The Promise of Computer-Adaptive Testing

The Quest for Information

Stuart Kahl | December 2015

Background

Educators want a lot from the tests they administer. They want them to be short, but accurate; and they want them to provide a lot of information on individual student learning. Unfortunately, these test characteristics are not easily achieved by a single instrument. Adaptive testing is a popular form of computer-delivered assessment. For efficiency in obtaining accurate total test scores for students, it is clearly an effective approach. However, the most commonly used computer adaptive models today have not been designed to produce more detailed information on the abilities of students.

Computer adaptive testing (CAT) has been around for decades. Designed to arrive at reliable total test scores as efficiently as possible, the approach has been particularly useful in the areas of professional licensure and certification, where all that is expected is a "pass-fail" determination. The underlying rationale for the approach is that there is no need to administer the easiest items to high-performing examinees or the hardest items to low-performing examinees since such items contribute little to the quality of the estimates of performance scores for those examinees—scores derived from the Item Response Theory (IRT) scaling and analysis employed in most CAT programs today.

With the simplest computer adaptive model, the computer selects an item for an examinee to answer based on that individual's performance on the previous item (either a "1" or a "0" since most computer adaptive test items are scored dichotomously). If an examinee answers an item correctly (1 point), then the next item will be a harder one. If the individual answers an item incorrectly (0 points), the next item will be an easier one. As the process continues, the computer zeroes in on the items that best correspond to the ability level of the individual examinee because IRT scaling puts examinees and items on the same scale. When the precision of the ability estimate for an examinee reaches an established threshold (i.e., when the standard error for the person's score gets below a pre-designated value), the computer stops administering items and reports a score.

> "... The quick turnaround time for scores is a result, not of the tests being adaptive, but rather of their being comprised solely of multiple-choice items, which require no human scoring."

In the K-12 educational arena, this traditional adaptive approach, with minor variations, has been used for such purposes as growth monitoring and early warning to identify students at risk of failure on

subsequent high-stakes tests in reading and mathematics. Because computer adaptive items/tests are vertically scaled and cover content and skills appropriate for multiple grade levels, the technique has the advantage of producing more reliable scores for students who would score in the extremes on fixed, grade-level tests, which typically include few items that discriminate well among those students.

Another attribute of CAT that educators value is the immediacy of scores. However, the quick turnaround time for scores is a result, not of the tests being adaptive, but rather of their being comprised solely of multiple-choice items, which require no human scoring. This is a limitation of traditional computer adaptive tests because it means the tests do not effectively assess many broader, higher-order skills. While advances in technology-enhanced items and automated scoring can improve this situation, to the extent that immediate reporting is a requirement, CAT will be limited to the types of items that can be efficiently machine scored and will shortchange important competencies better measured by extended, constructed-response items and performance tasks.

> “Uninformed local educators might think a student scoring at a certain level could correctly answer all the questions below that level but could not correctly answer any questions higher on the scale. Life should be so simple!”

Accepting this limitation of CAT, educators generally still want more than total test scores. They want detailed information on the academic strengths and weaknesses of their students. In this regard, adaptive testing has not yet reached its full potential. Efforts to infer more diagnostic information from the results of computer adaptive tests not designed to produce such data have been ill-conceived and have led to inappropriate decisions regarding the needs of individual students. The remainder of this paper discusses some of these missteps, as well as more promising approaches that are being or could be employed.

Item Mapping

As stated previously, IRT analysis, typically used for CAT, puts items and students on the same scale. Picture two columns of information: in the left-hand column is a list of student scaled scores from highest to lowest; to the right is a set of test items listed from hardest to easiest.

Looking across the two columns, a student scaled score corresponds to a test item. The correspondence is this: of the group of students who earned that scaled score, a certain percentage of them would answer the question correctly. Sixty-seven percent is commonly used. Items higher in the list would be answered by lower percentages of students in that group.

Item mapping is a technique that was used in conjunction with the National Assessment of Educational Progress (NAEP) to describe the general capabilities of groups of students scoring at different points on the NAEP scale in a subject. Picture the two columns of information again. The left-hand column shows the NAEP scaled scores, while the right-hand column contains released items that have been put on the same scale as the larger set of NAEP items. NAEP reports then could make statements such as, “Of all the students scoring at 280, two thirds of them could answer the following questions correctly . . .” This reporting gives readers a general idea of the capabilities of students scoring at a particular point on the score continuum.

