	CBA/CAT	P & P	Grade Levels	Multiple Choice (ELA & Math)	Constructed Response (ELA & Math)	Technology Enhanced (ELA & Math)	Performance Assessment (ELA & Math)	
ACT Aspire	22.00	28.00	Grades 3 - 10	●	◐	●	○	No constructed response test questions at Grade 9
Amplify Education, Inc.	NR	NR	NR	NR	NR	NR	NR	
College Board	NR	27.75	Grades 9 - 12	●	◐	○	○	No constructed response test questions in Mathematics; no constructed response test questions in ELA grades 9 and 10
CTB/McGraw-Hill	27.00	27.00	Grades 3 - 11	●	●	●	○	
Curriculum Associates LLC	11.00	NA	Grades 3 - 12	●	○	◐	○	No technology enhanced test questions in Grades 9-12
Discovery Education Assessment	NR	NR	NR	NR	NR	NR	NR	
Houghton Mifflin Harcourt/Riverside	20.00	25.00	Grades 3 - 12	●	●	●	◐	No performance assessments available for ELA
Measured Progress	NR	NR	NR	NR	NR	NR	NR	
PARCC	30.00	34.00	Grades 3 - 11	●	●	●	●	
Scantron	NR	NR	Grades 3 - 12	NR	NR	NR	NR	
Smarter Balanced	22.50	15.62	Grades 3 - 8, 11	●	●	●	●	
Triumph Learning	NR	NR	NR	NR	NR	NR	NR	

The Grade Levels column indicates that service providers reported having items of each type for only those grades. Additional notes about this are included in the Exceptions Column.

Additionally, a major driver of both cost and alignment is the number and type of constructed response items. Since the issues around these types of test questions are so pervasive, MDE staff determined it was necessary to display information about them in a separate table on pages 18-19.

- KEY:** ● — Appears to include this type of test question based on responses provided
 ◐ — Appears to include this type of question on some, but not all, subjects or grade levels. Please see the comment in the exception column
 ○ — Does not appear to include this type of test question based on responses provided
 NR — No response or did not indicate having an summative product

Constructed Response Standard Product Cost Implications

SUMMATIVE

INTRODUCTION

While multiple-choice and technology-enhanced test questions are types of items that are well-understood, easy to score, and comparatively cheap to produce, truly assessing the rigor and higher-order thinking skills required by career- and college-ready standards requires something more substantive. Any assessment solution that seeks to demonstrate the capability to measure and provide rich, student achievement and growth information on constructs such as writing, research, communicating reasoning and problem solving to the degree described in the CCSS, must offer test stimuli where students have the opportunity to do more than select 'a', 'b' or 'c'.

Examples of this idea include asking a student to summarize a reading passage in his or her own words, or write about the process he or she used to solve a math problem rather than just selecting the correct answer. Educators deserve strong data on how students are achieving and growing on these challenging topics. To attempt to learn more about what options are available now or in the near future to support this idea, the survey included questions specific to constructed-response items. Service providers were asked to list the number of constructed-response test questions that came with their standard product; numbers that are displayed in the following table.

Summative Assessment Constructed Response (CR) Test Questions Included in Per Student Cost Estimate									
Service Provider	Mathematics CR Test Questions				ELA CR Test Questions				Exceptions
	Hand Scored Short Answer	Hand Scored Extended Response	AI Scored Short Answer ¹	AI Scored Extended Response ¹	Hand Scored Short Answer	Hand Scored Extended Response	AI Scored Short Answer ¹	AI Scored Extended Response ¹	
ACT Aspire	0	4 -5*	1	0	0	1	1 -2	1	*No hand scored extended response test questions in Grade 9 mathematics
Amplify Education, Inc.	NR	NR	NR	NR	NR	NR	NR	NR	
College Board	0	0	0	0	NR	1**	NR	NR	**No hand scored extended response test questions in Grades 9-10 ELA
CTB/McGraw-Hill	4	1	0	0	3	1	0	0	
Curriculum Associates LLC	0	0	0	0	0	0	0	0	
Discovery Education Assessment	NR	NR	NR	NR	NR	NR	NR	NR	
Houghton Mifflin Harcourt/Riverside	3	1	0	0	3	2	0	0	
Measured Progress	NR	NR	NR	NR	NR	NR	NR	NR	
PARCC	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	
Scantron	NR	NR	NR	NR	NR	NR	NR	NR	
Smarter Balanced	0	4 - 7	0	0	5	1	0	0	
Triumph Learning	NR	NR	NR	NR	NR	NR	NR	NR	

¹ Artificial Intelligence

CONCLUSION

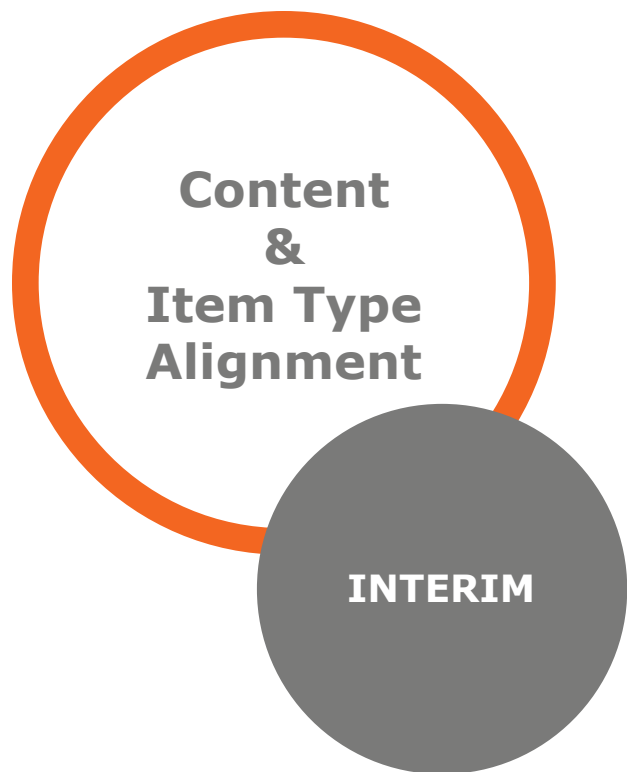
ACT Aspire, CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside and Smarter Balanced appear to include enough constructed-response items to measure student achievement deeply.

NOTES

As indicated in the text above and mentioned in other appropriate sections of this report, constructed-response test questions are considerably more expensive to

score than other types of test questions and student responses. Therefore, the survey included an opportunity for service providers to indicate whether or not they were able to provide additional constructed-response items beyond what they offered in their standard package, and a corresponding pricing structure. However, the portion of the survey seeking to gather information on this augmented option functioned differently, depending on the method the service provider used to complete the survey. As a result,

service providers interpreted the augmentation section differently and the information was not consistent or reliable. This was discovered as MDE staff began examining responses to this section and it was immediately evident that service providers interpreted this section in dramatically different ways. Therefore, the decision was made to not include information from the survey questions on augmented options (questions 70-73 in Appendix A).



INTRODUCTION

Interim assessments are given periodically throughout the school year. They provide information to educators about student learning and potential success with the summative assessments. The goal is to determine student achievement after instruction while there is still time to remediate areas in which students have done poorly. Michigan desires to have an interim assessment system that mirrors its summative counterpart and that uses an item pool that is independent from the summative assessment item pool.

