BRIEF #7: ALABAMA ASSESSMENT TASK FORCE

COMPUTER ADAPTIVE TESTING

Juan D'Brot and Scott Marion, Center for Assessment

January 16, 2018

Test developers, practitioners, and educators are often excited when faced with the possibilities associated with computer based testing (CBT). One such possibility includes that of large-scale computer adaptive testing (CAT). It is difficult to speculate when the first formal adaptive tests were administered, however they have been used for some time. The Binet IQ test[1] (now known as the Stanford-Binet) is a well-known fully adaptive test (i.e., examinees receive different questions based on correct or incorrect responses). It was first administered in 1905 and required one-on-one administration. Through the use of modern testing methods and CBT, adaptive tests can be administered simultaneously to students across a state (and in some cases, across a country).

Features of Adaptive Tests

Adaptive tests differentiate themselves from fixed-form tests by providing examinees with different questions depending on how they respond to test items or sets of items. However, all tests—fixed or adaptive—must adhere to certain technical requirements. These include reliability, fairness, and validity[2]. These concepts are described in more detail in relation to summative testing below.

- **Reliability**: Generally, reliability refers to the pieces of information that help us determine whether a test is precise, reliable, or consistent enough *for the intended use of that test*.
- **Fairness**: Fairness emphasizes that the test must be fair, accessible, and appropriate for all individuals in the intended population *for the intended use of that test*.
- **Validity**: Validity refers to the degree to which evidence and theory support the interpretations of test scores *for the intended use of that test*.

[1] Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. L'Année Psychologique, 11, 191-244.

[2] AERA, APA, & NCME, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

As we can see in the descriptions provided above, all arguments about reliability, fairness, and validity are based on the intended use of the test. Fixed-form tests for accountability prioritize content coverage, sufficient reliability and precision to make claims about proficiency, fairness about students taking the test, and generalizability of claims across students. Adaptive tests for accountability seek to make the same claims, but have a greater ability to be more precise, more efficient, more targeted with appropriate set of items for higher and lower performing students.

Types of Adaptive Tests

Adaptive tests have the ability to adapt to a student's ability by providing more or less difficult items based on student responses. Furthermore, adaptive tests tend to reduce barriers to motivation associated with test takers receiving items that are too difficult or too easy. However, the adaptivity offered by CAT differs based on the resources available to the development effort. A general conceptualization of CAT is provided in the figure below.

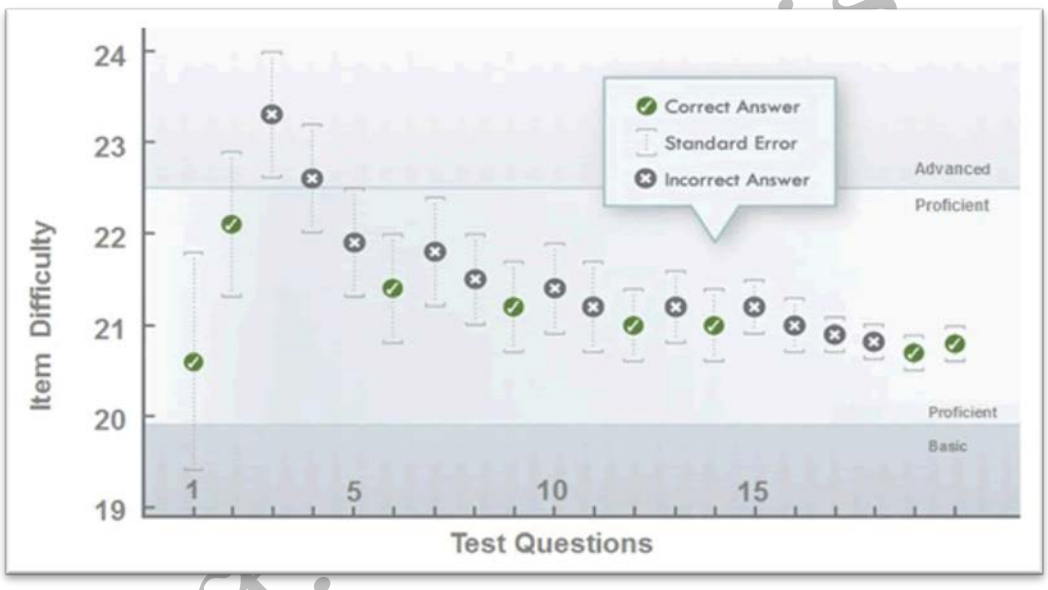

Figure 1. Common conceptualization of an adaptive test[3].

In reality, adaptive tests can vary in the level of adaptivity significantly. Four common adaptive approaches include (1) linear on the fly testing, (2) multi-stage testing, (3) multi-stage on the fly testing, and (4) computer adaptive testing. In these examples, the approaches are increasingly adaptive in nature. There are additional resources that are required to support increases in the adaptivity of a test. These include, but are not limited to an increased item pool, immediately scoreable items, increased research capacity to simulate CAT administrations, a CAT delivery

[3] From http://www.ascd.org/publications/educational-leadership/mar14/vol71/num06/The-Potential-of-Adaptive-Assessment.aspx

system, and appropriate software to account for additional analyses associated with CAT. The four sample types of adaptive testing are described in further detail below.

1. **Linear on the fly testing (LOFT)**: All items are selected at the start of the test (i.e., a fixed form) and are adapted based on prior test performance.
2. **Multi-stage testing (MST)**: Pre-determined forms are adapted to the student at pre-determined stages (e.g., after 15 items or after a cluster of topic-specific items). Students are routed to forms of varying difficulty based on performance in the previous stage.
3. **Multi-stage on the fly testing (MSOFT)**: Combines LOFT and MST testing. Forms are created on the fly at pre-determined stages (e.g., after 15 items or after a cluster of topic-specific items). Because forms are not pre-determined, it requires more items to populate optimized forms. However, it also reduces the exposure of items because fewer students see the same items.
4. **Computer adaptive testing (CAT)**: Fully individualized by student and adaptive at each item. It is the most precise and potentially the shortest test. If done well, it minimizes the exposure of items more than other types of adaptive testing. However, it requires the most investment and the largest pool of items with appropriate ranges of difficulty and complexity.

The benefits of CAT are maximized when CBT is supported fully throughout a state. It is possible to support comparable PPT and CAT scores through certain field test designs and administration approaches; however, the constraints are numerous. Supporting a dual mode assessment system with CAT requires an in-depth field test design, a robust research agenda, longer administration windows, a larger budget, and an extensive support plan for training and help-desk access.

Questions to Answer

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

1. In the December meeting, participants provided a preliminary recommendation to not implement computer adaptive testing. Given your recommendation based on the brief, *Considerations for Paper Pencil Testing vs. Computer-Based Testing*, should the Task Force continue to recommend excluding CAT from the assessment system?
2. If the state of Alabama were to support adaptive testing, what degree of adaptivity should be supported? Keep in mind the four types of adaptive testing and that fully adaptive tests should require 5-10x the intended test length to support sufficient items, depending on exposure estimates.
    a. Note: On the low end, that is only slightly more than supporting multiple forms in a single year, but backup—or breach—forms may be pushed to later years of an assessment program. Other factors to consider include an adaptive delivery system, an adaptive engine, and a vendor's research capacity.