Some Considerations Related to the Use of Adaptive Testing for the Common Core Assessments

Walter D. Way, Senior Vice President, Psychometric & Research Services, Pearson

Jon S. Twing, Executive Vice President, Assessment & Information, Pearson

Wayne Camara, Vice President, Research & Development, The College Board

Kevin Sweeney, Executive Director, Psychometrics, The College Board

Stephen Lazer, Vice President, Assessment Development, ETS

John Mazzeo, Vice President, Statistical Analysis & Psychometrics Research, ETS

February 2010

Preface

ETS, Pearson, and the College Board have formed a collaboration to explore how innovative approaches and best practices in high-quality assessments can be applied to the creation of a common assessment system. In a recently released paper, *Thoughts on an Assessment of Common Core Standards*, we raised some key assessment-design questions and discussed some ideas for a systematic high-level assessment design that satisfies many of the needs expressed by stakeholders. In this new paper, we extend our previous discussions by addressing the topic of adaptive testing in further detail. Our organizations have been involved with research, development, and operational implementation of adaptive testing for several decades. Because of this extensive expertise, we believe that we can offer unique insights based on the collective experience of our staffs.

In the spirit of our previous work, we offer our thoughts to stimulate discussion rather than to recommend a single, specific position. ETS, Pearson, and the College Board are excited to be a part of the national discussion on new assessment systems and hope this paper will contribute to the dialogue.

Introduction

ETS, Pearson, and the College Board believe that computerized adaptive-testing approaches hold significant promise within a common core assessment system. Our reasons for this position can be seen in the U.S. Department of Education (USED) notice related to the development and implementation of high-quality assessments based on common standards. This notice referred to three essential characteristics of summative assessments:

- Individual student achievement as measured against standards that build toward college and career readiness by the time of high school completion;
- Individual student growth (that is, the change in student achievement data for an individual student between two or more points in time); and
- The extent to which each student is on track, at each grade level tested, toward college or career readiness by the time of high school completion.

Because adaptive testing targets assessment to best measure each student's knowledge and skill, it is particularly effective in a summative assessment context for tracking individual student growth over time. Furthermore, it is highly compatible with the concept of vertically aligned standards and curricula that progress toward college and career readiness.

The USED notice referred to additional desirable characteristics of a common core assessment system, such as providing a fast turnaround of results, using technology effectively and appropriately, maintaining security, and providing access to the broadest range of students, with appropriate accommodations for students with disabilities and English-language learners. Adaptive testing seems especially well suited to achieve these characteristics. It offers additional advantages as well, including:

- Shorter testing times and more efficient use of computers;
- A more motivating assessment because content is administered at a more engaging level of difficulty, particularly for struggling students;
- Increased precision of measurement; and
- The potential to more effectively utilize open-ended/performance-based testing.

In general, adaptive testing seems to be consistent with many of the features desired in a common core assessment system. However, our experience also tells us that adaptive testing is not the "magic bullet" that some would consider it to be. We therefore believe the dialogue regarding the role of adaptive testing within a common core assessment system should strive to achieve a balance not unlike what many aspire to see for the assessment system itself.

In this paper, we discuss some important considerations related to the use of adaptive testing within a common core assessment system, particularly as used for summative purposes. Specifically, we address three considerations:

- Differences between adaptive testing and linear testing for high-stakes summative purposes;
- Using out-of-level content within adaptive testing; and
- The impact of content, item formats, and innovation on adaptive testing.

Adaptive Testing vs. Linear Testing

Summative testing in the United States has always placed a premium on producing scores on different versions of tests so that the results can be compared across administrations. The general term to describe this process is score linking. Test equating is the strongest and most restrictive form of linking. Following the application of an equating procedure, alternate forms of the same test yield scale scores that can be used interchangeably — even though they are based on different sets of items.