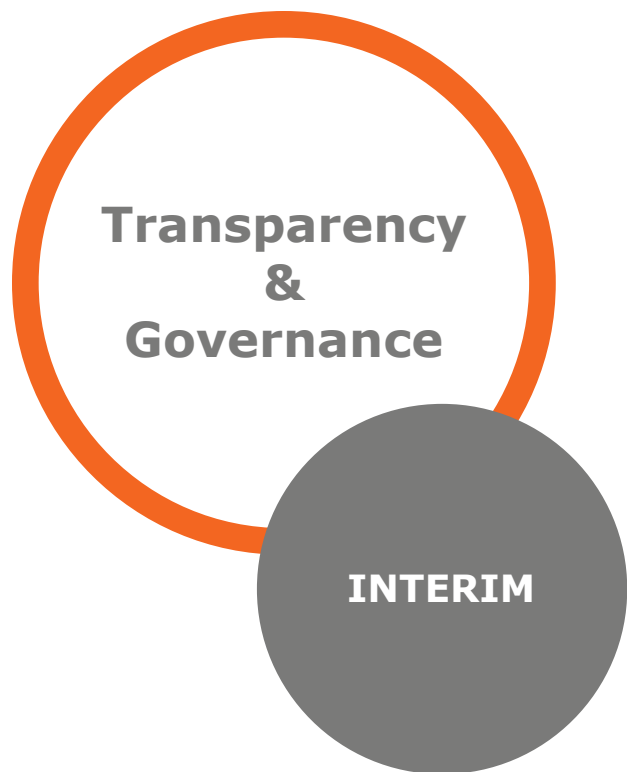
Multiple-choice and technology-enhanced item types are critical components of an assessment, in order to truly assess the rigor and higher-order thinking skills required from the CCSS and career and college readiness standards an assessment solution must offer a substantial number of constructed-response items as well. Please note that a detailed report on constructed response items is included in the cost section of this report. The quantity of constructed-response items will also be covered in both the cost and scoring and reporting sections of this report. Performance tasks provide insights into students’ depth of knowledge on important content because they require students to engage in authentic problem solving and to persevere through the multiple steps of the task. Comments on content alignment are based on survey responses only.

Service Provider	Content Alignment				Item Types		
	<i>Content aligned to the CCSS</i>	<i>Solution addresses all 5 content areas</i>	<i>Solution addresses all grade levels (G3-G11)</i>	<i>Qualifications for educators involved in alignment for content, diversity, and special populations</i>	<i>Standard item types (multiple choice and constructed response) will be available</i>	<i>Diverse set of technology-enhanced item types will be available</i>	<i>Performance tasks/assessments will be available</i>
ACT Aspire	○	○	●	○	○	○	○
Amplify Education, Inc.	○	●	●	○	●	●	●
College Board	NR	NR	NR	NR	NR	NR	NR
CTB/McGraw-Hill	○	●	●	○	●	●	●
Curriculum Associates LLC	○	○	○	○	●	●	○
Discovery Education Assessment	○	○	●	○	○	○	○
Houghton Mifflin Harcourt/Riverside	○	○	●	○	●	○	○
Measured Progress	●	●	●	○	○	○	○
PARCC	●	●	●	○	NR	NR	NR
Scantron	○	○	●	○	○	○	○
Smarter Balanced	●	○	●	○	●	●	●
Triumph Learning	○	●	●	NR	●	●	●

CONCLUSION

Amplify Education Inc., CTB McGraw-Hill, Measured Progress, PARCC, and Triumph report having all five content areas represented in interim assessments. Of these solutions, Measured Progress, and PARCC, demonstrated that their solutions were aligned to the CCSS through the survey. Five solutions (Amplify Education Inc., CTB McGraw-Hill, Measured Progress, Smarter Balanced, and Triumph) report the ability to offer standard item types, technology enhanced items, and performance tasks for the interim assessments they are building.

- KEY:**
- — Appears to fully meet requirements based on responses provided
 - — Unclear if meets or appears to partially meet requirements based on responses provided
 - — Does not appear to meet requirements based on responses provided
 - NR — No response or did not indicate having an interim product



INTRODUCTION

It is essential that Michigan’s educators and students have an assessment system that meets the unique needs of the state while providing opportunities for valid comparison with other states and large-scale assessment systems. This means that a balance must be found between customizability and compromise, with service providers (e.g., with off-the-shelf products) and other states (e.g., with multi-state consortia), in order to find the best solution for these two competing goals. Michigan is one of a few states that have significant experience with this challenge. Our current assessment programs include tests that are state-specific and completely customized (e.g., MEAP) and tests that are not customizable (e.g., the ACT, which is part of the Michigan Merit Examination) as they are administered across many states and therefore must be static for important reasons such as test security and comparability. Over the course of the months of testimony and debate around implementation of the Common Core State Standards, it was readily apparent that retaining Michigan control over elements such as personally-identifiable student data was crucial. This section of the report includes ratings for responses to survey questions documenting opportunities for Michigan educators and MDE staff to have a direct and substantive influence in the development and operational implementation of the assessments.

The ratings in this section’s table above were made in light of the purposes that interim tests are typically

Service Provider	Clear opportunities for Michigan educators to participate in...			Clear opportunities for MDE involvement in...			Clear evidence the State of Michigan retains sole and exclusive ownership of all student data
	Test question development processes	Bias/sensitivity and accessibility reviews of test questions	Test question scoring processes	Test design	Test question scoring administration and reporting processes	Technical quality processes	Retains sole and exclusive ownership of all student data
ACT Aspire	○	●	○	○	◐	○	○
Amplify Education, Inc.	○	○	○	◐	◐	○	●
College Board	NR	NR	NR	NR	NR	NR	NR
CTB/McGraw-Hill	●	●	●	◐	◐	●	●
Curriculum Associates LLC	○	○	○	○	◐	○	○
Discovery Education Assessment	◐	●	○	○	◐	○	○
Houghton Mifflin Harcourt/Riverside	●	●	○	●	●	●	●
Measured Progress	●	●	●	●	○	○	●
PARCC	●	●	●	●	●	●	●
Scantron	●	●	●	●	●	○	NR
Smarter Balanced	●	●	●	●	●	●	●
Triumph Learning	●	●	●	NR	NR	NR	●

designed to inform. For example, interim tests can be used to inform educator evaluations if all teachers in the same grade and subject administer them under the same conditions. Since it is likely that interim tests will be used for some accountability systems or purposes in Michigan schools, it is just as important that opportunities for involvement in test development and design be documented for them as for summative tests. If readers are interested in reviewing the specific survey questions and a particular service provider’s response, Appendix B contains the necessary information.