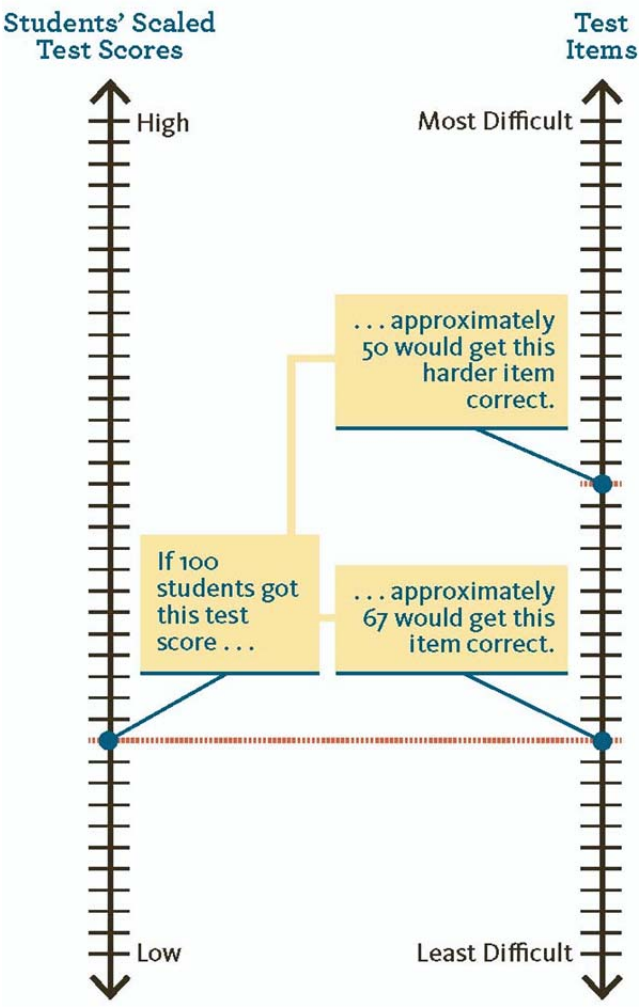
Probably because students taking computer adaptive tests are not administered the same items and because items in these tests are typically kept secure, providers of computer adaptive testing for schools and districts also offer item mapping

information to local educators. This includes released test items and information about how they line up relative to the reporting scale. This can be done because these released items, through past administrations to students along with items ultimately kept secure for operational use by schools, are placed on the same scale as the secure items.

Uninformed local educators might think a student scoring at a certain level could correctly answer all the questions below that level but could not correctly answer any questions higher on the scale. Life should be so simple! Slightly more knowledgeable educators might realize that students scoring at a particular level would have greater difficulty answering questions higher on the scale. But how much higher on the scale would one have to go to be relatively certain an item addresses a concept or skill for which a student needs additional instruction?

Think of a group of students with a particular score on a test. Then consider items above the ones corresponding to that scaled score. Going up that list of items, one encounters an item that 60 percent of the kids would be able to answer correctly, then an item 50 percent could answer correctly, then an item 40 percent could answer correctly. A large number of students in the group do not need help on these items. But which students do and which ones do not? Remember, these are released items not taken by the kids who took the computer adaptive test, but which have been scaled with all the items in the item pool to make item mapping possible.

Clearly, teachers who provide extra instruction to this group of students on the specific skills measured by such items would be wasting their efforts on many students, perhaps even half of them. Worse yet, equally large numbers of students who *do* need assistance with skills addressed by items lower on the scale would not get needed help as a result of decisions based on item mapping. NAEP appropriately uses probabilistic judgments via item mapping to describe the general capabilities of groups of students. However, teachers are mistaken if they believe item mapping is giving them actionable diagnostic information about individual students' specific skills.



Test Alignment

The quality of any test—adaptive or fixed—is partly a function of how well its items represent the target domain. The problems with some adaptive tests revolve around issues of alignment to content standards. Indeed, such issues were at the heart of NCLB peer reviews, which were not supportive of some adaptive testing programs for state accountability purposes. Before elaborating on these issues, it is appropriate to review four kinds of alignment conceived by Norman Webb of the University of Wisconsin.

Webb uses "standards" to refer to the broadest categories of content within a subject area, and

"objectives" to mean the next level of content/skill categories within standards. (Note: Different groups' content standards use different terminology for categories and subcategories of content. This document uses a variety of terms interchangeably.)

- ***Categorical concurrence*** refers to the commonality between the content categories of the standards and the content categories of the assessment [items]. However, in assessment alignment studies it means more than this; it refers to the extent to which items in a test can be matched to objectives within the standards.

- ***Depth of knowledge (DOK) consistency*** between standards and assessments refers to the match between the cognitive demand of items and the level of cognitive demand communicated by the wording of the objectives.
- ***Range of knowledge*** pertains to the breadth of coverage of topics or skills within standards and objectives.
- ***Balance of representation*** addresses the relative coverage of content categories (standards or objectives within standards) by items in a test—i.e., whether the categories are given emphasis consistent with pre-determined specifications.