By definition, adaptive tests given to different individuals vary in their statistical specifications and, as such, cannot be considered equivalent in the strictest sense. On the other hand, the theoretical basis of adaptive testing provides a commonly accepted rationale for linking and interpreting scores on a common scale. The trade-off is one where the statistical assumptions needed for equating in its strictest sense are given up in exchange for the increased flexibility and precision of measurement afforded by the adaptive testing methodology. It is important to understand this trade-off because it anticipates possible limitations and criticisms of adaptive testing. For example, critics may suggest that adaptive testing is problematic because all students are not completing the same items or exposed to the same content in the same manner. However, such an adaptive nature to item selection is important if we want to have sufficient items targeted to the specific ability levels of students who vary considerably in knowledge and achievement. This targeting is essential to provide accurate diagnostic information on individual students.

In contrast, paper and linear computer tests must include more items across a wider range of difficulty levels to accommodate the full range of ability levels of the group tested. Critics may also fault adaptive testing because item-context effects or item-order effects are not controlled by the adaptive testing algorithm. Or they might state that adaptive testing is not fair in the situation where wrong responses are made at the beginning of the adaptive test because of nervousness or carelessness. According to critics, this could cause an overly low estimate of a student's ability since, even if the student begins responding correctly, the computer is less likely to administer difficult enough questions for the student to demonstrate their true ability.

These criticisms are largely unique to adaptive testing. In traditional test settings involving linear test forms, item-context and item-order effects are typically controlled by administering the items used for equating in the same or very similar positions over time. Furthermore, student scores on traditional tests are typically based on how many questions are correct and are usually independent of the order in which correct or incorrect responses occur. Although we know of no research indicating that these criticisms accurately reflect problems on any of the adaptive testing programs currently in use, the possibility that they could affect the score for an individual test taker cannot be disproven.

Thus, it is important to acknowledge and carefully consider the threats to valid scores and interpretations that may affect adaptive testing. It is vital to design adaptive testing systems

to minimize these threats, and to set up effective monitoring systems once adaptive testing is in place. We value adaptive testing because we recognize the advantages that it can bring to the common core assessments. At the same time, we encourage the recognition that adaptive testing in the context of the common core assessments must be pursued cautiously and deliberately.

Out-of-Level Content Within Adaptive Testing

One of the challenges associated with the use of summative assessments in the era of the No Child Left Behind (NCLB) legislation has been assessing the skills and knowledge of highly advanced and struggling students. The emphasis on reaching proficiency and the methodologies established to track adequate yearly progress (AYP) dictated that statewide tests should focus on optimizing the accuracy of measurement in the area of the scale near the "proficient" cut score. As growth measures have begun to be considered and introduced as part of determining AYP, it has become clear that the highest- and lowest-achieving students are typically not being measured very accurately, even though they can be reliably classified above or below the proficient level. The limitations in this regard stem not only from a dearth of items in the fixed-form assessments targeted to students at the extremes, but also from the restrictions imposed by the requirement that only on-grade-level content standards can be assessed.

The design of the common core assessments for English-language arts (ELA) and mathematics is expected to facilitate the measurement of growth for individual students across time. In addition, the common core standards should be aligned with college and work expectations, inclusive of rigorous content and application of knowledge through high-order skills, and internationally benchmarked. In concept, the common core standards will be vertically aligned so that the K – 12 standards will cascade down from the college and career readiness standards. In practice, this will be difficult to achieve, particularly at the high school level, where content is delivered in discrete courses and all students within a grade do not necessarily take the same courses.

In grades three – eight, vertical alignment in content standards is best achieved by focusing on the nature of content linkages from one grade to the next. These should be articulated clearly and, if done successfully, provide the basis for building assessments that are also vertically aligned. If the linkages and learning progressions across grade levels are clear in a set of vertically articulated standards, it becomes possible to make specific content-referenced statements about how we expect students to progress toward college and career readiness as they move along in the educational system. At the high school level, linkages are less clear. For example, the content strands defined for algebra may have little in common with the content strands defined for geometry.

Particularly in grades three – eight, we believe that using adaptive testing to permit some assessment of off-grade-level standards can substantially improve the measurement of students who are performing significantly below or above the levels of typical students within a particular

grade. This approach to assessment is consistent with a philosophy that focuses on personalized learning and recognizes the hierarchical nature of knowledge and skills that are required for students to be college and career ready. It is also consistent with the increased emphasis on measuring growth rather than just status in summative testing used for accountability purposes. It anticipates the blurring of grade levels, which many predict will begin to occur in classrooms of the future as the United States implements new instructional practices that have been proven successful in other countries.