CONCLUSION

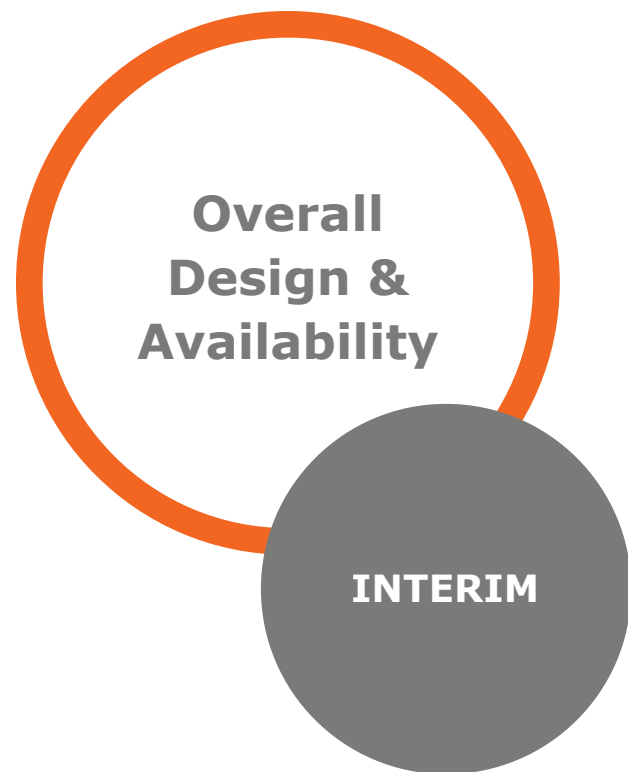
In addition to more service providers indicating that they have an aligned interim solution compared to this category for summative assessments, it is clear that

more opportunities exist for Michigan participation in test question development and governance activities. Based on the responses to the survey questions on this topic, CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside, Measured Progress, PARCC, Scantron, Smarter Balanced, and Triumph Learning all provide substantial opportunities for Michigan educator involvement in developing test questions. Lack of MDE input into essential areas of governance and design eliminates Measured Progress and Triumph Learning from the list. Since all the remaining service providers except Scantron affirm Michigan’s control over student data,

CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside, PARCC, and Smarter Balanced would be recommended for further consideration with regard to their interim assessment solutions.

It is also important to note that clear differences exist with regard to Michigan control of student data for these interim tests. In light of that critical factor, only CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside, PARCC, and Smarter Balanced would be recommended for further consideration based on the responses to this survey.

- KEY:** ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having an interim product



INTRODUCTION

Michigan is committed to building an interim assessment system that is completely internet-based, as well as providing paper-and-pencil test options while schools and districts continue to acquire and implement the technology required to administer online assessments. MDE believes strongly that the nature of computer-adaptive assessment, where each student receives a customized test event based on his or her performance, is the best solution for improving how student achievement and growth is measured. This is particularly true in the case of high-achieving students, and students that may be much lower than average due to factors such as disability, learning English, etc.

Michigan also desires an interim assessment system that would provide a great amount of flexibility and applications for Michigan educators. This would require that an interim assessment is available to be given at least twice a year, which would allow it to be used as an end-of-course, or a mid-year checkpoint as examples for educators and their students.

When reviewing the survey responses, MDE asked each respondent to note if their solution offered a computer-adaptive or computer-based test, as well as a paper-and-pencil option for administration. One key difference between computer-adaptive and computer-based

Service Provider	Overall Test Design				Availability
	<i>Solution will be available in a computer adaptive modality</i>	<i>Solution will be available in a computer based modality</i>	<i>Solution will include a comparable paper-pencil option.</i>	<i>Interim Solution(s) will have opportunity for multiple (at least twice per year) administrations.</i>	
ACT Aspire	○	●	●	●	●
Amplify Education, Inc.	○	●	●	●	○
College Board	○	○	●	NR	○
CTB/McGraw-Hill	●	●	●	●	●
Curriculum Associates LLC	●	○	○	◐	◐
Discovery Education Assessment	○	●	●	●	●
Houghton Mifflin Harcourt/Riverside	○	●	●	●	●
Measured Progress	○	●	●	●	●
PARCC	○	●	●	●	●
Scantron	◐	●	●	●	●
Smarter Balanced	●	○	●	●	●
Triumph Learning	◐	●	●	○	◐

assessments is that a computer-adaptive assessment scientifically selects items based on estimated student ability level; therefore a unique test is crafted for each student. A computer-based test is a fixed-form test (similar to paper-and-pencil) where every student will see the same items.

All but one of the solutions presented clearly offered an online administration option, although three, Curriculum Associates, CTB McGraw-Hill, and Smarter Balanced, offered an online computer-adaptive delivery of their assessment. Of those three, CTB McGraw-Hill and Smarter Balanced clearly offered multiple administration opportunities per year of their interim system.

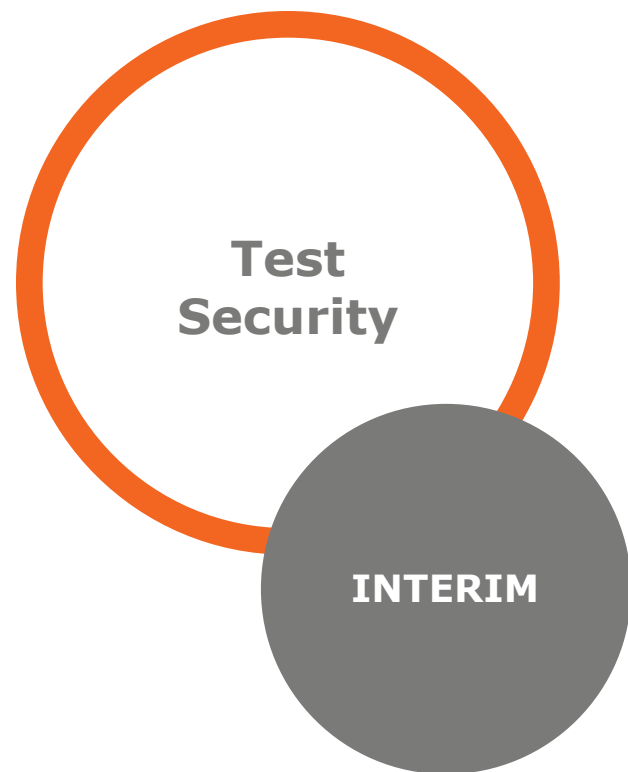
Based on the information provided in the survey, many of the solutions would be fully available by the 2014-2015 school year as desired. ACT Aspire, CTB McGraw-Hill, Houghton-Mifflin Harcourt Riverside, PARCC, Scantron, and Smarter Balanced all would have solutions ready within that timeframe.

CONCLUSION

Given the information provided, if Michigan desires an interim assessment solution that could be administered

in an online computer-adaptive system with a comparable paper-pencil alternative and offer multiple administrations per year, it appears that the solutions presented from CTB McGraw-Hill and Smarter Balanced would be prepared to meet those requirements. If Michigan decides that a computer-adaptive solution is not a requirement, then the solutions from ACT Aspire, Discovery Education, Houghton-Mifflin Harcourt Riverside, Measured Progress, PARCC and Scantron would also be suitable options.

- KEY:** ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having an interim product



INTRODUCTION

School accountability (including designation as priority or focus schools) and educator evaluation are high-stakes uses of Michigan’s next generation assessment. Test results from the next generation assessment must therefore be valid for such uses. A key to maintaining validity is the assurance that student performance reflects the learning that students have experienced rather than advance familiarity with test questions or receiving inappropriate assistance in obtaining a high score.

Timely provision of security-related data to MDE is critical in being able to monitor for security breaches and respond appropriately. MDE will need to be provided with timely access to security-related data in order to perform forensic analyses on the data for potential security breaches. Timely analysis is needed to initiate and conduct investigations, and (if possible) provide for re-testing, before the testing window closes in the case of a security breach. Providers were asked whether MDE would be provided with timely access to security-related data for analysis.

