Expectations of State Assessments: More Information from Shorter Tests

Stuart Kahl
RMC Research Corp.
February 2019

Educational testing companies and state testing officials hear the same concerns from local educators, policy makers, and the general public: the tests are too long and they don't provide enough useful information. In essence, consumers of test results want the tests to be shorter and yield more information that teachers can use immediately. You don't need to be a measurement expert to sense that this may not be possible. Nevertheless, this is where we are. So how did we get here?

Let me first explain that my perspective is that of an educator and a testing company executive who has been involved in state testing for several decades. However, the issues (and market demands) are the same whether we are talking about state tests or commercial tests purchased by districts from test publishers. I can say with certainty that we have allowed expectations of statewide and much district-level testing to go astray. This did not happen overnight. Instead, the situation evolved over time as stakes associated with state testing increased.

A Little History

NAEP-like State Assessments

In the early years of state testing, when many states had no testing programs at all, others tended to use either commercial norm-referenced tests or customized state assessments, some modeled after the National Assessment of Educational Progress (NAEP). In the early 1980s, some New England states had NAEP-like assessments. Like NAEP, these programs tested samples of students and did not report individual student or school results. (Actually, NAEP didn't produce state results at that time either.) They did report results of subgroups by gender, race, type of community, etc. Following NAEP's example, interpretive reports were produced, which discussed individual item results, drew conclusions about what students knew and could do and what they could not, and offered implications for instruction.

The results from these programs seemed to be of greatest interest to university educators in the various disciplines. A common concern about the programs was that the local educators did not feel the reported results reflected their own students' capabilities. When they heard of poor statewide performance relative to a particular skill, they felt "my students could do that." Of course, they were probably remembering their students' performance on their tests right after the tested material was taught. (For program evaluation purposes, I believe retention should matter.

This is one of the reasons I believe state end-of-year assessments have an important role to play in education.)

Census Testing

In the mid 1980s, these NAEP-like assessments were abandoned and replaced by "census testing," by which all students in three non-adjacent grades were tested – grades 4, 8, and 11, for example. With all students in target grades tested, student and school results were reported. This addressed a concern of the local educators in that it gave students a reason to take the testing seriously. But it also increased the stakes associated with the test results for local school personnel. Since tax payer dollars paid for the state assessments, the results were made public, and school scores became hot topics in newspapers. Local school boards weren't happy when their kids' performance did not compare well with that of students in neighboring or similar communities. Teachers and administrators were held accountable. Personally, I believe this was an appropriate level of stakes for accountability assessments, which also provided useful information to help local educators identify problems in their instructional programs and possible solutions.

One of the problems with testing just a few non-adjacent grades was that school results could fluctuate from year to year simply because of the varying abilities of the groups of students passing through the tested grades. The smaller the school, the more this was an issue. The official guidance to local educators was to not jump to conclusions too quickly on the basis of one year's results. Rather they should have been looking for patterns of results over a few years, before making major programmatic adjustments. The school reports ultimately depicted three-year's worth of results. This, of course, tested the patience of both educators and policy makers, especially in the first couple years of new programs.

"Authentic" Assessment

Another development during the mid 1980s was the somewhat scandalous "Lake Wobegon" effect characterized by the much-publicized inflation of scores on standardized tests, including state tests. While there were many causes of this situation, the possible cause that got considerable attention was cheating or "gaming the system" by school personnel. Regardless of the causes – whether it was educator behavior or technical issues related to norming or data analysis procedures – it was hard to justify the finding that ninety-five percent of elementary schools could claim their school performance was "above average." I believe that anti-standardized-testing sentiment ultimately led to curriculum specialists taking greater control of state testing and to what became known as the "authentic assessment" era. The late eighties and nineties saw the implementation of many innovative state testing programs involving such approaches as performance and portfolio assessments designed to tap higher order thinking.

ESEA Reauthorizations

The 1994 reauthorization of the Elementary and Secondary Education Act (ESEA) was the Improving America's Schools Act (IASA). The next reauthorization was the No Child Left Behind Act (NCLB), signed into law in January of 2002. Among other things, IASA required states to adopt rigorous standards, assess student performance relative to them in three non-adjacent grades, and hold schools accountable for improvement. It could be said that states were slow to meet the requirements of the law or that the US Department of Education was slow to enforce the law. Toward the end of the 20th century, the feds toughened up through compliance agreements with states, requiring them to fall in line within established timelines.