Subtest Scores

IRT scaling of items in a bank to be used for a computer adaptive test assumes there is a single, uni-dimensional construct being measured (e.g., general mathematical ability) and that every item is an estimator of that ability, even though the items address multiple subdomains (e.g., geometry, measurement, numerical operations, algebra). For a test designed only to generate a total test score, this assumption is reasonable, and the adaptive testing algorithm for item selection described earlier seems to work fine. That is, selecting "next items" solely on the basis of item difficulty (and randomly with respect to other considerations) yields reliable total test scores. However, such adaptive tests do not fare well in alignment studies.

While items in the master item bank may be aligned to standards categorically and exhibit a range of depth of knowledge, the items selected for any individual student's test likely fall short in the areas of balance of representation (for content and cognitive complexity categories) and range of knowledge (subdomain coverage). Nevertheless, these problems seem to balance each other out, so that the total test score can be trusted.

All this is to say that subtest (subdomain) scores should not be generated from an adaptive test applying the simplest, total-test-oriented algorithm. (Also, it should be noted, IRT scaling, with its assumption of uni-dimensionality, is not the only scaling approach that could be used for CAT.) Constraints could be added to the algorithm to assure better distribution of items across subdomains within an individual student's test, but this would likely only improve the content validity of the total measure—an outcome that can also be accomplished by the computer administering a short, balanced, fixed test before moving into the adaptive mode. Remember, efficiency is a goal of adaptive tests, so if a student's final test is comprised of only 20 to 30 items, the representation of concepts and skills in a subdomain is still not likely to be good.

> "One approach more certain to address alignment and reporting issues would be for an adaptive test to be comprised of separate adaptive subtests for each subdomain."

In mathematics, if there are four or five content categories (standards), then there might only be five or six items representing each and, with the other item-selection criterion (item difficulty) being applied, coverage of the subdomain would be poor. More constraints could be added to the algorithm to remedy this situation—e.g., allocations of items across topics (objectives) within categories—but the

more constraints that are added for this reason, the longer the test needs to be, and many more items would be required for the item bank from which the computer draws.

The more specific the reporting categories are (the subsets of items for which scores are generated and reported), the more diagnostic the reported information is. However, the subcategories at the first level (e.g., geometry or statistics within the domain of mathematics) are still very broad. Scores in such areas are more useful for identifying general areas of weakness in instructional programs than for diagnosing the difficulties individual students are encountering. And, as suggested above, reporting individual scores at an even more specific level would greatly increase testing time and quantities of items needed in the item pool.

One approach more certain to address alignment and reporting issues would be for an adaptive test to be comprised of separate adaptive subtests for each subdomain. Items in the large item pool would be scaled separately for each subdomain, and when addressing a particular subdomain, the adaptive algorithm would only access items belonging to that subdomain. For a comprehensive total test, the standard error "stop criterion" for each subdomain could be relaxed to shorten the test somewhat. For interim tests, schools could choose single, subdomain adaptive tests throughout the year to coincide with their instructional sequences.

"Subtests" with More Meaningful Underlying Continua

Many have suggested that instead of traditional subtest areas, computer adaptive tests should focus on well-defined learning progressions. Such an approach could result in student reports that look like a graphic equalizer display for a sound system—showing where a student is on each of 10 to 15 progressions in mathematics, for example. This approach has great potential, but it also has implications that should be considered relating to item development and inclusion in the item pool (or separate pools for progressions). Those implications are illustrated by the real situation described below.

Consider the scaling test Metametrics® equates to other reading tests, so the latter can yield Lexile® scores. The passages in the scaling test represent a range of text complexity; in fact, they are placed on a scale based on text complexity. Items associated with the passages are put on the same or a parallel scale. If the passages are listed vertically, the items associated with a particular passage should be clustered on the scale in the vicinity of where the passage appears. The item difficulties of items associated with two passages that are close on the scale would overlap some, but the item difficulties of items associated with two passages that are far apart on the scale should not overlap.

> "One approach to adaptive testing that offers greater control of the content within categories is 'multi-stage adaptive testing.' "

There should not be "outlier" items associated with a passage. Since the goal of Lexile scores is to put students on the same scale as the passages, outlier items would produce contradictory information and would not be helpful. Of course, this assumes there is a reasonable ordering of passages, which is indeed accomplished by the use of a text complexity formula. The item bank used for an adaptive test focused on a single learning progression should be thought of in the same way. The items in the bank associated with a particular concept or skill should be clustered in a vertical difficulty continuum somewhere near where the concept or skill appears on a continuum representing the progression. This can only be accomplished if there is a logical and reasonable ordering of concepts and skills in the progression, an ordering based on cognitive readiness and/or commonly accepted curricular/instructional ordering. Of course, concepts and skills have many manifestations in items of varying complexity or

sophistication. Nevertheless, outlier items associated with a concept or skill should be avoided.
Here's an extreme example to illustrate the outlier issue. Nobody would question that students need to master addition of whole numbers before being able to solve an algebraic equation. Yet, more ninth graders might succeed at solving an algebraic equation than adding thirty eight-digit whole numbers. The potential is great for all students to make careless addition errors in the latter, including those with a good command of addition of whole numbers. The point is that if a large set of items is going to be developed that doesn't specifically address well-defined progressions (as is the case for today's commonly used computer adaptive tests), then there will be items in the set that are "out of sync" with other items addressing the same construct. The out-of-sync items will therefore not be useful for a progression-based adaptive test.