CONCLUSION

Because maintaining test security is so integral to appropriate high-stakes use of Michigan’s next generation assessments, for interim assessments MDE qualified only those that clearly indicated that security related data would be provided to MDE in a timely manner. The only service providers meeting this criteria based on responses provided were CTB/McGraw-Hill, Houghton Mifflin Harcourt/Riverside, Measured Progress, PARCC, and Smarter Balanced.

Service Provider	Assessment Integrity and Security
	<i>MDE will be provided timely and adequate information needed to monitor and investigate test administration, including student level data and psychometric data to perform forensic and security analyses</i>
ACT Aspire	○
Amplify Education, Inc.	◐
College Board	NR
CTB/McGraw-Hill	●
Curriculum Associates LLC	◐
Discovery Education Assessment	◐
Houghton Mifflin Harcourt/Riverside	●
Measured Progress	●
PARCC	●
Scantron	◐
Smarter Balanced	●
Triumph Learning	◐

- KEY:** ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having an interim product



INTRODUCTION

Future scoring and reporting functions of state testing programs need to provide (1) faster, virtually instant, results back to the classroom; (2) more definition as to the depth of knowledge demonstrated by students on the content and standards being assessed; and (3) flexible testing reports that offer students, parents, teachers, and administrators the ability to access data specifically customized according to their individual learning, teaching, and/or evaluation needs.

To do this, we need systems designed to take the most efficient advantage of the data available.

Those who responded to the MDE’s request for information regarding their interim assessment offerings related to the Common Core State Standards were presented with a series of questions in two major areas regarding Scoring and Reporting. The two areas are Data Analysis Capabilities and Scoring, and Assessment Reporting.

In the area of Data Analysis Capabilities and Scoring, the focus was on vendor-provided products and data that would allow the MDE to run analyses verifying that the vendor results were sufficient and accurate measures, as well as provide the MDE with additional opportunities for research and evaluation using the supplied data.

There was also emphasis on the amount of input the State would have into the design of student-level and aggregate data sets, statistical procedures, and scoring protocols. Having opportunities at the design level make

Service Provider	Data Analysis Capabilities and Scoring		Assessment Reporting			
	<i>MDE will have sufficient information for verification and analysis done in-house, using vendor-provided products and data.</i>	<i>MDE will have direct influence on student and aggregate level data structures, psychometric procedures, and scoring procedures and protocols.</i>	<i>Reporting will be at a level sufficient to provide necessary information to educators, MDE, and to satisfy federal and state requirements.</i>	<i>Reporting of assessment results will be timely (i.e., significantly improved over results from current, paper-pencil tests).</i>	<i>MDE and schools/districts will be provided with all data underlying the reports and will have the capability to perform further analysis if desired.</i>	<i>Students who test with State-approved accommodations will receive the same menus and types of score reports provided to students in the general population.</i>
ACT Aspire	○	○	○	◐	○	●
Amplify Education, Inc.	○	○	◐	◐	○	●
College Board	NR	NR	NR	NR	NR	NR
CTB/McGraw-Hill	●	●	●	●	●	●
Curriculum Associates LLC	○	○	○	●	○	●
Discovery Education Assessment	◐	○	○	●	○	●
Houghton Mifflin Harcourt/Riverside	●	○	●	◐	○	●
Measured Progress	○	○	○	◐	○	●
PARCC	●	◐	●	NR	●	●
Scantron	○	◐	○	●	●	●
Smarter Balanced	●	●	●	●	●	●
Triumph Learning	○	○	NR	◐	●	NR

it possible to assure the service provider is implementing processes that are reliable, efficient, and valid for the intended purposes.

In Assessment Reporting, the areas examined include vendor provisions for:

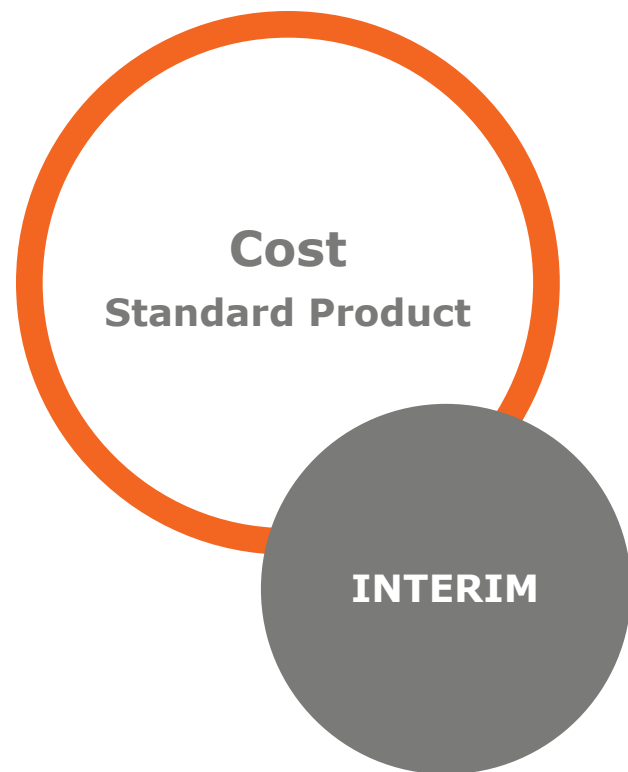
- reporting at a level sufficient to provide necessary information to educators, MDE, and to satisfy federal and state requirements.
- reporting of assessment results that will be timely (i.e., significantly improved over results from current, paper-pencil tests). The immediacy with which reports can be obtained following testing is of constant concern to our stakeholders at all levels. It is critical that new systems take advantage of the opportunities made available by computer-delivered testing.

- assurance that MDE and schools/districts will be provided with all data underlying the reports and will have the capability to perform further analysis if desired. Many schools want and need the capability to examine data in ways that serve their unique populations. This also assures that data will be available as needed to those involved in efforts where improvement is a critical priority.
- parity for students who test with state-approved accommodations to the extent they will receive the same menus and types of score reports provided to students in the general population.

CONCLUSION

The symbols displayed in the table on these pages provide a visual representation of how service providers offering an interim assessment product appear to meet the requirements for scoring and reporting. Based on responses provided, CTB/McGraw-Hill and the Smarter Balanced Assessment Consortium appear to fully meet requirements in all scoring and reporting categories.