Not seeing timely enforcement of IASA or enough improvement in student performance, the Bush administration and legislators in DC enacted NCLB. (Passed almost unanimously, NCLB also showed the country that attention was still being given to domestic issues in the wake of 9/11.) The new law had requirements regarding standards and assessments that were much the same as those of the preceding reauthorization. For example, in addition to rigorous standards and assessments mentioned above, IASA 1994 had already referred to adequate yearly progress (AYP), but left it to the states to determine how AYP was defined and how schools' failure to attain AYP would be dealt with. NCLB, however, added "teeth" to the law by defining AYP in terms of interim targets on the way toward 100 percent proficient students by 2014 – an impossible goal for a variety of reasons – and by identifying consequences for not showing adequate progress at yearly or two-year intervals.

IASA also required individual student interpretive and descriptive reports. NCLB added "diagnostic" to the list of report adjectives (another impossibility) and that the reports were to be provided "as soon as practicable after the assessment is given." Reasonable testing time and budgetary considerations allow a state end-of-year summative assessment to cover only a thin sampling of a year's worth of curricular content and skills. As a result, there is nothing remotely diagnostic about what information these tests could provide about individual students' capabilities. (More on that in a later section on test length and the use of results.)

In a 2009 piece I put in *Education Week*, I facetiously mimicked statements of gubernatorial candidates in one mid-Atlantic state:

> *Let's replace our current accountability assessment with a single, summative, formative, adaptive, diagnostic general achievement test that measures growth and yields immediate results that teachers can use right away to modify their instruction.*

Despite the absurdity of this thinking, many states and local education agencies, in seeking testing contractors, were issuing (and continue to issue) requests for proposals (RFPs) reflecting overly ambitious expectations of their tests.

As mandated by the law, in 2005-06 states implemented NCLB testing in seven grades (3 through 8 and once in high school) instead of three grades. Interestingly, this expansion helped address the problem I described earlier about year-to-year fluctuation in school scores from

testing in a few non-adjacent grades. However, it also caused more teachers to experience the impacts of high-stakes state testing. Additionally, it permitted the computation of student academic "growth" based on state testing, at least in two subjects in some grades.

In the early years of NCLB, teachers were getting hammered for their students' lackluster performance. Some of the criticism leveled at them was based on a lack of understanding some legislators and others had regarding assessment programs and their results. I recall one congressman noting that the percentage of students at one grade level in his state who scored proficient or above one year was approximately the same as the percent proficient for the same students in the next grade the following year. His publicly stated conclusion … they haven't learned a thing! This kind of thinking didn't sit well with local educators. They knew that on average their students achieve a year's "growth" academically during the course of a school year. And they wanted credit for this. As it turns out, NCLB ultimately allowed growth to be taken into account in determinations of AYP. Currently, academic growth is a school quality indicator in many state plans. Aggregating growth statistics across students for purposes of accountability is reasonable. However, methodologies for computing growth and interpretations of meaning underlying growth statistics have many issues that go well beyond the scope of this paper. Nevertheless, information on student growth is high on the list of things local educators want from testing.

Race to the Top and ESSA

As if NCLB didn't raise the stakes of state testing enough, the Obama administration's Race to the Top program, initiated in 2009, took the stakes to a still higher level. That program provided considerable funding to states (and ultimately districts) for innovation and school reform efforts. However, there were significant strings attached to that funding. To receive it, states had to adopt common college and career-readiness standards, implement common assessments measuring student performance relative to those standards, and weigh student achievement data heavily in the evaluation of teachers.

Given that the only college and career readiness standards that were in the works at the time were the Common Core State Standards for English language arts and mathematics and that the RttT program was also undertaking the creation of large state assessment consortia to develop and initially implement common assessments addressing those standards, the Race to the Top was, in effect, requiring states to adopt the Common Core and belong to one of the assessment consortia in order to receive RttT funding. While it is reasonable for the performance of a teacher's students to somehow play a role in the teacher's evaluation, the strict RttT requirement led to shaky, and in some cases just plain bad, practices. There were (and still are) issues of fairness associated with the statewide use of "value-added models" in teacher evaluations and even greater issues associated with meeting the RttT requirement in subjects and grades in which there are no statewide tests that could be used for the computation of student growth.

