BRIEF #3: ALABAMA ASSESSMENT TASK FORCE

MEASURING STUDENT LEARNING: ITEM TYPES FOR ALABAMA'S NEXT ASSESSMENT SYSTEM

Scott Marion & Juan D'Brot, Center for Assessment

January 12, 2018

If you want to know what a test measures, look at the items!

While this oft-repeated axiom in education is a bit of an over-simplification, it is more true than not. Items and tasks are the tools that we use to elicit student responses to support inferences about what students know and can do. The information produced from test items is the foundation of a validity argument in support of test scores. Therefore, the quality of test items and tasks builds or detracts from the credibility of the assessment system in the eyes of students, educators, parents, and the public. Importantly, test item development is one of the major cost drivers of a state testing program, so in addition to the primary focus on quality; we must also focus on getting and maintaining quality as efficiently as possible. In this brief, we discuss the following:

- ✓ An overview of the types of items and tasks that can be included on a summative test
- ✓ The opportunities and challenges associated with each of the commonly used item types
- ✓ Considerations for how to balance the tradeoffs

Overview of items and tasks

Large-scale test items historically have been classified into two very broad categories of items: selected- and constructed- response. Selected-response includes the ubiquitous multiple-choice items, but can also include a variety of related item types or arrangements such as item clusters and evidence-based selected-responses (two-part multiple-choice items). Constructed-response items or tasks can range from very short responses of a few words to multi-hour or even multi-day activities. These "extended" constructed-response items/tasks share many features with performance-based tasks, but it is very rare to include extended performance tasks on end-of-year state assessments due to time requirements and cost. With the advent of computer-based testing, we have seen a new class of items, often referred to as technology-enhanced items. The

common feature of such items is that they rely on the digital environment to support interactions among students and the content in ways that are just not possible on paper. This is an area that is rapidly developing, but holds tremendous promise for improving the ways in which we measure student learning.

Opportunities and challenges

Test design is an exercise in optimization under constraints. The same is true for item development. Every choice involves considerable tradeoffs. The name or class of item means less than what the responses to the item tell us about student performance. We must keep in mind the following questions as we consider our choices:

- ✓ What are we trying to measure?
- ✓ How will this type of item help us to measure these learning targets well?
- ✓ What is a close enough approximation to what we really want to measure?
- ✓ What resources are available?
- ✓ Will the assessment be given solely on computer or split between computer and paper/pencil?
- ✓ What are the potential intended positive and unintended negative consequences associated with our choice of items?

In the table that follows, we highlight the opportunities and potential shortcomings with the various item types;

Item Type	Opportunities	Challenges
Multiple-choice	Multiple-choice items have a long track record of success and efficient use. The student is presented with a prompt and is asked to select from among 4-5 response options (generally). The field has developed robust measurement models for scoring, scaling, and evaluating multiple-choice items and they are able to generate a considerable amount of "measurement information" quite efficiently. Multiple-choice items are often presented to students as being independent of one another, but they can are also grouped as clusters or testlets around a scenario or reading passage.	The major challenge with multiple-choice items is that they are limited in the complexity of thinking they can elicit from students. While the field has generally advanced beyond populating tests with items that call on rote memory, many multiple-choice items still rely on factual and procedural information. Consequentially, many are concerned that if the accountability test is populated with multiple-choice items, teachers may try to mimic such approaches in their classrooms at a cost to deeper learning.
Evidence-based selected-response	These items are essentially two-part multiple-choice items where the student answers a multiple-choice item, but then answers a second item in the pair to "explain" their original answer. Such items have considerable promise for going beyond the generally lower levels of thinking called for in typical multiple-choice items.	These items have been used on PARCC and Smarter Balanced and once some of the kinks were worked out, they have been somewhat effective. The scoring rules associated with the item type are still tricky (e.g., does the student have to get the first item right in order to get the second one correct?) and the measurement modeling (creating and maintaining score scales) is less straightforward.
Short constructed-response	Short constructed-response items ask students to generate a written response that is generally a paragraph or less or to solve a fairly straightforward problem in mathematics. When designed well, such items are very effective at generating complex thinking from students that goes beyond multiple-choice items.	The challenges associated with these items has to do with cost of scoring, if they have to be scored by a human rater, they require more testing time than multiple-choice items, and tend to generate fewer points (test information) per minute than multiple-choice items. Some short constructed-response items can be scored effectively by computer, but most cannot at this point.

Item Type	Opportunities	Challenges
Extended constructed-response	Extended tasks are best at probing strategic and deep thinking by students. They often require between 30-90 minutes each. They are often the most authentic types of items because they can better draw on real-world scenarios or problems than other types of items and tasks. Importantly, such extended-response tasks send a powerful signal for the types of activities that we would like to see teachers use in their classrooms.	Extended-response items are expensive to score, except in cases where writing responses can be scored efficiently by computer (becoming more prevalent). As the name suggests, such items require considerable time and including several such tasks on a test can greatly increase testing time. Finally, because such tasks can be memorable and time consuming, they pose challenges for field-testing and test equating.
Technology-enhanced items	As the name implies, these items rely on the technology platform to enhance the interactions between the student and the content. The field is still new, but progressing rapidly. Early versions of TEIs were not much more than video clips on multiple-choice items, but newer items allow for sophisticated simulations that require students to think deeply in order to respond to the question. These items offer considerable promise for advancing our measurement capacity in a cost-efficient manner.	The obvious challenge is that technology-enhanced items require students to be testing in a digital environment. The more "enhanced" the item, the harder it is to create a "paper clone" that can be administered to students still testing on paper and, therefore, the greater the threats to comparability. Further, the field is still learning about the technical measurement properties about the more innovative item types and how such items contribute to our understanding of what students know and can do. Another risk with TEIs is that many schools do not have enough of a digital footprint to allow students the opportunity to learn in a digital environment so the only time they experience such procedures and approaches in on the state test.

Wrestling with tradeoffs

If this was an easy choice, we wouldn't need a task force! Based on our experience, we think that a balance of item types where the state can capitalize on the advantages of each type while trying to minimize the unintended negative consequence or other risks of the item type is a

prudent approach. Of course, finding that balance is the real challenge! We encourage the State to consider exploring technology-enhanced items to the degree that digital capacity and digital opportunities for learning can be expanded.

Questions to Answer

We would like the Task Force to weigh in on the following questions:

1. What proportion of the test, in terms of time, points, and number of items do you think should be represented by **short constructed-response items**?
2. What proportion of the test, in terms of time, points, and number of items do you think should be represented by **extended constructed-response items**?
3. How does the task force feel about requiring the potential test contractor to include **technology-enhanced items** on the test?
4. Is the task force interested in exploring items such as **evidence-based selected-response** or other non-computer innovations?