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response or did not indicate having an interim product



INTRODUCTION

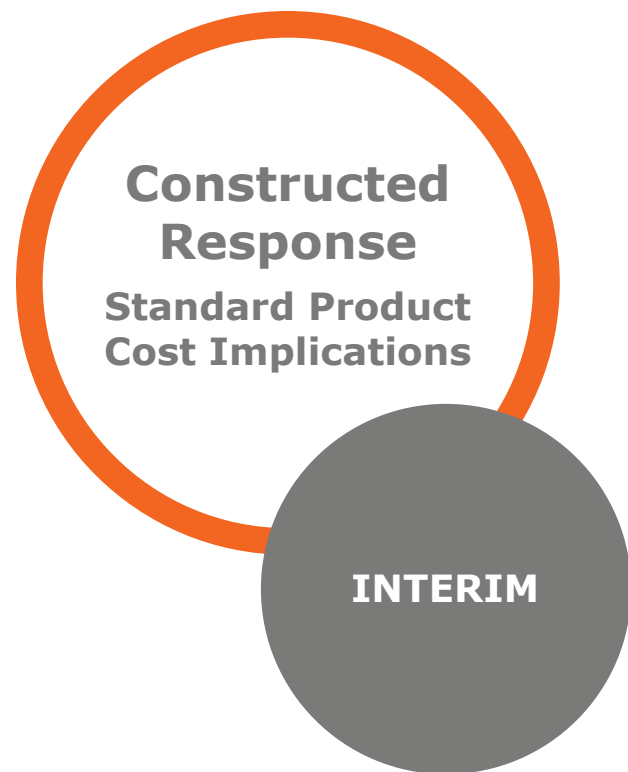
This table displays the average, per-student cost for the standard interim products offered by each service provider. While most offered thorough solutions for most of the desired grade span (3-11) indicated in the resolution, there was some degree of variability. Average cost was generated by taking the mean price for the computer-based assessment/computer-adaptive assessment (CBA/CAT) solution offered across all grades. This table is provided as an informational snapshot, to which MDE staff did not attempt to assign ratings. While these proposed costs give some idea as to which products are likely to be more or less expensive in a general sense, the information gathered by the survey is insufficient to determine an accurate cost model. As noted in the introduction to the report, that level of detailed information can only be produced by going through the full, formal state procurement process. The Grade Levels column of this section's table indicates that service providers reported having items of each type for only those grades. Additional notes about this are included in the Exceptions Column.

Additionally, a major driver of both cost and alignment is the number and type of constructed response items. Since the issues around these types of test questions are so pervasive, MDE staff determined it was necessary to display information about them in a separate table on pages 32-33.

Interim Assessment Per Student Cost (Standard Product)							Exceptions
Service Provider	Average per student cost CBA/CAT	Types of Test Questions Included					
		GradeLevels	Multiple Choice (ELA & Math)	Constructed Response (ELA & Math)	Technology Enhanced (ELA & Math)	Performance Assessment (ELA & Math)	
ACT Aspire	7.00	Grades 3 - 12	●	○	◐	○	No technology enhanced test questions in mathematics; no technology enhanced test questions in ELA grades 9-10
Amplify Education, Inc.	4.25	Grades 3 - 12	●	●	●	●	
College Board	NR	NR	NR	NR	NR	NR	
CTB/McGraw-Hill	13.00	Grades 3 - 12	●	●	●	●	
Curriculum Associates LLC	11.00	Grades 3 - 8	●	○	●	○	
Discovery Education Assessment	8.00	Grades 3 - 11	●	○	○	○	
Houghton Mifflin Harcourt/Riverside	10.00	Grades 3 - 12	●	●	●	○	
Measured Progress	5.70	Grades 3 - 11	●	●	●	◐	No performance assessments in grades 9-11
PARCC	NR	NR	NR	NR	NR	NR	
Scantron	5.00	NR	●	●	○	○	
Smarter Balanced	4.80	Grades 3 - 11	●	●	●	●	
Triumph Learning	20.00	Grades 3 - 12	●	●	●	●	

The Grade Levels column indicates that service providers reported having items of each type for only those grades. Additional notes about this are included in the Exceptions Column.

- KEY:**
- — Appears to fully meet requirements based on responses provided
 - ◐ — Appears to include this type of question on some, but not all, subjects or grade levels. Please see the comment in the exception column
 - — Does not appear to meet requirements based on responses provided
 - NR — No response or did not indicate having an interim product



INTRODUCTION

While multiple-choice and technology-enhanced test questions are types of items that are well-understood, easy to score, and comparatively cheap to produce, truly assessing the rigor and higher-order thinking skills required by career- and college-ready standards requires something more substantive. Any assessment solution that seeks to demonstrate the capability to measure and provide rich, student achievement and growth information on constructs such as writing, research, communicating reasoning and problem solving to the degree described in the CCSS, must offer test stimuli where students have the opportunity to do more than select 'a', 'b' or 'c'.

Examples of this idea include asking a student to summarize a reading passage in his or her own words, or write about the process he or she used to solve a math problem rather than just selecting the correct answer. MDE feels very strongly that educators deserve strong data on how students are achieving and growing on these challenging topics. To attempt to learn more about what options are available now or in the near future to support this idea, the survey included questions specific to constructed-response items. Service providers were asked to list the number of constructed-response test questions that came with their standard product; numbers that are displayed in the following table.

Interim Assessment Constructed Response (CR) Test Questions Included in Per Student Cost Estimate									
Service Provider	Mathematics CR Test Questions				ELA CR Test Questions				Exceptions
	Hand Scored Short Answer	Hand Scored Extended Response	AI Scored Short Answer ¹	AI Scored Extended Response ¹	Hand Scored Short Answer	Hand Scored Extended Response	AI Scored Short Answer ¹	AI Scored Extended Response ¹	
ACT Aspire	0	0	0	0	0	0	0	0	
Amplify Education, Inc.	66 - 171	16 - 59	0	0	5 - 21	19 - 50	0	0	
College Board	NR	NR	NR	NR	NR	NR	NR	NR	
CTB/McGraw-Hill	12 - 34	2 - 8*	2 - 20	0	12 - 19	8 - 10	2 - 20	0	*No hand scored extended response test questions for Grade 12 math
Curriculum Associates LLC	0	0	0	0	0	0	0	0	
Discovery Education Assessment	0	0	0	0	0	0	0	0	
Houghton Mifflin Harcourt/Riverside	0	0	3	1	0	0	3	1	
Measured Progress**	16	8	0	0	0	8	0	0	Information is for Grades 3-8 only; NR for High School
PARCC	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	
Scantron	0	0	0	0	0	0	0	0	
Smarter Balanced	2***	4 - 5****	0	3***	5	1	0	0	***No short answer test questions in math grades 3-8; extended response test questions in math are hand scored for grades 3-8 and AI scored for grades 9-12
Triumph Learning**	60 - 70	35 - 75	0	0	60 - 70	35	0	0	Information is for Grades 3-8 only; NR for High School

¹ Artificial Intelligence

NOTES

As indicated in the text above and mentioned in other appropriate sections of this report constructed-response test questions are considerably more expensive to score than other types of test questions and student responses. Therefore, the survey included an opportunity for service providers to indicate whether or not they were able to provide additional constructed-response

items beyond what they offered in their standard package, and a corresponding pricing structure. However, the portion of the survey seeking to gather information on this augmented option functioned differently, depending on the method the service provider used to complete the survey. As a result, service providers interpreted the augmentation section differently and the information was not consistent

or reliable. This was discovered as MDE staff began examining responses to this section and it was immediately evident that service providers interpreted this section in dramatically different ways. Therefore, the decision was made to not include information from the survey questions on augmented options (questions 70-73 in Appendix A).