All this being said, the stakes for local educators became higher than ever. An unintended consequence of this situation was a significant increase in the use by districts of commercial or consortium-provided interim or benchmark assessments. School personnel wanted early warning about students or areas within school subjects that required additional attention before the high-stakes tests were administered, and they wanted more information on the specific learning needs of students. Is there any wonder over testing occurred and ultimately became a big concern of students, parents, and even the teachers themselves?

Needless to say, there was a great deal of pushback relative to the Race to the Top requirements and its consequences – none greater than that among legislators in Washington. The result – provisions of the Every Student Succeeds Act (ESSA), passed into law in 2015. The law expressly forbids the Secretary [of Education] to dictate or incentivize particular accountability systems, standards, assessments, or teacher evaluation requirements. While the general NCLB requirements for standards and assessments remain much the same in the new law, decisions about how to use the test data for purposes of accountability and evaluations of teacher effectiveness rest with the states. ESSA has also offered states some flexibility with respect to assessment innovations.

Many states have dropped out of the two major state assessment consortia and have implemented their own assessment programs. As for standards, many have continued to implement the Common Core State Standards or have adopted Common-Core-like standards. I should mention here that the development of the Common Core standards was not a federal initiative, although the Race to the Top requirements led many to believe the standards were federally-mandated national standards. Consequently, they became a political hot potato or lightning rod for critics of federal over-reach. They were actually the result of a collaborative state effort.

While most educators are probably well aware of the recent influences on state testing highlighted above, perhaps some of the early influences are new to them. And perhaps the discussion above provides a useful perspective even to the recent history. It suffices to say that the entire history of state assessment programs is one of continually increasing stakes associated with the test results. While ESSA represents a minor retreat to the pre-NCLB era, high stakes accountability assessment appears to be here to stay. The pressures of those stakes on educators have led to unreasonable expectations of traditional state summative measures as well as to unproductive, even inappropriate, practices relative to the use of their results. These are things we can and should address, and we can start by dealing with the demand for more information from shorter tests.

Test Length and the Use of Results

Reliability and Validity

State assessment officials and testing experts will tell you that state tests are already as short as possible, given the requirements for information they are to provide. Just as a population survey

during election season relies on the quality of its sample of people, so does a test rely on the quality of the sample of test items it includes. One aspect of that quality is the sample size. Are there enough test items to give us faith in our results – a sense that we would get a similar result if we tested again? That's reliability. Testing specialists know how long a test needs to be to achieve an acceptable level of reliability. A fifty-item multiple-choice test sampling from a subject area domain at a particular grade produces reliable total test scores for individual students. If the test includes some constructed-response items each worth more than a single point, then the test should have a total possible point value of approximately fifty. These guidelines can vary depending on the size of the domain to which one wants the results to generalize and on the stakes associated with the results.

A reliable test measures something well, but faith in the results also requires that it measures the "right stuff." That's a major consideration for test validity. Just as one wouldn't want a population survey prior to a presidential general election to sample members of only one political party, one wouldn't want an eighth-grade, end-of-year summative math test to include only statistics items. This is why state assessment instruments are subjected to alignment studies that assure "balance of representation." Similarly, it is important to assure that the items within a math subcategory, say statistics, are not mostly testing the same specific concept or skill, such as reading bar graphs. So alignment studies also check on "range of knowledge" covered within subcategories of a subject domain. And these studies also evaluate "depth of knowledge" to be sure the items are not all low-level recall items, but instead give appropriate attention to higher order cognitive skills as well.

Subtest Scores

Subtest scores for individual students are not as reliable as total test scores because a subtest includes only a fraction of the number of items needed to produce reliable total test scores. School scores (averages of students' total test scores or of their subtest scores) are more reliable because a lot of the individual student measurement error is random and cancels out when student scores are aggregated. Nevertheless, the ten to twelve items (or points) in a state test's subtest still represent only a limited sampling of the content of the subtest area. The interpretation and use of subtest results need to be undertaken with caution.

Turnaround Time for Results

Local educators and others believe it takes far too long for them to receive the results of their state tests. They may be right to some extent. However, given the thin sampling of a whole year's worth of curricular content that the typical state test of tolerable length can provide, it is not reasonable to expect results that have immediate use in guiding instruction. The primary use for which these tests are designed is for evaluating whole instructional programs, thereby guiding program improvement efforts. Thus, it might be the next school year when students can benefit from the most appropriate use of the results. Now, if the results aren't getting back to schools in time for implementing program improvements the next year, that's another matter. Also, there's

the "out-of-sight-out-or mind" factor. It's not a great idea to send test results out long after students, parents, and even teachers have forgotten about the testing.