From a psychometric standpoint, adaptive testing is a natural vehicle for permitting the measurement of off-grade-level content in the assessment system. Given clear and coherent vertically articulated standards, adaptive testing can administer off-level items to students in flexible and non-intrusive ways. It avoids the labeling that was traditionally associated with students taking out-of-level forms in norm-referenced testing because *all* students can begin at the same place and have their assessment branch as appropriate based on their performance. Moreover, in an adaptive system where student results and the vertically articulated standards are expressed on the same common scale, each student can be compared to the same "on-grade" standards regardless of whether or not off-grade items were used in their particular adaptive test. Given such a scale, if one fifth-grade student received only fifth-grade items, another fifth-grade student received some sixth-grade items, and yet another fifth-grader received some fourth-grade items, the adaptive testing results for all three students could be compared to the same fifth-grade performance standards.

At several of the recent public meetings sponsored by USED on common core assessments, advocates of students with disabilities and English-language learners argued against including alternate assessments with content at a lower level than that administered to general education students. This could be construed as an indictment of off-grade-level testing and, in the context of NCLB-era summative testing, such a position is understandable. It has been such a long and hard-fought battle to have these students included in appropriate instruction and assessment that there is a reluctance to consider any compromise that might jeopardize the rigor of the curriculum offered to them. But our vision of adaptive testing supports a rigorous, vertically aligned curriculum and a set of standards for *all* students. Although we recognize the continued importance of grade level as the organizing basis for curriculum, instruction, and assessment, the common core standards appear to be focused on growth and the acquisition of the knowledge and skills that are ultimately needed to prepare *all* students for college and the workplace. This requires us to rethink what "off-grade-level" means in terms of both instruction and assessment.

Some experts may contend that it does not really matter whether off-grade-level content is included in a common core adaptive testing system. We acknowledge that it is possible to implement adaptive testing using pools of items that are confined to grade-level standards. However, our experience with adaptive testing has demonstrated that it is very challenging to develop item pools that are robust enough to accurately measure students at the extreme levels

of knowledge and skills. Sharing items from item pools targeted at adjacent grade levels can significantly bolster the ability of item pools to accurately measure students at the extremes.

In the end, using adaptive testing for the common core assessments that span to off-grade-level content will require a change to the Elementary and Secondary Education Act (ESEA) legislation and/or an articulation of the common core standards that permits the issue to be finessed. We hope that one of these possibilities emerges, as we believe that adaptive testing with the common core assessments will be greatly enhanced if some spanning of grade-level content can be done.

Impact of Content, Item Formats, and Innovation on Adaptive Testing

Content is not any less important in adaptive testing than it is in linear testing. Clearly, the standards upon which the common core assessments will be based must be appropriately covered within an adaptive testing framework. But assessment content is more than just what is assessed; it is also how. It is universally acknowledged that the computer has the potential to assess what students know and can do in new and better ways. As such, the common core assessments should incorporate innovation to the greatest extent possible. This clearly has implications for adaptive testing. In short, it must be possible for adaptive testing to work with more than traditional multiple-choice questions.

Unquestionably, adaptive testing works most efficiently in subject areas where content is measured using discrete questions that can be objectively scored right or wrong. When sufficient item pools exist, adaptive tests can consistently satisfy content blueprints and accurately measure students across a wide range of knowledge and skill levels. However, when questions are associated with a common stimulus or problem scenario, the content and statistical characteristics associated with individual questions become confounded with the presentation of the stimulus, passage, or scenario. This introduces a number of complexities that the adaptive testing algorithm must be programmed to deal with, and these complexities impact efficiency.

The development of innovative items and tasks will likely further complicate the deployment of adaptive testing for the common core assessments. Assuming such item types can even be scored by computer, they are likely to require more than simple right-or-wrong scoring. For such item types, it may be possible to utilize psychometric models that permit partial-credit scoring. Even so, the details of how an adaptive algorithm selects and scores innovative items or tasks may have to be quite different than an adaptive test that is composed entirely of items scored right or wrong.