INTRODUCTION

Michigan is committed to the inclusion of ALL students, including students with disabilities (SWD) and English language learners (ELLs), in large-scale assessment and accountability systems. Assessment results should not be affected by disability, gender, ethnicity, or English language ability, and all students should have an opportunity to receive valid scores for summative and interim assessments. To ensure validity, assessments must promote an equitable opportunity for ELLs, SWDs, and general education students. The challenge of how to include all students in these assessments brings accessibility issues to the forefront. The purpose of the Accessibility Category is to ensure that all students have the supports and tools they require in order to fully access Michigan’s assessment system. There are two types of accessibility features: Assessment Accommodations and Universal Tools. Assessment Accommodations are used to change the way students access a test without changing the content being assessed. In other words, accommodations equalize entry to the test without giving the student an unfair advantage, or altering the subject matter. For example, a blind student could access the test in Braille rather than print, and an English language learner may require test questions be translated into their primary language. Universal Tools can be used by any student who needs minor supports, such as a highlighter, magnifying device, or notepad. A series of questions aimed at determining the availability of these accessibility features for summative and interim assessments were included in the survey. Please refer to Appendices A and C.

Based on the results of the categorization process, the following is a list of the responders in rank order, top-to-bottom, who had the most offerings meeting Michigan’s accessibility requirements for their respective summative and/or interim assessments:

Service Provider	Accommodations for English language learners (ELLs)					Accommodations for Students with Disabilities (SWD)					Accessibility Tools	Translation Languages	Accommodations/ Reported Scores
	Embedded text-to-speech	English Glossing	Foreign Language Glossing	Full translation of test questions into language other than English	Universal accommodations	Embedded text-to-speech	Embedded video in ASL (human)	Refreshable braille	Print-on-demand tactile graphics	Universal accommodations			
ACT Aspire	●	○	○	●	●	●	◐	◐	◐	●	◐	○	●
Amplify Education, Inc.	○	○	○	●	●	○	○	○	○	●	◐	○	●
College Board	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	○	◐	●
CTB/McGraw-Hill	●	○	○	◐	●	●	○	○	○	●	◐	○	●
Curriculum Associates LLC	○	○	○	○	◐	○	○	○	○	◐	◐	○	●
Discovery Education Assessment	○	○	○	●	○	○	○	○	○	○	NR	○	●
Houghton Mifflin Harcourt/Riverside	○	○	○	○	●	○	○	○	○	●	◐	○	●
Measured Progress	●	○	○	○	●	●	○	○	○	●	◐	○	●
PARCC	●	●	●	●	●	●	●	●	●	●	◐	○	●
Scantron	○	○	○	●	○	○	○	○	○	◐	◐	○	●
Smarter Balanced	●	●	●	●	●	●	●	●	●	●	◐	●	●
Triumph Learning	○	○	○	○	●	○	○	○	●	●	◐	○	NR

Number of accessibility features meeting requirements	
Smarter Balanced	12
PARCC	11
ACT Aspire	6
CTB/McGraw-Hill	5
Measured Progress	5
Amplify Education, Inc.	4
Houghton Mifflin Harcourt/Riverside	3
Triumph Learning	3
Discovery Education Assessment	2
Scantron	2
College Board	1
Curriculum Associates LLC	1

CONCLUSION

Two of the respondents, Smarter Balanced and PARCC, provided sufficient evidence that their product meets all of Michigan’s expectations for providing appropriate accommodations on their respective assessments for ELLs and SWDs. None of the respondents met all requirements for universally-provided tools. Smarter Balanced was the only respondent to report they currently provide the required languages for translation.

All respondents except for Triumph Learning indicated they meet Michigan’s requirements for reporting valid scores for students using State-approved accommodations on their respective assessments.

- KEY:** ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response

Technical Requirements

INTRODUCTION

Service providers were asked to indicate the device types and operating systems supported by their Computer Adaptive Testing solution. Service providers were also asked to provide the bandwidth requirement for each testing site. These factors have a significant effect on the level of school technology readiness as well as the overall cost to schools and districts.

All of the Service providers that responded, with the exception of Triumph Learning, indicated that their online testing system supports Windows XP/7 desktop and laptop testing devices. Since Windows XP is still widely used in Michigan schools, it is critical that the online testing system provide support for these devices.

All service providers that responded indicated that their online testing system supports Mac OSX desktop and laptop testing devices. According to MTRAx, a technology readiness survey tool, OSX devices are also common among Michigan schools. Therefore, it is critical that the online testing system provides support for these devices.

An increasing number of schools are adopting Chromebook devices for student instructional and assessment use. Seven of the responding service providers indicated that their online testing system supports Chromebook as a testing device. iPads are also widely used in Michigan schools for instruction and assessment. Six of the responding service providers indicated that their online testing system supports the iPad as a testing device.

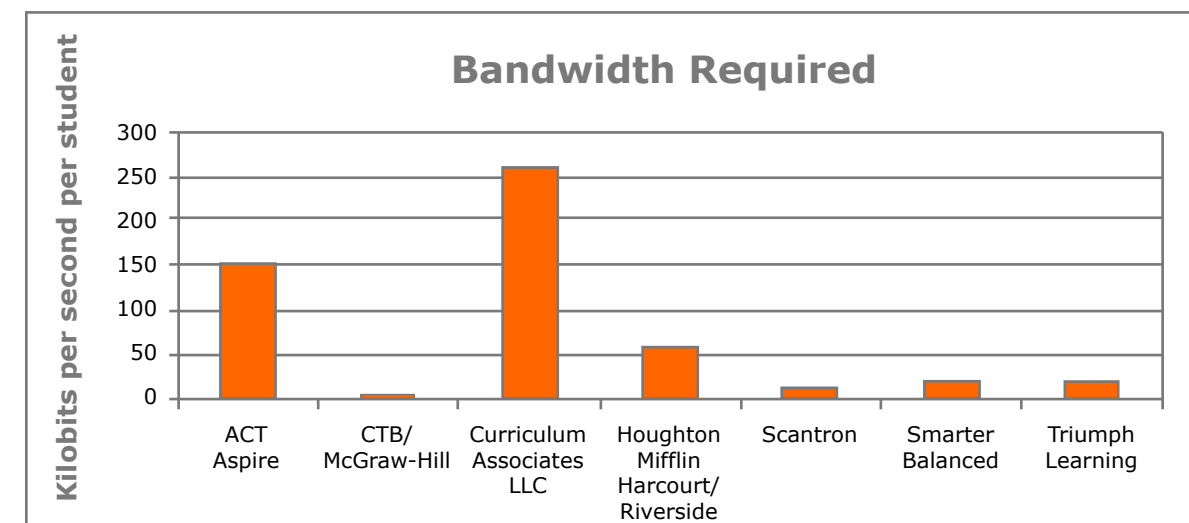
Service providers were asked if MDE would have a formal decision-making role with the ability to have a direct influence on the operating systems and technology platforms supported by their online testing system. Only PARCC and Smarter Balanced indicated MDE would have influence on the operating systems supported. Houghton Mifflin Harcourt/Riverside, PARCC, and Smarter Balanced indicated MDE would have influence on the technology platforms supported.

CONCLUSION

Many Michigan schools and districts have begun deployment of a variety of student-level mobile devices including Chromebooks and tablets. In many schools, these mobile devices are actually replacing the traditional computer lab configuration. Best practice calls for students to use the same device for both instruction and assessment. Therefore, the online testing system needs to support not only desktops and laptops but also Chromebooks and tablets (running iOS, Android, and Windows 8). Additionally, some schools have limited internet bandwidth available, which may limit the number of students that can test simultaneously.

Of the service providers that responded, Discovery Education Assessment and Smarter Balanced appear to meet the overall criteria regarding technical requirements.

