BRIEF #5: ALABAMA ASSESSMENT TASK FORCE

FIELD TESTING AND MAXIMUM TEST TIME FOR ADMINISTRATION

Juan D'Brot and Scott Marion, Center for Assessment

January 12, 2018

A hotly debated topic in large-scale testing is that of testing time. What is the maximum amount of time that a student should spend taking a standardized testing to reflect their grade-level mastery? Arguments for and against longer tests are prevalent and include concepts like creating more authentic assessment (i.e., requiring longer tests) and minimizing interruptions to instruction (i.e., shorter tests). While testing time is an important factor to consider for an administration, it is important not to confuse the timing and length of a single end-of-year summative assessment (typically a very small percentage of available instructional time) when compared to assessments required by a school and district. Depending on the dual state and district requirements, the number of tests can be larger than expected. Thus, test length is a key consideration in test design.

Factors that Impact Test Length

As with most aspects of test development, there are tradeoffs associated with each decision along the way. There are several factors that impact test length, but not all will be covered by this brief. Some factors include, but are not limited to

- Content coverage
- Item types*
- Desired reliability
- Subscore reporting
- Adaptivity
- Field test design*
- Public perception

For the purposes of this brief, only those with asterisks will be discussed. The remaining factors will be covered by other Task Force discussions and decisions. We will then consider the set of factors to specify the overall test design and length.

Item Types

We will only cover item types briefly, as they are a primary focus in another brief. However, item types are a major driver of test time. The test will be longer if an assessment calls for

inquiry-based tasks or performance tasks that seek to measure analysis and problem-solving. The trade-off to consider is whether the additional information gained from the assessment justifies the increased amount of time students spend on that particular task. Often, an in-depth task will comprise multiple pieces of evidence, multiple questions, and multiple responses. Thus, an in-depth task should yield greater amounts of information. However, a very well-specified task could take multiple sessions or days to fully answer. These tasks have been used successfully in the past but require buy-in across the spectrum from the classroom to the state office. We ask you to keep in mind the role item types play in test length as you review the remainder of this brief.

Field Testing

Field testing, or the act of testing items in actual situations reflecting intended use, seek to provide an initial view of reliability and validity. Traditionally, items can be field tested in one of two ways: through (1) stand-alone field tests or (2) embedded field tests.

Stand-alone field tests are those instances when items are tested out in an independent administration. Stand-alone field tests are traditionally optional for districts and schools. In some states, mandatory field tests have been deployed, which can lead to public resistance. The benefit of stand-alone field tests is that they can accelerate test development timeframes[1] but student motivation may be diminished. That is, stand-alone field tests must be supported through their own administration events, which can lead to educators and students becoming aware that the administration is for a field test. Test length itself is not a concern, but the presence of a separate "summative" testing event could be.

Embedded field tests place the field test items into the administration of another operational assessment. The primary benefit of this approach is that educators and students do not know what items are operational and what items are for the field test. This mitigates issues regarding decreased student motivation on field test items. However, it does increase the length of the assessment. Depending on the number of test items necessary to field test, "blocks" of items are usually administered to different students in the same grade (e.g., any one student would receive 10 additional field test items in addition to the typical set of items on a form, but those blocks would differ by student). Embedded field tests can also be used to create vertical scales or to link different assessments.

While field testing is not the only issue to consider around testing time, understanding the role and impact of field tests is key to determining their value in the test length conversation. The following figure presents three conditions: (1) an operational (OP) test, (2) an OP test with an embedded field test (FT), and (3) an OP test with a stand-alone FT.

[1] See brief on test timeline: *Timeline for Test Development and Administration.*

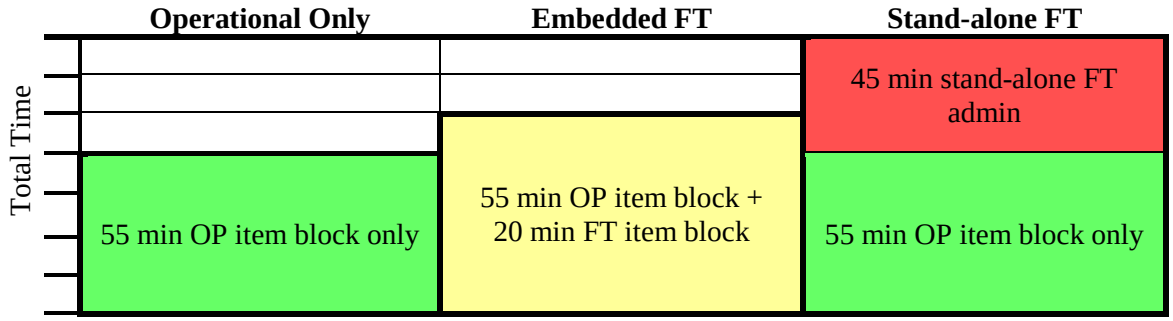

Figure 1. Visualizing test length in embedded vs. stand-alone field testing

Based on the figure above, you can see the differences in the impact of a single administration's testing time using embedded field testing compared to the overall testing time for a stand-alone field test.

Questions to Answer

Based on the information presented above, the Task Force should be prepared to address the following questions:

1. Field testing is a necessary part of assessment development. With a lower-risk timeline (see *Timeline for Test Development and Administration Brief*), there are more options for field testing. Should the new assessment prioritize embedded field testing if possible?
2. Considering what you know about field testing and the other briefs on content coverage and item types, how long should students spend on each end-of-year summative assessment?