In ELA, the draft common core standards for college readiness recognize skills and knowledge in reading, writing, speaking, and listening. Similar standards are almost certain to cascade down the K – 12 span. The multidimensional nature of these skills will present challenges for using adaptive testing with ELA. In addition, computer-delivered ELA assessments will almost certainly

require students to produce samples of writing and speech. These performance samples are easily captured by computer, but it may or may not be possible to score them by computer.

In mathematics, there are currently examples of online constructed-response items that vary greatly in the complexity of response that is required of students. One end of the spectrum might consist of answer-only items, in which students provide only the final result of their work, such as a numeric response, mathematical expression, or plotted line. The other end of the spectrum might be extended-response items requiring students to embellish their final results with justification or reasoning using a choice of text, equations, graphs, and/or freehand drawings. Such constructed-response items might also vary in whether they involve single or multiple parts; in the latter case, students might respond to multiple screens with a single stimulus and multiple dependent or independent stems. Furthermore, subsequent parts may or may not depend on the answer to previous parts.

Notwithstanding the continued development and refinement of automated scoring algorithms for constructed-response and related item types, we believe that plans for adaptive testing within the common core assessments will have to adopt approaches to adaptive testing that are different from those employed in the world of "all-multiple-choice" tests in order to incorporate innovative item types and tasks that cannot be scored without human intervention. Although the resulting system may not have all of the same features currently associated with many adaptive test applications, the essential benefits of adaptive testing for a summative assessment system can still be realized.

For example, an adaptive system might be designed so that item types that cannot be scored in real time are administered after the computer has obtained a reasonably accurate estimate of a student's knowledge and skills from previously administered items that can be scored by machine. The test could administer some number of machine-scorable items and then, based on student performance to that point, choose among some number of innovative items or tasks previously calibrated by human scoring, administering the easier ones to lower-performing students and the more difficult ones to higher-performing students. Ideally, a distributed human-scoring system could efficiently route the responses to trained raters and minimize the lag time between the completion of the test and score reporting.

As our discussion suggests, we believe that commonly held notions of adaptive testing will change as innovation develops and assessments change. Fortunately, there is an active research base that is exploring the application of new psychometric models in the context of innovative assessment. As the utility of these models for high-stakes operational assessments becomes established, the approaches we use to deliver and score these assessments will evolve and improve. The common core assessments provide a unique opportunity to introduce and refine technology and innovation. One strategy for cultivating evolution and improvement is through the development of platforms that integrate balanced assessment systems. With such integration, innovation and experimentation can be encouraged in the lower-stakes components of the system, and those enhancements that prove effective can be efficiently

transitioned to the summative component. We therefore can imagine summative adaptive testing evolving over time from dynamically selecting items for the assessment to, perhaps, dynamically managing the assessment itself.

Summary

Although adaptive testing has been in use for many years in various settings (i.e., certification in technology and medicine, nursing licensure, military aptitude, higher education admissions, and interim school- and district-based programs), it has only recently begun to be used in summative statewide programs in K – 12. We believe the common core assessments can introduce adaptive testing on a large scale, as part of both a high-stakes summative component and an interim component. In this paper, we have discussed three specific topics related to a dialogue about the use of adaptive testing within the common core assessments. The key points from this discussion are summarized below:

1. There are trade-offs with choosing to use adaptive testing over conventional testing. Adaptive testing leads to more accurate measurement of individuals and better supports the measurement of growth over time, but does not produce tests that are equated in the strictest statistical sense. It is important to understand the threats to the validity of scores and interpretations resulting from adaptive testing, and to design adaptive testing programs in ways that allow these threats to be monitored and addressed.

2. The effectiveness of adaptive testing for measuring student growth can be enhanced by including off-grade-level items within the on-grade-level item pools. The development of common core standards characterized by a well-articulated vertical alignment leading to college and career readiness will serve to support this concept.

3. The common core assessments will require that adaptive testing expand to include more of the content that is important to assess. This requirement will lead to new adaptive testing approaches, scoring technologies, and supporting psychometric models.

While there are other important considerations related to adaptive testing for the common core assessments that deserve discussion, our focus on the three topics above was motivated by a premise that these are “big picture” concerns. Details related to the concrete steps needed to develop and implement a common core adaptive testing system are most deserving of explication and evaluation, and efforts along those lines should, no doubt, follow.