Service Provider	The online testing system supports the use of						
	Windows XP/7 desktops/laptops	Windows 8 desktops/laptops	Mac OS X desktops/laptops	Chrome OS laptops (Chromebooks)	iOS tablets (iPads)	Android tablets	Windows 8 tablets
ACT Aspire	●	●	●	●	●	●	○
Amplify Education, Inc.	●	○	●	●	●	○	○
College Board	NR	NR	NR	NR	NR	NR	NR
CTB/McGraw-Hill	●	●	●	○	○	○	○
Curriculum Associates LLC	●	●	●	●	○	○	●
Discovery Education Assessment	●	●	●	●	●	●	●
Houghton Mifflin Harcourt/Riverside	●	●	●	○	●	●	○
Measured Progress	NR	NR	NR	NR	NR	NR	NR
PARCC	NR	NR	NR	NR	NR	NR	NR
Scantron	●	○	○	●	○	○	○
Smarter Balanced	●	●	●	●	●	●	●
Triumph Learning	◐	●	●	●	●	●	○



*Only those service providers that responded with a bandwidth requirement are displayed.

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response

Formative Assessment Resources

INTRODUCTION

The formative assessment process differs from summative and interim assessments in many ways. Fundamental to understanding these differences is knowing how and when formative assessment is used.

In 2006, Michigan education representatives collaborated with other state education leaders, Council of Chief State School Officers (CCSSO), and national and international experts on formative assessment to develop a widely cited definition of formative assessment:

“Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievements of intended instructional outcomes.” (CCSSO FAST SCASS 2006)

The importance of this definition is that it is compatible with research showing such practices to be an important driver of student learning gains. At the core of the formative assessment process is that it takes place during instruction to support student learning while learning is developing. This is a distinct difference from summative and interim assessment that are intended to assess students after an extended period of learning. Simply giving students an assessment in the classroom does not mean that the assessment is formative. Use of assessment evidence requires teachers to gain insights into individual student learning in relation to standards and to make instructional decisions and to use descriptive feedback to guide next steps. In addition, during the formative assessment process, student involvement is an essential component. Teachers seek ways to involve the student in “thinking about their thinking” (metacognition) to use learning evidence to close the gap and get closer to the intended learning target.

While formative assessment is not a new idea, teachers are not typically trained on it in-depth. Simply putting resources and tools into teacher hands is not sufficient. Sustained professional development is needed to apply sound formative assessment practices. This is why reviewing MDE staff included Professional Development as a rating category.

Service Provider	Compatible Definition	Online Availability	Variety of Classroom resources/tools/ strategies	Professional learning opportunities	Resources aligned to quality criteria	MI educator submission process	Cost Indicators
ACT Aspire	NR	○	NR	NR	NR	NR	NR
Amplify Education, Inc.	◐	◐	●	●	◐	○	No additional costs based on responses provided
College Board	NR	NR	NR	NR	NR	NR	NR
CTB/McGraw-Hill	●	●	●	◐	◐	●	Additional costs based on response provided
Curriculum Associates LLC	○	○	NR	NR	NR	NR	NR
Discovery Education Assessment	●	●	●	●	●	●	No additional costs based on responses provided
Houghton Mifflin Harcourt/Riverside	●	○	NR	NR	NR	NR	NR
Measured Progress	●	●	◐	●	●	◐	Additional costs based on response provided
PARCC	◐	●	◐	◐	◐	◐	NR
Scantron	●	●	●	●	◐	●	Additional costs based on response provided
Smarter Balanced	●	●	●	●	●	●	No additional costs based on responses provided
Triumph Learning	●	●	●	○	○	○	NR

CONCLUSION

Based on a review of survey responses, it appears that CTB/McGraw-Hill, Discovery, Measured Progress, Smarter Balanced, and Scantron may meet all or most of the stated requirements. Each indicates an online repository, a compatible definition of formative assessment, availability of classroom tools and professional learning resources, and opportunities for Michigan educators to submit additional resources. However, closer examination of resources and services that support educator understanding and use of the formative assessment process is encouraged.

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response

Local Implications

INTRODUCTION

Many of the preceding sections focus on global aspects for how various products or solutions were designed or intended to function. The MDE believes a consideration that must be given equal weight is the set of implications for Michigan districts and schools that come with each solution. The opportunity to implement a new assessment system, especially in light of the shift from paper-pencil test to those delivered by computer, means that careful examination of several issues is important to determine if the transition will add or remove a significant amount of the burden that comes with secure, large-scale and high-stakes testing. In order to maintain the validity of test results, it is critical that standardized processes be in place and adhered to by test administrators so that the tests remain secure and the results uncompromised. What that principle in mind, four primary factors (Test Security, Test Design, Platform Availability and Bandwidth Requirements) are presented here in light of the potential they have to substantially increase or reduce burden on local districts and schools, depending on how they are implemented. In order to express the rationale for why these elements are so important, they are reiterated here, as opposed to only in the preceding sections where they originally appear.

FACTOR DESCRIPTIONS

Test Security Deploying a large number of comparable forms can assure that few students see a particular set of test questions, significantly reducing the potential for cheating. Computer adaptive testing (CAT) takes this further in that each student sees a unique test form matched to his or her performance, dramatically

reducing the opportunities for cheating. If cheating occurs, identifying the extent, and providing additional testing opportunities places a significant burden on local districts and schools affected by the compromised test questions. A large number of test forms can help to reduce this risk for schools. CAT can substantially mitigate this risk. As described in the Overall Design & Availability and Test Administration & Security sections, providers responded to questions regarding number of forms and use of CAT. Providers meeting thresholds for multiple forms and/or CAT were: CTB/McGraw-Hill, Curriculum Associates, PARCC, and Smarter Balanced.

Test Design The design (e.g., CAT vs. fixed-form) of tests delivered via computer has another substantial implication with regard to test administration. In order to maintain test security, because students see the same set of test questions, fixed-form testing requires that all students be tested on the same day (or the same small set-of-days). This scenario would require every student to have a suitable device. This student to device ratio is a major cost driver for local districts and schools in moving from paper-and-pencil testing to online testing. Because each student taking a CAT test sees a unique test form, CAT allows for a long testing window, in turn making it possible for local districts and schools to move testing online even without one-to-one student-to-device ratios. Districts that will not be ready, even for this low bar of technology readiness, will need to have a paper-and-pencil option available. As described in the overall design & availability section, the two providers with CAT solutions with a paper & pencil option are CTB/McGraw-Hill and Smarter Balanced. While Scantron and Triumph Learning noted similar solutions it was unclear if they would meet Michigan’s needs as they only noted CAT solutions at certain grades.

Platform Availability In order to take advantage of technology purchases already made by local schools and districts, the solution adopted for Michigan must support the widest possible array of computing devices. The fewer platforms that are supported, the fewer the number of students that will be able to take the tests online, or the more new devices local schools and districts will need to purchase to make the move to online testing. As described in the technical requirements section, the providers indicating adequate availability on a wide variety of platforms include: Discovery Education Assessment, Smarter Balanced, and ACT Aspire.