Traditional adaptive testing models based on item difficulty can yield more diagnostic information if they are adaptive at the learning progression level. However, if such tests are developed, outlier items should be excluded from the associated item pools. As an aside, there are adaptive testing models that do not base identifying a student's "next question" solely on item difficulty. Children's Progress has produced K-3 computer-adaptive testing based on concept/skill networks. Showing where students are in such networks clearly has diagnostic value. Similarly, a federally funded state consortium, the Dynamic Learning Maps (DLM) Alternate Assessment System Consortium, is developing computer adaptive tests based on learning maps.

Multi-Stage Adaptive Tests

While balance of representation across reporting categories can be readily addressed by a constraint to the adaptive testing item-selection algorithm based on item difficulty, representation of content within a category (range of knowledge) cannot. Multiple constraints at another level (one possible solution) would necessitate a much larger item pool—and an item pool for traditional adaptive testing must be very large already.

One approach to adaptive testing that offers greater control of the content within categories is "multi-stage adaptive testing." With this approach, the algorithm selects sets of items or testlets, instead of individual items for a particular examinee. The first stage would consist of a short fixed test taken by all students. At the next stage, a testlet would be selected for each individual student based on his/her performance on the initial fixed test. The second stage could also involve separate testlets for different subdomains. More stages would be possible, although a lot could be accomplished by a total of three.

Whether focused on traditional reporting categories or well-defined, researched learning progressions, the bank of testlets would represent different levels of content/complexity or different parts of a learning progression. There would be some overlap in the content/complexity of adjacent testlets to accommodate students for whom the initial algorithmic decision is not so clear-cut based on performance on the short fixed test. The hand-picking of items for testlets by testing professionals assures that the range of knowledge associated with a testlet is not a matter of chance. Similarly, the purposeful selection of items for testlets can address a concern some have about adaptive tests not allowing very low-performing students the opportunity to answer grade-level questions. The testlet developer can select easy grade-level questions for a testlet, rather than hoping a computer will not simply pick lower-than-grade-level items to satisfy a need for easier items for a particular student.

Human item selection can also assure that no items in a testlet key others in that testlet or are simply clones of others. That is, it avoids poor testing practices that are distinct possibilities when the computer selects items based only on item difficulty—problems more likely to occur with added constraints on a computer-based item selection algorithm. Another significant benefit of multi-stage adaptive testing is that many fewer items are required for the item pool from which testlets are created. Furthermore, for lower-stakes interim testing, teachers could view the items/testlets.

Conclusions

Historically, computer adaptive tests were designed to produce total test scores as efficiently as possible. Attempts to glean diagnostic information from such testing were often inappropriate, leading to unjustified decisions by teachers. However, computer adaptive testing has tremendous potential to provide better information for teachers to use in making instructional decisions through the effective use of constraints to the item selection algorithm, multi-stage adaptive tests, and a focus on learning progressions or maps as opposed to item difficulty without consideration of content.

The least constrained item selection algorithms for adaptive testing yield tests that have significant problems related to content alignment—specifically balance of representation across and range of knowledge within content categories. These tests can produce reliable total test scores, but no finer grained reporting that can be relied upon. Adding a constraint calling for selected items to be spread across categories improves the balance of representation, but not range of knowledge. Thus, subtest scores would still be spurious. Additional constraints could improve the range of knowledge within categories, but the tests would have to be longer, and the item banks supporting them significantly larger.

The most meaningful and useful information for teachers that computer adaptive testing could yield would result from adaptive tests focused on well-defined continua of content and skills (learning sequences or progressions). Development and selection of items for the supporting item banks would have to be purposeful, but where a student scores on a particular progression would provide far better information on what the student has mastered and what he or she should tackle next.

Multi-stage computer adaptive testing combines the best features of adaptive and fixed-form testing. The adaptive nature of the second and any subsequent stages (selecting testlets based on previous stage results) assures that the results would be reliable for students in any part of the ability distribution—even the extremes. Development of the testlets by professionals can ensure there are no alignment problems or items that don't belong together within any testlet. The testlets in the later stage(s) can span meaningful learning continua or progressions, thereby maximizing the quality of information teachers can use to make instructional decisions.