A lot of folks attribute long turnaround times to the human scoring of students' constructed responses and essays. While it is true that human scoring of student work takes more time than computer scoring of multiple-choice item responses and other response formats designed for machine scoring, the human scoring process for state assessments usually takes weeks, and not the months of turnaround time about which people are concerned. One of the biggest time consumers in getting assessment results produced and disseminated is getting complete, accurate data files that are necessary before final analyses can be completed. When state assessments were primarily paper-based, it could take many weeks, even months, to track down response documents from all students eligible for testing. Despite clear instructions and timelines for the return of materials to testing contractors, there were always schools whose materials were delayed or returned improperly. Even with online testing the primary mode of test administration nowadays, there are still students (more in some states than in others) who are administered paper-based tests. And, of course, there is still the necessity to reconcile every school's enrollment information with tests taken. Nevertheless, late summer should be the latest that a state's spring testing results are made available.

Online Testing

One of the unfortunate approaches to reducing turnaround time for reports has been to reduce the amount of constructed-response testing. As suggested above, this action didn't really solve the problem. Actually, it was more a matter of cost savings, since human scoring is much more expensive than machine scoring whether the testing is paper-based or computer-based. One of the oft stated arguments for online testing is the quick turnaround of results. But that is only a reality if all item responses are machine-scoreable. And even then, some results that rely on those complete, accurate final data files would not be available right away.

What is lost with most online testing and computer scoring is the adequate assessment of deeper learning. While there have been some innovations in the development of machine-scorable approaches, only some of the technology-enhanced items (TEIs) we see do more than multiple-choice items to tap higher-order thinking skills. Dragging a response to another area on a computer screen is not appreciably different from marking a bubble for that response.

Many educators, parents, and policy makers are enamored with computer-adaptive testing (CAT). By this method, as a student progresses through a test, "next questions" are selected for the student to answer based on his/her performance on the previous questions. In this way, a test is "tailored" to a student's ability. Such a test is designed to zero in on a student's overall ability (and a total test score) as efficiently as possibly – meaning with as few items as possible. However, CAT often has problems with alignment as described earlier. Even if there are programmed requirements to assure some minimum number of items in different subcategories of a subject area, there can be range-of-knowledge issues within the categories. If a total test

score is the only score of interest, then machine-scoreable CAT is a reasonable approach to testing provided deeper learning is not a major interest. Still, subtest data for individual students is highly suspect.

In lieu of subtest data, some companies try to use a technique called item mapping to address the need for diagnostic information from testing. Item mapping was an approach used by the National Assessment of Educational Progress to describe the capabilities of groups of students, in terms of items on which two-thirds of the students at a particular ability level would likely succeed. This is not an approach for the diagnosis of individual student specific learning gaps. For more on CAT and item mapping, see The *Promise of Computer Adaptive Testing: The Quest for Information*.

The Appropriate Use of End-of-Year State Tests

At the risk of repeating the following point too often, state summative tests are not designed to provide 1) information teachers can use for immediate instructional decisions or 2) diagnostic information about individual students. The thin sampling of a year's content by a limited number of test items does not lend itself to these uses. Besides, it's a teacher's job to know far more about their students' individual learning gaps than they could learn from a state test.

However, these tests can provide excellent information for the purpose of program evaluation, including the monitoring of the effectiveness of new programmatic initiatives. For these uses, quick turnaround of results is not necessary, which means there is no reason related to turnaround time to avoid measures of deeper learning that can best be measured by items requiring human scoring.

State testing for program evaluation and improvement (and school accountability) should raise questions that could require further investigation. For example, "Why did one subgroup of students perform worse than another (either within the school or in the state) when their performances should have been similar?" or "Why do our students underperform others in a particular subtest area?" Finding the answers to such questions leads to corrective actions and program improvement. Actual improvement or the success of new initiatives (e.g., efforts to stress higher-order thinking or training in the instructional process of formative assessment) should be reflected in subsequent state testing results -- sometimes sooner, sometimes later. Some improvements may take a few years to bear fruit – time to impact a critical number of teachers or a critical number of students for an extended period of time.

External summative assessments (state assessments) are a necessary component of a balanced assessment system. Designed and used appropriately, they can lead to improved educational experiences of students. Other components of a balanced assessment system can accomplish the same, but in different ways. See the RMC post: *What does 'balance" mean in a balanced assessment system?*