Service Provider	Test Security	Test Design	Platform Availability	Bandwidth Requirements
ACT Aspire	○	○	◐	○
Amplify Education, Inc.	○	○	○	○
College Board	○	○	○	○
CTB/McGraw-Hill	●	●	○	●
Curriculum Associates LLC	●	○	○	○
Discovery Education Assessment	○	○	●	○
Houghton Mifflin Harcourt/Riverside	○	○	○	○
Measured Progress	○	○	○	○
PARCC	●	○	○	○
Scantron	○	○	○	●
Smarter Balanced	●	●	●	●
Triumph Learning	○	○	○	●

Bandwidth Requirements To maximize the number of students who can take assessments online without significant costs put toward increased bandwidth, the solution provided must require minimal bandwidth. As described in the technical requirements section, providers responded to a question about the bandwidth required for each student taking a test. The MDE review team qualified only those that were reasonably near the lowest requirement listed by any provider. MDE leniently qualified, based on our experience, providers requiring less than 50kbps per student. The providers meeting this threshold were: CTB/McGraw-Hill, Scantron, Smarter Balanced, and Triumph Learning.

CONCLUSION

In these four primary areas driving local implications, ACT Aspire, Curriculum Associates, Discovery Education Assessment, PARCC, Scantron and Triumph met one threshold, CTB/McGraw-Hill met three thresholds, and Smarter Balanced met all four.

KEY: ● — Appears to fully meet requirements based on responses provided
 ◐ — Unclear if meets or appears to partially meet requirements based on responses provided
 ○ — Does not appear to meet requirements based on responses provided
 NR — No response



SUMMARY CONCLUSIONS

This report on options for assessments aligned with the Common Core State Standards (CCSS) contains a substantial amount of information on the current status of a number of potentially viable solutions. Each element (i.e., Summative, Interim and Formative) required for a balanced, rigorous and fair system of measuring student achievement and growth currently exists or will be operational in the near future. However, since many components of the solutions presented for consideration are not yet fully operational, and none of the solutions currently provides all three components, a definitive recommendation for a full-service system is difficult.

Additionally, assessments used to inform the state accountability system are subject to review by the U.S. Education Department, as part of the No Child Left Behind Act of 2001. This review requires the state to demonstrate how well the tests match the purposes to which they are being applied. In order to do this, Michigan needs to be able to have complete information on all aspects of the development, administration, scoring and reporting of the assessment. Therefore, multiple survey questions that formed the basis of the content of this report attempted to capture the degree to which Michigan can participate in or have opportunities to thoroughly understand aspects of the assessments proposed by service providers.

As Michigan moves forward with new levels of accountability for districts, schools and for teachers, the Department believes strongly that Michigan educators and assessment experts must have opportunities to inform the design of the tests. This includes how test questions will be developed and scored, results will be reported and how the technical adequacy will be documented. MDE must have access to sufficient documentation to permit staff with content and assessment expertise to evaluate the quality of processes used to develop and implement each aspect of the system.

As important as ensuring LEAs have access to high-quality, secure summative (once-yearly) assessments is the need to provide high-quality interim (pre-post or more often) assessments and formative resources and tools (to provide professional learning to educators regarding gathering and using data to inform day-to-day instruction). LEAs currently procure interim assessments and formative assessment resources individually or in small groups (e.g., across Intermediate School Districts). This small-scale procurement is costly and creates significant challenges with regard to comparability. It will be much more cost-effective for the state to provide interim assessments and formative assessment resources online to LEAs, freeing up local resources and helping to ensure comparability across the state. This is essential as Michigan moves forward with implementing reforms such as educator evaluations. Another key factor is whether a provider's solution will increase or decrease the burden on local districts. Multiple questions addressed these issues as well.

Finally, a major driver is cost. As noted in this report's introduction, the cost information captured in this report only serves as a limited benchmark for off-the-shelf products. The only way to truly determine specific and detailed costs, at the student level or otherwise is to complete the full state procurement process. As all providers were within a reasonable ballpark on prices, all were identified as meeting this threshold. The full state procurement process has been completed recently for all aspects of test development, and is in the final stages of being completed for test administration. This process (from issuing an RFP through signing contracts with successful service providers) currently takes approximately eighteen months. The contracts currently in place or that are being finalized are scheduled to expire after the spring of 2016.

MDE had been proceeding with implementation of the CCSS and participating in the development of Smarter Balanced for three years, with the aim of ensuring at least one viable option for an assessment system aligned to CCSS is available to the state. At the time that these contracts were being prepared, Smarter Balanced was the only viable option available to the state, and as this report demonstrates, it remains the only viable option that can satisfy all of the multiple needs for test security, student data privacy, a Michigan governance role, Michigan educator involvement, minimizing local burdens, cost effectiveness, Michigan access to all data to allow for verification, and so on. Because Smarter Balanced was designed primarily by state assessment directors who understand these needs, this should not be a surprising result.

RECOMMENDATIONS

The state procurement process is lengthy in great part because there are appropriate protections built into the system. It also takes significant time, once a contract is signed, for vendors to get systems in place to serve the needs of Michigan students, schools, districts, and the state. Because of these time constraints, adopting a different solution at this time will result in not having an assessment for the 2014-15 school year, and would likely result in not having an assessment for the 2015-16 school year, putting MDE in violation of both state and federal law. This is the case even with continuing forward with MEAP, where development has been ceased to avoid unnecessary costs.

As the current contracts expire after the spring of 2016, it presently takes approximately eighteen months to complete the formal state procurement process, and it takes time for a new contractor to put systems in place, MDE recommends developing and issuing a new RFP in late 2014 that incorporates information from this report. Contracts put in place from that RFP process will be geared toward delivering summative, interim and formative solutions beginning with the 2016-17 school

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- The Council of Chief State School Officers (2013). *CCSSO Assessment Quality Principles*. Washington, DC: Author.
- The Council of Chief State School Officers & The Association of Test Publishers (2013). *Operational Best Practices for Statewide Large-Scale Assessment Programs-2013 Edition*. Washington, DC: Author.
- U.S. Department of Education. (2009). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. [Revised December 21, 2007 to include Modified academic achievement standards. Revised with technical edits, January 12, 2009] Washington, DC: Author.

year. By issuing a new RFP in the fall of 2014, we hope that more providers will be able to put forward a product that can meet Michigan's needs at that time. MDE is agnostic regarding what solution is ultimately chosen for the 2016-17 school year and beyond, as long as it meets Michigan's needs.

FINAL NOTE

One potential avenue for assessing student achievement against the Common Core that (due to the timeframe) could not be produced for this report is that MDE has the capability to develop a customized, high-quality assessment in-house. Technology solutions such as MDE's Item Banking System, Secure Site, and existing service provider systems are already in place. MDE is currently developing a suite of interim assessments across grades 3-12 in science and social studies, to support reform efforts such as educator evaluations. This path gives the state complete control over such things as alignment to standards and test administration procedures, and makes sense in those content areas as there are no multi-state consortia or test companies that have developed tests specifically to measure Michigan's science and social studies content.

MDE has not been pursuing this path for English language arts and Mathematics, as the in-house approach does not permit the state to take advantage of the resources available from a consortium of states working on rich solutions to measure the same content (i.e., the Common Core State Standards). For example, the MDE-developed science and social studies will have a limited number of item types (e.g., constructed-response) and resources to support professional development. Without substantial new funding, MDE would not be able to develop its own assessment with the rich item types necessary to adequately measure the level of knowledge and skill described by the CCSS. The economies of scale provided by working with a consortium have allowed Michigan to avoid those substantial new funding needs.



Phone: 1-877-560-8378
Website: www.michigan.gov/baa
Email: baa@michigan.gov